

INFERRING CELLULAR MECHANISMS OF
TUMOR DEVELOPMENT FROM TISSUE-SCALE
DATA: A MARKOV CHAIN APPROACH

DISSERTATION

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften der
Technischen Universität Dresden

von

Dipl.-Math. Thomas Buder

geboren am 14.12.1986
in Bad Muskau

Eingereicht am 04.04.2018

Verteidigt am 28.08.2018

Gutachter:

Prof. Dr. Dietmar Ferger
Prof. Dr. Niko Beerenwinkel

ABSTRACT

Cancer as a disease causes about 8.8 million deaths worldwide per year, a number that will largely increase in the next decades. Although the cellular processes involved in tumor emergence are more and more understood, the implications of specific changes at the *cellular* scale on tumor emergence at the *tissue* scale remain elusive. Main reasons for this lack of understanding are that the cellular processes are often hardly observable especially in the early phase of tumor development and that the interplay between cellular and tissue scale is difficult to deduce. Cell-based mathematical models provide a valuable tool to investigate in which way observable phenomena on the tissue scale develop by cellular processes. The implications of these models can elucidate underlying mechanisms and generate quantitative predictions that can be experimentally validated. In this thesis, we infer the role of genetic and phenotypic cell changes on tumor development with the help of cell-based MARKOV chain models which are calibrated by tissue-scale data.

In the first part, we utilize data on the diagnosed fractions of benign and malignant tumor subtypes to unravel the consequences of genetic cell changes on tumor development. We introduce extensions of MORAN models to investigate two specific biological questions. First, we evaluate the tumor regression behavior of pilocytic astrocytoma which represents the most common brain tumor in children and young adults. We formulate a MORAN model with two absorbing states representing different subtypes of this tumor, derive the absorption probabilities in these states and calculate the tumor regression probability within the model. This analysis allows to predict the chance for tumor regression in dependency of the remaining tumor size and implies a different clinical resection strategy for pilocytic astrocytoma compared to other brain tumors.

Second, we shed light on the hardly observable early cellular dynamics of tumor development and its consequences on the emergence of different tumor subtypes on the tissue scale. For this purpose, we utilize spatial and non-spatial MORAN models with two absorbing states which describe both benign and malignant tumor subtypes and estimate lower and upper bounds for the range of cellular competition in different tissues. Our results suggest the existence of small and tissue-specific tumor-originating niches in which the fate of tumor development is decided long before a tumor manifests. These findings might help to identify the tumor-originating cell types for different cancer types.

From a theoretical point of view, the novel analytical results regarding the absorption behavior of our extended MORAN models contribute to a better understanding of this model class and have several applications also beyond the scope of this thesis.

The second part is devoted to the investigation of the role of phenotypic plasticity of cancer cells in tumor development. In order to understand how phenotypic heterogeneity in tumors arises we describe cell state changes by a MARKOV chain model. This model allows to quantify the cell state transitions leading to the observed heterogeneity from experimental tissue-scale data on the evolution of cell state proportions. In order to bridge the gap between mathematical modeling and the analysis of such data, we developed an R package called CELLTRANS which is freely available. This package automatizes the whole process of mathematical modeling and can be utilized to (i) infer the transition probabilities between different cell states, (ii) predict cell line compositions at a certain time, (iii) predict equilibrium cell state compositions and (iv) estimate the time needed to reach this equilibrium. We utilize publicly available data on the evolution of cell compositions to demonstrate the applicability of CELLTRANS. Moreover, we apply CELLTRANS to investigate the observed cellular phenotypic heterogeneity in glioblastoma. For this purpose, we use data on the evolution of glioblastoma cell line compositions to infer to which extent the heterogeneity in these tumors can be explained by hierarchical phenotypic transitions. We also demonstrate in which way our newly developed R package can be utilized to analyze the influence of different micro-environmental conditions on cell state proportions.

Summarized, this thesis contributes to gain a better understanding of the consequences of both genetic and phenotypic cell changes on tumor development with the help of MARKOV chain models which are motivated by the specific underlying biological questions. Moreover, the analysis of the novel MORAN models provides new theoretical results, in particular regarding the absorption behavior of the underlying stochastic processes.

ACKNOWLEDGMENTS

I would like to express my gratitude to Prof. Anja Voß-Böhme and Prof. Andreas Deutsch for almost unlimited advice and support during many years. In particular, I thank both for the possibility to pursue my own ideas and very valuable, countless hours of scientific discussions and honest feedback for my work.

I would like to thank all colleagues in the department *Innovative Methods of Computing* (IMC) at the *Centre for Information Services and High Performance Computing* at TU Dresden for very helpful comments and discussions in many seminars and for creating a productive and pleasant working atmosphere.

I also would like to express my thanks to Fabian Schwarzenberger for proof-reading this thesis and for valuable support in improving some of the figures.

Special thanks go to Barbara Klink for priceless explanations of biological and medical coherences which helped to design mathematical models that tackle specific biological questions. I also appreciate the time that I spend together with Michael Seifert in one office and would like to thank him for interesting discussions and the idea to develop CELLTRANS.

I also would like to thank Prof. Dietmar Feger and Prof. Niko Beerenwinkel for accepting reviewing this thesis.

I'd like to say a big thank you also to M. for always bringing a smile on my face.

Last but not least, I would like to express my sincere gratitude to my family for their support not only during the time in which this thesis emerged but during my whole life. This thesis is dedicated to them.

Thank you!

CONTENTS

1	INTRODUCTION	1
1.1	Background on cancer biology	1
1.2	Scope & outline of this thesis	4
2	MATHEMATICAL AND MODELING FOUNDATIONS	9
2.1	MARKOV chains	9
2.2	The MORAN model	17
2.3	A MORAN model on a crypt-like structure	25
I	MORAN MODELS FOR THE INFERENCE OF GENETIC CELL CHANGES ON TUMOR DEVELOPMENT	37
3	EVALUATION OF TUMOR REGRESSION IN PILOCYTIC ASTROCYTOMA	39
3.1	Clinical characteristics of pilocytic astrocytoma	39
3.2	A MORAN model for pilocytic astrocytoma growth and progression	41
3.3	Results & clinical implications	57
3.4	Discussion & outlook	59
4	PATTERNS OF TUMOR PROGRESSION PREDICT TUMOR-ORIGINATING NICHES	65
4.1	Tumor progression on the cellular and the tissue scale	65
4.2	Mathematical model	67
4.3	Results	76
4.4	Discussion	80
II	THE ROLE OF PHENOTYPIC CELL CHANGES IN TUMOR DEVELOP- MENT	83
5	QUANTIFYING STOCHASTIC CELL STATE TRANSITIONS	85
5.1	Characteristic cell state compositions in cancer	86
5.2	Introducing CELLTRANS	87
5.3	Case studies demonstrating the applicability of CELLTRANS	94
5.4	Discussion	109
6	QUANTIFICATION OF STATE TRANSITIONS BETWEEN GLIOBLAS- TOMA CELLS	113
6.1	Glioblastoma multiforme and the cancer stem cell hypothesis	113
6.2	Material & methods	115
6.3	Results	117
6.4	Discussion	122
7	CONCLUSIONS	125
7.1	Summary	125
7.2	Novelty and relevance of the results	126

7.3 Outlook	127
A APPENDIX	129
A.1 Residual cerebellar postoperative pilocytic astrocytoma volume and outcome	129
A.2 CELLTRANS vignette	130
A.3 Additional model predictions for GBM cell state transitions	134
BIBLIOGRAPHY	139

LIST OF FIGURES

Figure 1	The classical MORAN model.	17
Figure 2	State space of the MORAN model on a crypt-like structure.	28
Figure 3	Intra-crypt short-range dynamics.	29
Figure 4	States and dynamics in the TGP process.	43
Figure 5	Decomposition of the TGP process.	46
Figure 6	The absorption probabilities in the absorbing states N and E of the TGP process in dependency of the risk coefficient γ	54
Figure 7	Tumor regression in the TGP process.	62
Figure 8	Parameter estimation from epidemiological data and derivation of the PA-regression-function.	63
Figure 9	Non-existence of an extent of resection threshold in PA.	64
Figure 10	MORAN dynamics with different spatial cell arrangements.	68
Figure 11	Tumor progression types and patterns.	71
Figure 12	The absorption probability α_{1D}^N of the 1D MORAN model as function of the risk coefficient.	78
Figure 13	Estimated tumor-originating niche sizes based on tumor progression patterns.	79
Figure 14	General CELLTRANS application workflow.	89
Figure 15	CELLTRANS model predictions for colon cancer cell lines.	99
Figure 16	CELLTRANS model predictions and validations.	106
Figure 17	Influence of non-pure initial cell state compositions on the predictions of CELLTRANS.	107
Figure 18	Cell state transitions and regularization of the matrix root.	110
Figure 19	Interaction graphs of the estimated transition matrices of GBM cell state transitions.	116
Figure 20	Predicted GBM equilibrium composition and the original tumor composition.	118
Figure 21	How to read data into CELLTRANS.	131
Figure 22	Options for plots in CELLTRANS.	133
Figure 23	Estimated and measured glioblastoma multiforme (GBM) phenotypic compositions from experiments with two different initial compositions.	136

LIST OF TABLES

Table 1	Overview of the components in the cell-based MARKOV chain models in this thesis.	7
Table 2	The absorption probabilities and times of MORAN models on different structures.	34
Table 3	Comparison of absorption probabilities (3.9) and the results of 10000 simulations of trajectories of the tumor growth and progression (TGP) process.	53
Table 4	Predicted regression probability for cerebellar pilocytic astrocytoma (PA) based on the absolute residual tumor size.	59
Table 5	Comparison of the approximate absorption probabilities and the results of 10000 simulations of trajectories of the 1D MORAN model.	77
Table 6	Tumor progression patterns in dependency of the spatial cell arrangement in the 1D MORAN model.	78
Table 7	Estimation of the homeostatic competition range N in different tissues.	81
Table 8	Main applications of CELLTRANS and corresponding implemented functions.	94
Table 9	Predictions of CELLTRANS for different choices of the time step length τ for data from the case study in [147].	104
Table 10	Predictions of CELLTRANS for different choices of the time step length τ for data from the case study in [40].	104
Table 11	Definition of 16 distinct GBM cell phenotypes.	115
Table 12	Single marker estimates and equilibrium composition under the assumption of marker-independent transitions of GBM phenotypes in normoxia.	120
Table 13	Single marker estimates and equilibrium composition under the assumption of marker-independent transitions of GBM phenotypes in hypoxia.	121
Table 14	Novel analytical results for the MORAN models in this thesis.	127
Table 15	Residual cerebellar postoperative pilocytic astrocytoma volume and outcome.	129
Table 16	Estimated transition probabilities per normoxia cell cycle length of 3 days in % between the 16 GBM cell phenotypes in normoxia.	134
Table 17	Estimated transition probabilities per 3 days in % between the 16 GBM cell phenotypes in hypoxia.	135

Table 18	Estimated transition probabilities in % between GBM cell phenotypes in normoxia per 3 days assuming independent cell state transitions.	137
Table 19	Estimated transition probabilities in % between GBM cell phenotypes in hypoxia per 3 days assuming independent cell state transitions.	138

ACRONYMS

- CSC cancer stem cell
- FACS fluorescence-activated cell sorting
- GBM glioblastoma multiforme
- PA pilocytic astrocytoma
- TGP tumor growth and progression

INTRODUCTION

Cancer is, in essence, a genetic disease. — Bert Vogelstein [140]

Contents

1.1	Background on cancer biology	1
1.1.1	The role of genetic cell changes in tumor development	2
1.1.2	The role of phenotypic cell changes in tumor development	3
1.2	Scope & outline of this thesis	4

1.1 BACKGROUND ON CANCER BIOLOGY

Cancer as a disease causes about 8.8 million deaths worldwide per year which represents 15.7% of all causes for human deaths [141]. Moreover, about 14.1 million new cases are diagnosed worldwide per year and this number is expected to increase to 24 million by 2035 [92]. This indicates that cancer is a major challenge and will be even more relevant in the future. From a historical perspective, many hypotheses have been formulated about the underlying mechanisms of cancer development including infectious diseases, dysregulated differentiation or a defective immune system. However, it took until the discovery of the exact structure of the DNA in 1953 until scientists understood that all these mechanisms are responsible for genetic alterations within the DNA leading to uncontrolled proliferation of cells. Nowadays, it is known that also environmental and other non-genetic factors, such as phenotypic properties of cells, play a crucial role in cancer development.

Cancer cells can be regarded as abnormal and exhibit typical hallmarks such as resisting cell death, enabling replicative immortality or inducing angiogenesis [48]. In the context of this thesis, the role of genetic and phenotypic cell alterations in the acquirement of these hallmarks are of particular interest.

1.1.1 *The role of genetic cell changes in tumor development*

CELLULAR MULTISTEP PROCESS. It is widely accepted that most cancers develop by a *multistep process* of genetic changes on the cellular scale. In this process, *tumor progression* describes irreversible changes in the characteristics of cells which are obtained in a sequential manner and lead to a genetically altered subpopulation of cells with these new characteristics [103]. Usually, the starting point of this process is the acquirement of a single cancer-relevant genetic alteration within a single cell. Subsequently, this cell can undergo *clonal expansion*, i.e. it gives rise to descendants by cell divisions. This increase of cancer cell numbers leads to *tumor growth*. Further genetic alterations within this growing cancer cell population can accumulate such that the proliferation of these cells gets more and more uncontrolled. However, the possible death of the organisms is in most cases not caused by this local tumor growth, but a consequence of *metastasis* which describes a process in which cancer cells acquire the ability to spread to and start growing in a different organ [127].

Importantly, the characteristic numbers of genetic changes in the multistep process of tumor progression differ between cancer types. For example, the development of malignant colon cancer requires about seven of these steps [139] whereas the development of retinoblastoma, a cancer within the eye, requires two mutations in a certain gene [37]. In contrast, pilocytic astrocytoma, a low-grade brain tumor which is most often diagnosed in young adults and children, can be induced by a single specific mutation and is therefore often designated as single-step disease [63]. These examples indicate that there are fundamental differences in the genetic characteristics between different cancer types and that cancer cannot be regarded as *one* disease.

CELL COMPETITION & TUMOR REGRESSION. Genetically altered cells that arise in the cellular multistep process of cancer progression do not necessarily induce tumor formation but are exposed to competition with its corresponding wild-type cells. This competition can be roughly divided into a neutral and a selection phase depending on the stage of progression. The genetic alterations do not confer a proliferative advantage to the tumor cells in the neutral phase but do so in the selection phase [144, 146]. Within the neutral phase, the cell competition can be regarded to take place within normal tissue homeostasis. Hence, the range of this competition is determined by the underlying tissue structure suggesting that the spatial cell arrangement has a large influence on clonal expansion of cells. In particular, the spatial arrangement of cells influences where other cells can place its offspring and thereby determines to which extent cells can clonally expand within a tissue. Moreover, genetically altered cells must not go extinct in order to induce tumor formation. This establishment can be achieved by clonal expansion to a sufficiently large cell population [75]. Importantly, a tumor cell population can also fail to establish and there are

tumor types for which spontaneous regression of clinically detectable sizes is a common phenomenon, e.g. pilocytic astrocytoma. However, which mechanisms determine whether these tumors regress or not is an open question. We provide a model-based approach to evaluate tumor regression of pilocytic astrocytoma in Chapter 3.

EMERGENCE OF BENIGN AND MALIGNANT TUMOR SUBTYPES. Tumors can be classified with respect to the grade of tumor progression into *benign* and *malignant* stages. Malignant tumors are able to invade nearby tissue or to metastasize whereas benign tumors lack these abilities. Moreover, malignant tumor cells usually proliferate faster than benign tumor cells so that, in general, malignant tumors grow faster than benign tumors. A well-studied example of tumor progression from a benign stage to a malignant stage is the development of colon cancer. Here, a stepping stone is the formation of benign adenomas which requires two genetic alterations in both alleles of the APC tumor suppressor gene [111]. Subsequently, adenomas can progress to malignant colon cancer due to further malignant alterations. From a clinical point of view, benign tumor stages offer an opportunity for detection before malignant progression. One observes different progression types, namely with and without detectable benign precursor stages. Tissue-scale data on the ratios of the two progression types exhibit large differences between cancers. In which way these progression patterns on the tissue-scale emerge from the processes on the cellular scale is difficult to infer since the early phase of this multistep process is hardly observable. In Chapter 4, we shed light on this early phase of the multistep process of cancer development on the cellular scale with a MARKOV chain model that is calibrated with epidemiological data on benign and malignant tumors on the tissue scale.

1.1.2 *The role of phenotypic cell changes in tumor development*

PHENOTYPIC EQUILIBRIUM OF CANCER CELL POPULATIONS. During tumor growth and progression, the normal composition of a tissue is destroyed and tumor cells emerge which exist in different phenotypic states with different functional properties. The ability of cancer cells to change this phenotype is referred to as *phenotypic plasticity* and induces a highly dynamic overall tumor composition which exhibits a broad range of heterogeneity between different patients [18].

Similarly to normal tissue homeostasis, the dynamic process of phenotypic plasticity leads in many tumors to a equilibrium with respect to the proportion of cancer cell phenotypes [87]. Cancer cell populations have the ability to maintain this phenotypic equilibrium which has been observed *in vivo* and *in vitro*. In these experiments, subpopulations of cells purified for a given cell phenotype return to the composition of cell state proportions of the original tumor over time. Understanding the phenotypic state transitions would allow to predict the

evolution of cell state proportions. In Chapter 5, we develop a modeling approach and an R software package to estimate phenotypic state transitions from cell line experiments.

STOCHASTIC AND CANCER STEM CELL MODEL. The phenotypic heterogeneity of tumors can be obtained from tissue-scale data by measuring the proportions of phenotypes. In order to understand the emergence of this heterogeneity one has to understand in which way it evolves from the underlying processes on the cellular scale. The cancer stem cell model suggests that tumors are hierarchically organized and follow a unidirectional differentiation process with cancer stem cells (CSCs) at the top of the hierarchy. Only the CSC sub-population has the potential to induce tumor formation due to the ability to self-renew and differentiate [18]. In contrast, the stochastic model postulates that potentially every cell can induce and maintain tumors. In detail, it is proposed that the exposure to certain combinations of intrinsic and extrinsic factors is responsible that any cell can acquire tumorigenic properties [142]. Major challenges in the experimental investigation of these two models are the reliable identification of cancer stem cells, the possibility that both models are not mutually exclusive and that tumors might potentially contain more than one CSC clone [76]. In this thesis, we contribute to a better understanding of the emergence of phenotypic heterogeneity in glioblastoma. We quantify the underlying phenotypic cell state changes of glioblastoma cells from experimental tissue-scale data in Chapter 6.

1.2 SCOPE & OUTLINE OF THIS THESIS

SCOPE. Although the above described cellular processes that are involved in tumor emergence are more and more understood, their specific role in tumor emergence on the *tissue* scale remains elusive. Main reasons for this lack of understanding include the fact that the cellular processes are often hardly observable, especially in the early phase of tumor development, and that the precise interplay between cellular and tissue scale is difficult to infer. Here, mathematical modeling is a valuable tool to evaluate assumptions and to make predictions which can be experimentally tested. A remarkable advantage of this approach is the ability to derive unexpected or even counter-intuitive principles which have not been discovered from a biological perspective.

The design of mathematical models depends on the specific biological question of interest. Complex intra-cellular processes determine the fate of a single cell and the interactions between cells have an impact on phenomena on the tissue scale. In this thesis, we concentrate on the role of cellular mechanisms on tumor-related observables on the tissue scale. Single cell processes like division, death and the acquirement of genetic alterations are highly complex and dependent on many intra-cellular processes. It is therefore not possible to precisely

predict the occurrence of these processes which suggests to utilize stochastic approaches.

In this thesis, we infer the role of genetic and phenotypic cell changes on tumor development with the help of MARKOV chain models which describe cellular processes and are calibrated by tissue-scale data. In particular, we tackle the following specific biological questions within the tumor context:

- In which way do *cellular* processes like competition, *genetic* alterations and proliferation and death determine whether tumors on the *tissue scale* grow, progress or even spontaneously regress?
- At which point of the *cellular* multistep process of cancer and the clonal expansion of *genetically* altered cells is the fate of tumor development decided?
- What is the role of *phenotypic cell state changes* of cancer cells in the establishment of a *tissue-scale* equilibrium with respect to different cell phenotypes in tumors?
- To which extent is the observed phenotypic heterogeneity on the *tissue scale* caused by hierarchical phenotypic cell changes and what is the role of different environmental conditions in determining this heterogeneity?

In the first part of this thesis, we utilize data on the diagnosed fractions of benign and malignant tumor subtypes in order to shed light on the consequences of genetic cell changes on tumor development. The second part of this thesis is devoted to investigate the role of phenotypic plasticity of cancer cells. In particular, we present an approach to estimate the frequencies of phenotypic cell state changes of cancer cells from tissue-scale data on the evolution of phenotypic cell population compositions.

OVERVIEW. In Chapter 2, we introduce the mathematical basics for the models in this thesis. We shortly present the most important definitions and results for MARKOV chains in discrete and continuous time. Moreover, we present the MORAN model and provide a summary of established theoretical results for this model class. We also present a first novel result by introducing a MORAN model on a crypt-like structure and derive the absorption probabilities and absorption times of this model. This finding extends existing mathematical results on MORAN models.

In Chapter 3, we evaluate spontaneous tumor regression in pilocytic astrocytoma (PA) which is the most common brain tumor in children and young adults. This tumor is usually benign and often follows an indolent course. The treatment of choice is resection and the prognosis is very favorable if total excision can be achieved. However, due to the location of the tumor, in many cases only

partial resection is possible. Partially resected PA can spontaneously regress, regrow or even progress to a more aggressive type. We develop a mathematical model which describes the growth, progression and regression of PA. Our analysis of the model allows to quantitatively predict the chance for tumor regression in dependency of the remaining tumor size. This prediction has the potential to provide decision support to clinicians after partial resection of benign PA. Furthermore, our results imply that there is no resection threshold for PA below which no survival advantage is provided. This finding stands in contrast to malignant brain tumors where such a threshold has been experimentally shown.

In Chapter 4, we investigate the hardly observable early phase of tumor development on the cellular scale. Cancers exhibit different types of malignant progression, namely with and without clinically detectable benign precursor stages. Clinical data shows that progression patterns, i.e. the frequencies of the two progression types, are tissue-specific but the role of the cellular multistep process of cancer development therein remains unclear. We introduce a model which is based on competition between wild-type and tumor cells and assume that a certain amount of tumor cells is needed for tumor emergence. We estimate this number by fitting the model to data on the diagnosed ratios of benign and malignant tumor subtypes. Our model predicts that this number is considerably small compared to the overall tumor mass and largely depends on the tissue type. Our results can be interpreted as existence of a tissue-specific tumor-originating niche in which the fate of tumor development is decided long before a tumor becomes detectable.

The second part of this thesis is devoted to the investigation of the influence of phenotypic plasticity of tumor cells in tumor development. In Chapter 5, we develop CELLTRANS, an R package to quantify stochastic cell state transitions from cell state proportion data from fluorescence-activated cell sorting (FACS) and flow cytometry experiments. The package is based on a mathematical model in which cell state alterations occur due to stochastic transitions between distinct cell states whose rates only depend on the current state of a cell. CELLTRANS is an automated tool for estimating the underlying transition probabilities from appropriately prepared data. We point out potential analytical challenges in the quantification of these cell transitions and explain how CELLTRANS handles them. The applicability of CELLTRANS is demonstrated on publicly available data on the evolution of cell state compositions in cancer cell lines.

In Chapter 6, we investigate the observed phenotypic heterogeneity in a certain brain tumor called glioblastoma (GBM) with particular emphasis on the question whether this heterogeneity is maintained by a reversible or an irreversible process on the cellular scale. For this purpose, we utilize data on the temporal evolution of GBM cell line compositions from FACS experiments under both normoxic and hypoxic conditions. We utilize CELLTRANS in order to quantify the state transitions in time between 16 different cancer cell phenotypes in GBM which are defined on the basis of potential stemness markers. More-

chapter	# cell types	competition	genetic changes	space	phenotypic changes
2	2	yes	no	yes	no
3	3	yes	yes	no	no
4	3	yes	yes	yes	no
5	$n \in \mathbb{N}$	no	no	no	yes
6	16	no	no	no	yes

Table 1: Overview of the components in the cell-based MARKOV chain models in this thesis.

over, we investigate the influence of different environmental conditions on the estimates and enquire whether the cell state transitions are dependent on the current marker expression. Table 1 provides an overview of the components in the cell-based MARKOV chain models in this thesis.

Finally, Chapter 7 provides a summary of the main results in this thesis. Furthermore, we discuss possible extensions of the modeling approaches that have been developed in this thesis and give an outlook for further research.

2

MATHEMATICAL AND MODELING FOUNDATIONS

Contents

2.1	MARKOV chains	9
2.1.1	MARKOV chains in discrete time	9
2.1.2	MARKOV chains in continuous time	15
2.2	The MORAN model	17
2.3	A MORAN model on a crypt-like structure	25
2.3.1	A crypt-like structure and state space of the model	25
2.3.2	Model dynamics & analysis	28
2.3.3	Derivation of the absorption probabilities and absorption times	30
2.3.4	Comparison with MORAN models on lattice structures	33

2.1 MARKOV CHAINS

In this section, we will introduce the basic theory of MARKOV chains on finite or countable state spaces based on the books [13] and [5].

2.1.1 MARKOV chains in discrete time

2.1 Definition (Time-homogeneous MARKOV chain in discrete time) Let $(X_n)_{n \in \mathbb{N}_0}$ be a discrete time stochastic process with countable state space S . Then, $(X_n)_{n \in \mathbb{N}_0}$ is called a MARKOV chain if for all $n \in \mathbb{N}_0$ and states $i_0, \dots, i_{n-1}, i, j \in S$ it holds that

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i) \quad (2.1)$$

whenever both sides are well-defined. A MARKOV chain is called time-homogenous if for all $n \in \mathbb{N}$ and $i, j \in S$

$$\mathbb{P}(X_{n+1} = j \mid X_n = i) = \mathbb{P}(X_n = j \mid X_{n-1} = i),$$

i.e. the probability of a transition from state i to j is independent of the time step $n \in \mathbb{N}$. Moreover,

$$p(i, j) := \mathbb{P}(X_{n+1} = j \mid X_n = i)$$

defines the transition probability from state i to j with $i, j \in S, n \in \mathbb{N}_0$.

The random variable X_0 is called initial state and its distribution w ,

$$w(i) = \mathbb{P}(X_0 = i), \quad i \in S,$$

is called initial distribution.

The property (2.1) is called MARKOV property. When the state space is finite we speak of the transition matrix $P = (p(i, j))_{i, j \in S}$. As we will consider only time-homogeneous MARKOV chains in this thesis, we do not use the phrase time-homogeneous and refer to the processes simply as MARKOV chains. Note that the distribution of a discrete time MARKOV chain is determined by its initial distribution and its transition probabilities.

CLASSIFICATION OF STATES. For a discrete time MARKOV chain, we say that there is a possible path from state i to state j if there is a sequence of states $i = i_0 \rightarrow i_1 \rightarrow \dots \rightarrow i_n = j$ such that for all transitions along the path one has $p(i_{l-1}, i_l) > 0, l = 1, \dots, n$. We will also use the phrase that state j is accessible from state i . We say that two states $i, j \in S$ communicate if there is a possible path from i to j and from j to i . A communicating class is a maximal set of states C such that every pair of states in C communicates with each other. A set of states C is called closed if $\sum_{j \in C} p(i, j) = 1$ for all $i \in C$. A state $i \in S$ is absorbing if $p(i, i) = 1$, i.e. the single-element set $\{i\}$ is a closed communicating class. Moreover, a MARKOV chain is said to be irreducible if there is only one communication class, i.e. if it is possible to get from any state to any other state.

The return time for a state $i \in S$ is defined by

$$T_i = \inf\{n > 0 \mid X_n = i\}$$

with $T_i = \infty$ if $X_n \neq i$ for all $n \geq 1$ and allows to introduce the concepts of recurrent and transient states.

2.2 Definition (Recurrence and transience) For a discrete time MARKOV chain on S a state $i \in S$ is called recurrent if

$$\mathbb{P}(T_i < \infty \mid X_0 = i) = 1,$$

and is called transient otherwise. A recurrent state $i \in S$ is called positive recurrent if

$$\mathbb{E}(T_i \mid X_0 = i) < \infty,$$

and otherwise it is called null recurrent.

Hence, a state is recurrent if and only if the probability of ever returning to it is one given that the MARKOV chain starts in that state. It can be shown that the states within a communication class are either all positive recurrent, null recurrent or transient [13, Chapter 2]. All states of an irreducible MARKOV chain are therefore either positive recurrent, null recurrent or transient and we call it a positive recurrent, null recurrent or transient chain, respectively.

LIMITING BEHAVIOR. The period of a state is an important property to investigate the long-term behavior of a MARKOV chain.

2.3 Definition (Period) For a discrete time MARKOV chain $(X_n)_{n \in \mathbb{N}}$ on state space S a possible loop of length n is a sequence of states $i_0, i_1, \dots, i_n \in S$ with $i_0 = i_n$ and

$$p(i_0, i_1) \cdot p(i_1, i_2) \cdot \dots \cdot p(i_{n-1}, i_n) > 0.$$

With

$$D_i = \{n \in \mathbb{N} \mid \text{there exists a possible loop of length } n \text{ with } i_0 = i_n = i\}$$

the period of state i is defined as the largest number dividing all numbers in D_i with the convention that $D_i = \infty$ if D_i is empty. A state with period 1 is called aperiodic.

Importantly, all states within a communication class have the same period [13, Chapter 2]. Hence, an irreducible, aperiodic MARKOV chain consists of a single communication class of period 1.

For aperiodic, recurrent MARKOV chains the long-term behavior does not depend on the initial distribution.

2.4 Theorem For an irreducible, recurrent and aperiodic discrete time MARKOV chain $(X_n)_{n \in \mathbb{N}}$ on S , any state $i \in S$ and any initial distribution it holds that

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = j) = \frac{1}{\mathbb{E}(T_j \mid X_0 = j)}.$$

The right-hand side is thereby defined as 0 if $\mathbb{E}(T_j \mid X_0 = j) = \infty$.

Proof. The proof can be found in [5, Chapter 1.2]. □

2.5 Definition (Invariant measure) A nontrivial vector $\pi = (\pi(j))_{j \in S}$ is called an invariant measure for the transition probabilities $P = (p(i, j))_{i, j \in S}$ of a MARKOV chain on S if

$$\pi(j) = \sum_{l \in S} \pi(l) p(l, j).$$

If π is a probability vector (i.e. all entries non-negative which sum to one) we call it invariant distribution for P .

The following result states that for irreducible and recurrent MARKOV chains there is a unique invariant measure.

2.6 Theorem (Existence and uniqueness of invariant measures) For an irreducible, recurrent MARKOV chain $(X_n)_{n \in \mathbb{N}}$ on S there is an (up to multiplication) unique invariant measure which can be normalized to a unique invariant distribution if and only if $\mathbb{E}(T_i | X_0 = i) < \infty$ for all $i \in S$, i.e. if the MARKOV chain is positive recurrent.

Proof. The proof can be found in [13, Chapter 3]. □

Note that irreducible MARKOV chains with a finite state space are positive recurrent [13] and thereby exhibit unique invariant distributions.

The following statements complete the characterization of the limiting behavior of discrete time MARKOV chains needed for the scope of this thesis.

2.7 Theorem (a) For an irreducible, aperiodic and positive recurrent MARKOV chain $(X_n)_{n \in \mathbb{N}}$ on S it holds for $j \in S$ that

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = j) = \pi(j) = \frac{1}{\mathbb{E}(T_j | X_0 = j)},$$

where $\pi = (\pi(j))_{j \in S}$ denotes the unique invariant distribution.

(b) For a null recurrent state j of a MARKOV chain with any initial distribution it holds that

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = j) = 0.$$

(c) For a transient state j of a MARKOV chain with any initial distribution it holds that

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n = j) = 0.$$

Proof. For the proof of part (a), see [13, Chapter 3] and for the proofs of parts (b) and (c), see [5, Chapter 1]. □

For the characterization of the limiting behavior of irreducible and periodic MARKOV chains, see for example [13, Chapter 4].

ABSORPTION PROBABILITIES AND TIMES. Consider a MARKOV chain on state space S with transition probabilities $P = (p(i, j))_{i, j \in S}$. We define the hitting time of a set of states $A \subseteq S$ by

$$H^A := \inf\{n \geq 0 : X_n \in A\}.$$

The hitting probability of ever reaching a state in A starting from state $i \in S$ is defined as

$$\alpha^A(i) := \mathbb{P}\left(H^A < \infty \mid X_0 = i\right)$$

and is also called absorption probability if A is closed. A general result regarding the calculation of the hitting probabilities of MARKOV chains and a corresponding proof presented in [99] follows.

2.8 Theorem The vector of hitting probabilities $\alpha^A = (\alpha^A(i))_{i \in S}$ of a MARKOV chain on state space S reaching a set of states A satisfies

$$\begin{aligned} \alpha^A(i) &= 1, \quad i \in A, \\ \alpha^A(i) &= \sum_{j \in S} p(i, j) \alpha^A(j), \quad i \notin A. \end{aligned}$$

If $\beta(i)$ is another solution of these equations, then $\beta(i) \geq \alpha^A(i)$.

Proof. Clearly, if $i \in A$, then $H^A = 0$ and therefore $\alpha^A(i) = 1$. If $i \notin A$, then $H^A \geq 1$ and $H^A = 1 + \hat{H}^A$ where \hat{H}^A denotes the remaining time until A is reached. By conditioning on the first step of the chain

$$\mathbb{P}\left(H^A < \infty \mid X_0 = i\right) = \sum_{j \in S} \mathbb{P}\left(\hat{H}^A < \infty \mid X_1 = j\right) p(i, j).$$

Moreover, it holds that

$$\mathbb{P}\left(\hat{H}^A < \infty \mid X_1 = j, X_0 = i\right) = \mathbb{P}\left(\hat{H}^A < \infty \mid X_1 = j\right)$$

since \hat{H}^A only depends on $X_k, k \geq 1$, and also

$$\mathbb{P}\left(\hat{H}^A < \infty \mid X_1 = j\right) = \mathbb{P}\left(H^A < \infty \mid X_0 = j\right) = \alpha^A(j)$$

because the chain is homogeneous.

Using the above, one obtains

$$\mathbb{P}\left(H^A < \infty \mid X_0 = i\right) = \alpha^A(i) = \sum_{j \in S} p(i, j) \alpha^A(j).$$

Let now $\beta(i)$ be another solution. Then

$$\beta(i) = \sum_{j \in A} p(i, j) + \sum_{j \notin A} p(i, j) \beta(j).$$

Hence,

$$\begin{aligned}
\beta(i) &= \sum_{j \in A} p(i, j) + \sum_{j \notin A} p(i, j) \left(\sum_{k \in A} p(j, k) + \sum_{k \notin A} p(j, k) \beta(k) \right) \\
&= \sum_{j \in A} p(i, j) + \sum_{j \notin A} p(i, j) \sum_{k \in A} p(j, k) + \sum_{j \notin A} p(i, j) \sum_{k \notin A} p(j, k) \beta(k) \\
&= \mathbb{P} \left(H^A = 1 \mid X_0 = i \right) + \mathbb{P} \left(H^A = 2 \mid X_0 = i \right) + \sum_{j \notin A} p(i, j) \sum_{k \notin A} p(j, k) \beta(k)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\beta(i) &\geq \mathbb{P} \left(H^A = 1 \mid X_0 = i \right) + \mathbb{P} \left(H^A = 2 \mid X_0 = i \right) \\
&= \mathbb{P} \left(H^A \leq 2 \mid X_0 = i \right).
\end{aligned}$$

Repeating the substitution leads to $\beta(i) \geq \mathbb{P} \left(H^A \leq n \mid X_0 = i \right)$ for all $n \geq 0$. By letting $n \rightarrow \infty$, one obtains

$$\beta(i) \geq \mathbb{P} \left(H^A \leq \infty \mid X_0 = i \right) = \alpha^A(i).$$

□

The mean number of steps until a set of states A is reached can be described by the mean hitting time

$$\tau^A(i) := \mathbb{E} \left(H^A \mid X_0 = i \right),$$

which is also referred to as absorption time if A is closed. It can be derived in the following way.

2.9 Theorem The vector of mean hitting times $\tau^A = (\tau^A(i))_{i \in S}$ of a MARKOV chain on state space S reaching a set of states A satisfies

$$\begin{aligned}
\tau^A(i) &= 0, \quad i \in A, \\
\tau^A(i) &= 1 + \sum_{j \notin A} p(i, j) \tau^A(j), \quad i \notin A.
\end{aligned}$$

If $\beta(i)$ is another solution of these equations, then $\beta(i) \geq \tau^A(i)$.

Proof. If the starting state $i \in A$, then $H^A = 0$ and $\tau^A(i) = 0$. If $i \notin A$, then $H^A = 1 + \widehat{H}^A$ as in the proof of Theorem 2.8. By conditioning on the first step of the chain

$$\begin{aligned}
\mathbb{E} \left(H^A \mid X_0 = i \right) &= 1 + \mathbb{E} \left(\widehat{H}^A \mid X_0 = i \right) = 1 + \sum_{j \in S} \mathbb{E} \left(H^A \mid X_0 = j \right) p(i, j) \\
&= 1 + \sum_{j \notin A} p(i, j) \tau^A(j).
\end{aligned}$$

The second part of the proof is provided in [99, Chapter 1.3].

□

In the first part of the proofs of Theorems 2.8 and 2.9 conditioning on the first step of the MARKOV chain leads to necessary conditions for the hitting probabilities and times given by a system of linear equations. This technique is called first step analysis and regularly used in the analysis of discrete time MARKOV chains. The resulting systems of linear equations in Theorems 2.8 and 2.9 are also sufficient conditions for the hitting probabilities if the state space S is finite. Then, the systems of linear equations exhibit a unique solution which is not the general case with an infinite state space [13].

2.1.2 MARKOV chains in continuous time

In this subsection, we define and summarize the most important properties of MARKOV chains in continuous time on a countable state space S without proofs. For a comprehensive introduction and the corresponding proofs, see e.g. [13, Chapter 8].

2.10 Definition (Time-homogeneous MARKOV chain in continuous time) Let $(X_t)_{t \geq 0}$ be a stochastic process in continuous time with countable state space S . Then, $(X_t)_{t \geq 0}$ is called continuous time MARKOV chain if for every $k \in \mathbb{N}$, all states $i, j, i_1, \dots, i_k \in S$, all times $t, s \geq 0$ and $s_1, \dots, s_k \geq 0$ with $s_l \leq s$ for all $l \in \{1, \dots, k\}$ it holds that

$$\mathbb{P}(X_{t+s} = j \mid X_s = i, X_{s_1} = i_1, \dots, X_{s_k} = i_k) = \mathbb{P}(X_{t+s} = j \mid X_s = i) \quad (2.2)$$

whenever both sides are well-defined. The continuous time MARKOV chain is called time-homogenous if the right-hand side of (2.2) is independent of s .

Moreover, the family $\{P(t)\}_{t \geq 0}$ given by

$$P(t) = \{p(i, j, t)\}_{i, j \in S}$$

with

$$p(i, j, t) = \mathbb{P}(X_{t+s} = j \mid X_s = i)$$

is called transition semigroup of the time-homogenous MARKOV chain in continuous time.

Since we only regard time-homogenous MARKOV chains, we will omit the phrase time-homogenous in the following.

The distribution $\mu(t) = (\mu_i(t))_{i \in S}$ with $\mu_i(t) = \mathbb{P}(X_t = i)$ at time t can be obtained from the initial distribution $\mu(0)$ by

$$\mu(t)^\top = \mu(0)^\top P(t).$$

A continuous time MARKOV chain is determined by its initial distribution and the transition semigroup. We only consider continuous time MARKOV chains

with corresponding transition semigroups that are right continuous in the origin, i.e. $\lim_{h \rightarrow 0+} p(i, j, h) = p(i, j, 0)$ for all $i, j \in S$ or $\lim_{h \rightarrow 0+} P(h) = I$, where I denotes the identity matrix. Note that the continuity at the origin implies continuity at any $t \geq 0$.

For any state $i \in S$ there exists

$$q_i := \lim_{h \rightarrow 0+} \frac{1 - p(i, i, h)}{h} \in [0, \infty]$$

and for all $i, j \in S$ with $i \neq j$

$$q(i, j) := \lim_{h \rightarrow 0+} \frac{p(i, j, h)}{h} \in [0, \infty).$$

With $q(i, i) := -q_i$ for every $i \in S$,

$$Q = (q(i, j))_{i, j \in S}$$

is called the generator of the continuous time MARKOV chain. The generator can be regarded as the derivative at 0 of the matrix function $t \mapsto P(t)$, i.e.

$$Q = \lim_{h \rightarrow 0+} \frac{P(h) - P(0)}{h}.$$

A state is called absorbing if $q(i, i) = 0$ which is equivalent to $p(i, i, t) = 1$ for all $t \geq 0$. The generator is called stable if $-q(i, i) < \infty$ for all $i \in S$ and is called conservative if

$$-q(i, i) = \sum_{\substack{i \in S \\ i \neq j}} q(i, j)$$

for all $i \in S$.

The jump times $(\tau_n)_{n \in \mathbb{N}_0}$ of a continuous time MARKOV chain are defined by

$$\tau_0 = 0, \quad \tau_{n+1} = \inf\{t \geq \tau_n : X_t \neq X_{\tau_n}\}, \quad n \in \mathbb{N}.$$

We only regard non-explosive processes, i.e. for which

$$\tau_\infty := \sup_n \tau_n = \infty \quad \text{a.s.}$$

meaning that there are not infinitely many jumps in finite time. This holds in particular for a continuous time MARKOV chain on a finite state space.

We conclude this chapter by introducing the embedded chain for MARKOV processes with stable and conservative generators.

2.11 Theorem (Embedded MARKOV chain) For a continuous time MARKOV chain on S with a stable and conservative generator $Q = (q(i, j))_{i, j \in S}$ the sequence

$(Y_n)_{n \in \mathbb{N}_0}$ of visited states is a discrete time MARKOV chain on S called embedded chain with transition probabilities

$$p(i, j) = \begin{cases} -\frac{q(i, j)}{q(i, i)}, & i \in S \setminus A, j \neq i \\ 0, & i \in S \setminus A, j = i \\ 0, & i \in A, j \neq i \\ 1, & i \in A, j = i \end{cases}$$

where $A = \{i \in S \mid q(i, i) = 0\}$ contains the absorbing states.

2.2 THE MORAN MODEL

In the first part of this thesis, we will utilize extensions of the so-called MORAN model in order to describe the cellular processes involved in tumor development. MORAN models are MARKOV chains that can be formulated in both discrete and continuous time. In this section, we introduce the classical MORAN model and corresponding extensions. Moreover, we provide an overview of applications of this model class within the tumor context.

THE CLASSICAL MORAN MODEL. The MORAN model was introduced by P. A. P. Moran in 1958 [98] and describes the evolution of two competing finite populations. The population size is constant of size N and each individual is either of type A or type B . In the context of cancer modeling, the different types can represent healthy and cancer cell populations. At each time step, an individual is uniformly chosen at random to reproduce and gives birth to a single offspring. Whenever such a birth occurs, another individual, including the one giving birth, is uniformly chosen to be replaced by the new offspring. The process is visualized in Figure 1.

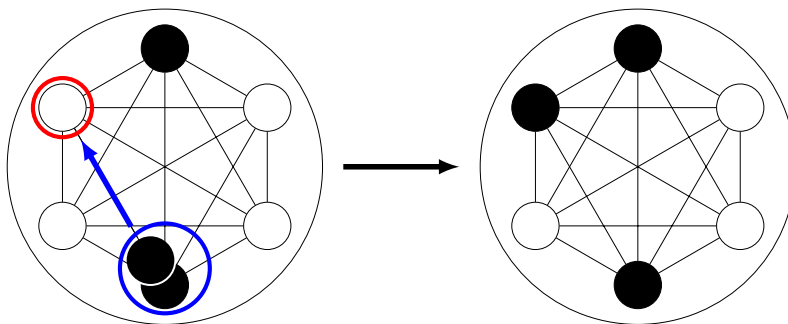


Figure 1: The classical MORAN model. At each time step, an individual is uniformly chosen at random to reproduce (blue circle). The offspring of this individual then replaces a uniformly chosen other individual (red circle).

Several versions of this model have been proposed including continuous time versions in which individuals reproduce with constant rates or in which individuals cannot both give birth and die in a single step.

Mathematically, the classical MORAN model is defined as follows.

2.12 Definition The classical MORAN model is a discrete time MARKOV chain $(X_n)_{n \in \mathbb{N}_0}$ with state space $S = \{0, 1, \dots, N\}$ in which X_n describes the number of individuals of type B at time n . The transition matrix is thereby given by $P = (p(k, l))_{k, l \in S}$ with

$$p(k, l) = \begin{cases} \frac{N-k}{N} \frac{k}{N}, & 1 \leq k \leq N-1, l = k-1, \\ \frac{k^2}{N^2} + \frac{(N-k)^2}{N^2}, & 0 \leq k \leq N, l = k, \\ \frac{k}{N} \frac{N-k}{N}, & 1 \leq k \leq N-1, l = k+1, \\ 0, & \text{else.} \end{cases} \quad (2.3)$$

The probability of a one-step transition from state k to $k+1$, $1 \leq k \leq N-1$, in the classical MORAN model is composed as follows. The probability that an individual of type B is chosen to reproduce is $\frac{k}{N}$ since there are k individuals of type B and one is uniformly chosen. The offspring of this individual replaces an individual of type A with probability $\frac{N-k}{N}$ because there are $N-k$ type A individuals present. The product of these probabilities yields the corresponding entry in equation (2.3).

This process describes a neutrally evolving population since the probabilities for an increase and a decrease of the state are equal. The states 0 and N are absorbing states of the process while the states $1, \dots, N-1$ are transient. Since the state of the process changes by at most one in a time step, the MORAN model can be regarded as generalized birth-death process. Usually, the length of one time step is set to $\frac{1}{N}$ such that during a unit time interval N individuals are chosen for reproduction and N are chosen to be replaced. This choice allows to interpret the number of time steps as generations of the population.

An important characteristics in the dynamics of the MORAN model is the absorption probability in states 0 and N describing the fixation of one of the individuals within the population. Moreover, the absorption time in an absorbing state allows to deduce the time-scale of fixation. Formally, these quantities can be described as follows. Let

$$\tau = \min\{n : X_n \in \{0, N\}\}$$

be the time of fixation of one of the individuals. Then, the absorption probability in state N when starting the process in state $i \in S$ can be described by

$$\alpha^N(i) = \mathbb{P}(X_\tau = N | X_0 = i).$$

The following result states that the fixation probability in state N is equal to the initial frequency of type B individuals, see also [100].

2.13 Theorem In the classical MORAN model, the fixation probability of the type B population equals the initial frequency of type B , i.e.

$$\alpha^N(i) = \frac{i}{N}. \quad (2.4)$$

Proof. The proof utilizes *first step analysis* which is a standard technique to analyze the absorption behavior of MARKOV chains, see [112]. Note that we already introduced this technique in the proofs of Theorems 2.8 and 2.9. Applying first step analysis here yields for $j = 1, \dots, N - 1$

$$\alpha^N(j) = \frac{1}{2}\alpha^N(j+1) + \frac{1}{2}\alpha^N(j-1)$$

which is equivalent to

$$\alpha^N(j+1) - \alpha^N(j) = \alpha^N(j) - \alpha^N(j-1).$$

Since $\alpha^N(0) = 0$

$$\alpha^N(j+1) - \alpha^N(j) = \alpha^N(1).$$

Summation for $j = 0, \dots, i - 1$ yields

$$\alpha^N(i) = \sum_{j=0}^{i-1} \alpha^N(1) = i\alpha^N(1).$$

Finally, for $i = N$ and since $\alpha^N(N) = 1$ it follows

$$\alpha^N(N) = N\alpha^N(1) = 1, \quad \text{hence } \alpha^N(1) = \frac{1}{N} \text{ and } \alpha^N(i) = \frac{i}{N}.$$

□

This result reflects that the stochastic process $(X_n)_{n \in \mathbb{N}_0}$ describing the classical MORAN model is a symmetric random walk.

Regarding the absorption times, the following results hold.

2.14 Theorem The expected absorption time measured in units of generations in either absorbing state in the classical MORAN model when starting with i type B individuals is approximately

$$\mathbb{E}(\tau | X_0 = i) \approx -N \left(\frac{i}{N} \ln \left(\frac{i}{N} \right) + \left(1 - \frac{i}{N} \right) \ln \left(1 - \frac{i}{N} \right) \right)$$

Moreover, the expected absorption time conditioned on absorption in state N can be approximated by

$$\mathbb{E}(\tau | X_0 = i, X_\tau = N) \approx -\frac{N \left(1 - \frac{i}{N}\right)}{\frac{i}{N}} \ln \left(1 - \frac{i}{N}\right).$$

The derivation of these results is provided in [24, chapter 1.5.1].

SELECTION AND MUTATIONS. The classical MORAN model describes the evolution of neutrally evolving populations, i.e. none of the individuals exhibits a reproductive advantage. Considering a cancer context, malignant cells usually have an advantage at some stage of tumor progression. Moreover, the important process of genetic alterations which leads to the progression of individual cells is not taken into account within the classical model. Therefore, it is reasonable to extend the classical MORAN model with *natural selection* and *mutations*.

From a modeling perspective, natural selection can be described in different ways. Here, we will introduce the approach from [145, chapter 9]. The individuals of type A have fitness 1 and the individuals of type B fitness $r > 0$. As in the classical MORAN model, one individual is uniformly chosen to be replaced in each time step. The relative chance for being chosen for reproduction is 1 for type A individuals and r for type B individuals. Note that the choice of the fitness parameter r allows to incorporate advantageous ($r > 1$), disadvantageous ($r < 1$) and neutral ($r = 1$) type B individuals.

Mutations can be described within the framework of the MORAN model as follows. Whenever an individual is chosen for reproduction, there is some constant probability that the offspring is different from the original individual. In the simplest case, only mutations from type A to type B individuals are possible. In each reproduction event of a type A individual, the offspring mutates to a type B individual with probability u .

2.15 Definition The MORAN model with selection parameter $r > 0$ and mutation parameter $u, 0 < u < 1$, is a discrete time MARKOV chain $(X_n)_{n \in \mathbb{N}_0}$ with state space $S = \{0, 1, \dots, N\}$ and transition matrix $P = (p(k, l))_{k, l \in S}$ given by

$$p(k, l) = \begin{cases} \frac{(1-u)(N-k)}{N-k+rk} \frac{k}{N}, & 1 \leq k \leq N-1, l = k-1, \\ 1 - p(k, k-1) - p(k, k+1), & 1 \leq k \leq N-1, l = k, \\ \frac{u(N-k)+rk}{N-k+rk} \frac{N-k}{N}, & 0 \leq k \leq N-1, l = k+1, \\ 1 - u, & k = l = 0, \\ 1, & k = l = N, \\ 0, & \text{else.} \end{cases} \quad (2.5)$$

The probability $p(k, k + 1)$ for an increase of the state in the MORAN model with selection and mutations is thereby composed as follows. There are two possibilities for such an increase. First, the offspring of a type B individual can replace an individual of type A . Since there are $N - k$ individuals of type A within the population, the probability that one of them is chosen for replacement is $\frac{N-k}{N}$. The probability that an individual of type B reproduces is proportional to its frequency within the population and its relative fitness, i.e.

$$\frac{rk}{N - k + rk}.$$

Hence, the probability of this first case is

$$\frac{rk}{N - k + rk} \frac{N - k}{N}.$$

Second, the state can also increase by a reproduction of a type A individual with subsequent mutation of the offspring to a type B individual which replaces an individual of type A . The probability for this event is given by

$$\frac{(N - k)u}{N - k + rk} \frac{N - k}{N}.$$

The sum of the probabilities for both possibilities yields the entry within the transition matrix (2.5). The probability $p(k, k - 1)$ is obtained similarly. In contrast to the classical MORAN model, the MORAN model with selection and mutations has only one absorbing state N , i.e. fixation of type B individuals will eventually occur.

From an application point of view, the processes mutations and selection can often be assumed to take place on different time-scales. In general, the mutation probabilities are very small. Once a single type B individual appears due to a mutation, it typically dies out or reaches fixation *before* a new mutation occurs. Therefore, it is often justified to treat both processes separately, i.e. for $X_n > 0$ one can argue that it is reasonable to assume $u = 0$. In this special case, one obtains the MORAN model with selection (and without mutations) and the probability of fixation in state N becomes a quantity of particular interest.

2.16 Theorem In the MORAN model with selection (and without mutations), the fixation probability of the type B population with fitness parameter $r > 0, r \neq 1$, and starting the process with i type B individuals is given by

$$\alpha^N(i) = \frac{1 - \left(\frac{1}{r}\right)^i}{1 - \left(\frac{1}{r}\right)^N}. \quad (2.6)$$

Proof. The proof is analogous to the proof of Theorem 2.4 and also utilizes first step analysis. However, since we will need to apply this technique later in this thesis again, we demonstrate it once more.

First step analysis allows to derive the system of equations

$$\alpha^N(j) = \frac{p(j, j+1)}{p(j, j-1) + p(j, j+1)} \alpha^N(j+1) + \frac{p(j, j-1)}{p(j, j-1) + p(j, j+1)} \alpha^N(j-1)$$

with $1 \leq j \leq N-1$ and $u = 0$ in the transition matrix (2.5).

Multiplying by $p(j, j-1) + p(j, j+1)$ and rearranging yields

$$\alpha^N(j+1) - \alpha^N(j) = \frac{p(j, j-1)}{p(j, j+1)} (\alpha^N(j) - \alpha^N(j-1)) = \frac{1}{r} (\alpha^N(j) - \alpha^N(j-1))$$

Using $\alpha^N(0) = 0$ and iterating implies

$$\alpha^N(j+1) - \alpha^N(j) = \alpha^N(1) \left(\frac{1}{r}\right)^j.$$

Summation for $j = 0, \dots, i-1$ yields

$$\alpha^N(i) = \sum_{j=0}^{i-1} \alpha^N(1) \left(\frac{1}{r}\right)^j = \alpha^N(1) \frac{\left(\frac{1}{r}\right)^i - 1}{\frac{1}{r} - 1}.$$

Since $\alpha^N(N) = 1$ it follows

$$\alpha^N(1) = \frac{1 - \frac{1}{r}}{1 - \left(\frac{1}{r}\right)^N}, \text{ hence } \alpha^N(i) = \frac{1 - \left(\frac{1}{r}\right)^i}{1 - \left(\frac{1}{r}\right)^N}.$$

□

Note that one obtains the fixation probability of the classical MORAN model by letting $r \rightarrow 1$ with L'HOSPITAL'S rule .

MULTI-TYPE MORAN MODEL. So far, we have only introduced MORAN models with two individuals. Next, we will consider a population of individuals that can be of three different types. For the purpose of this thesis, these individuals represent three different cell types within the cancer context, namely

- *wild-type* cells, which represent healthy cells and have fitness 1,
- *type-1* cells, which are cancer cells with fitness r having a specific genetic alteration that is acquired from wild-type cells at division with mutation probability u ,
- *type-2* cells, which are cancer cells with fitness R emerging by progression from type-1 cells at division with mutation probability v .

In applications it is sensible to assume that the mutations are unidirectional, i.e. there are no back-mutations. Within this setting, type-1 cells can represent benign tumor cells which occur by mutations from wild-type cells and that might have no selective advantage ($r = 1$) or might have a selective advantage ($r > 1$) compared to wild-type cells. Wild-type and type-1 cells evolve by the transition rates given by equation (2.5). Progression of type-1 cells can lead to malignant type-2 cells. Within the multi-type MORAN model, cancer development can be described as occurrence of a type-2 cell which inevitably leads to cancer due to a very large fitness advantage R .

One question in this multi-type model is the distribution of the random variable τ_2 which describes the time when the first type-2 individual occurs. A comprehensive derivation of this distribution is provided in [100, Chapter 12]. For the purpose of this thesis, it is sufficient to describe the different scenarios that can lead to the occurrence of the first type-2 individual. These scenarios will play an important role in this thesis to describe the occurrence of benign and malignant tumor subtypes in Chapter 3 and to represent different paths of tumor progression in Chapter 4.

If we assume that type-1 individuals are neutral, i.e. $r = 1$, the following scenarios in dependency of the three parameters population size N and mutation probabilities u and v can be distinguished.

- *Sequential fixation:*

$$N \ll \sqrt{v}$$

In this case, type-1 individuals are typically able to fixate within the system before a type-2 individual occurs. In other words, the fixation of the mutant occurs sequentially: first, the type-1 individuals fixate, and subsequently type-2 individuals occur and fixate. Note that, in general, several type-1 individuals will die out but eventually one successful type-1 mutant will emerge which reaches fixation.

- *Stochastic tunneling:*

$$\frac{1}{\sqrt{v}} \ll N \ll \frac{1}{u}$$

In this case, a type-2 individual will occur due to a mutation of a type-1 individual before the family of type-1 individuals reaches fixation. Hence, the type-2 individuals reach fixation without prior fixation of the type-1 individuals. The term tunneling in this context was first introduced in [95].

- *Quasi-deterministic scenario:*

$$N \gg \frac{1}{\sqrt{u}}$$

In this scenario, type-1 cells arise immediately and the number of type-1 cells grows linearly with time.

Here, $a(N) \ll b(N)$ means that $\lim_{N \rightarrow \infty} \frac{a(N)}{b(N)} \rightarrow 0$, i.e. the above parameter regimes provide an asymptotic classification of the model behavior. For rigorous proofs of these results see [100, Chapter 12].

SPATIAL MORAN MODELS. The models introduced so far can be regarded as space-free models in which no neighborhood relations between the individuals are considered. Hence, the dynamics allows that each individual can be replaced by the offspring of any other individual which might be valid if one considers a homogeneously mixing population.

In the cancer context, this assumption might be valid in liquid tumors, e.g. in the modeling of leukemia. However, solid tissues exhibit characteristic neighborhood relations between cells. Formally, these neighborhood relations can be described by a graph in which the vertices describe individual cells and the edges indicate an adjacency relation between the corresponding cells. Here, one can choose regular lattice structures in 1D or 2D but also non-regular structures can be utilized. For example, cancer development in tube-like tissues like the mammary ducts of the breast, or renal tubules could be viewed as approximately one-dimensional structure due to the aspect ratio of tube radius vs. length. On the other hand, the squamous epithelium of the cervix, the bladder or the oral cavity can be viewed as two-dimensional lattice [36]. Note that the neighborhood relations between the individuals in the classical MORAN model can be formally described by a complete graph.

The first theoretical results of a spatial MORAN model with selection have been introduced in [70]. In this work, a one-dimensional neighborhood relation between the cells is considered and results concerning the fixation probability of a single cell with fitness r compared to fitness 1 of wild-type cells are presented. The main result in this work is that advantageous and disadvantageous mutants, i.e. mutants with a fitness $r \neq 1$, have a smaller fixation probability compared to the space-free MORAN model. This first approach to take spatial cell neighborhood relations into account illustrates the important role of spatial aspects.

Subsequent works consider three-type MORAN models with mutations on d -dimensional lattices. In [26], there are three types of cells (0, 1, and 2) all with fitness 1 which compete with each other by MORAN dynamics, and type i cells mutate to type $i + 1$ cells at rate $u_i, i = 1, 2, 3$. Under these assumptions, the waiting time to create a cell of type 2 is characterized. Interestingly, it turns out that only in the one-dimensional case space has a substantial influence on the waiting time [26]. Moreover, analogous results for three-type MORAN models on d -dimensional lattices have been derived in the case in which mutants have a selective advantage [25].

APPLICATIONS OF MORAN MODELS IN THE CANCER CONTEXT. Although the vast majority of the literature investigates theoretical aspects of the MORAN

model and its extensions, there are also specific applications to questions within the cancer context. For example, in [58, 102] the population genetics of tumor suppressor genes is investigated with a three-type MORAN model. In detail, the probability that a single cell with two inactivated alleles of a tumor suppressor gene has arisen by time t in a population of N cells is derived in these works. In [131], two spatial extensions of the MORAN model are introduced to study the effects of motility on the ability of a new mutant cell to invade a pre-existing tumor. Summarized, the results in [131] indicate that a mobility phenotype can be dominant and able to invade a background of cells with higher reproduction rates suggesting that migration is a key modulator of fitness.

Tomasetti et al. [132] used a diffusion approximation of the MORAN process to calculate the expected number of passenger mutations that accumulate in the precancer phase. They found that at least half of the somatic mutations in tumors of self-renewing tissues occur before the onset of neoplasia by using DNA sequencing data.

In [36], the authors investigate how various properties of a potential premalignant field, or cancer field, depend on microscopic cellular properties of the tissue. For this purpose, a spatial MORAN model is considered in which cells can acquire successive random mutations which confer selective advantages, reproduction occurs at rates proportional to cellular fitness, and reproduction results in neighbor replacement at random. With this model, cancer field dynamics in tobacco-related head and neck cancer is quantified in a subsequent work [117].

Recently, Temko et. al [130] utilized a MORAN model to explain the known protective effect of full-term pregnancy in early adulthood for estrogen receptor-positive breast cancer later in life.

In part I of this thesis, we will introduce further specific biological applications of extended MORAN models within the cancer context. In these applications, the absorption probabilities and times of the underlying stochastic processes will be of great importance. In order to emphasize the role of the spatial cell arrangement in MORAN models on these quantities, we will derive the absorption behavior of a novel spatial MORAN model in the next section.

2.3 A MORAN MODEL ON A CRYPT-LIKE STRUCTURE

2.3.1 *A crypt-like structure and state space of the model*

MOTIVATION. The neighborhood relations between the individuals in most spatial MORAN models are represented by regular lattices in one dimension [71] or two and three dimensions [29, 35]. As many authors have theoretically shown, the specific choice of these relations has a large influence on the model dynamics, in particular on the absorption behavior of the underlying stochastic process. For example, it has been shown that non-regular structures often imply a higher absorption time compared to regular structures [33, 90]. Moreover, the

absorption probability in state N of three-type MORAN models with mutations on regular structures is the highest if all cells can potentially compete with each other and the lowest for a one-dimensional cell arrangement [26]. While these results are in accordance with the intuition regarding the behavior of the models, there are also counter-intuitive results. For example, in [51] it is analytically shown that the absorption time can decrease when the underlying structure gets less regular. Hence, the absorption behavior of a spatial MORAN model constitutes a theoretically relevant question.

In order to extend the theoretical results of the absorption behavior of MORAN models and to motivate the modeling approach for the applications in part I of this thesis, we introduce a spatial MORAN model on a non-lattice structure. In detail, we introduce a crypt-like structure which is motivated by the human colon [55]. The epithelial layer of the human colon consists of a single sheet of columnar epithelial cells which are characterized by finger-like invaginations into the underlying connective tissue to form the basic functional unit of the intestine, the so-called *crypt*. A stem cell population is located at the base of each crypt which is capable of regenerating all intestinal cell types [55]. The human colon consists of several millions of these crypts which represent separated sub-structures. Hence, the neighborhood relations of the stem cells can be described in a twofold manner by considering both the neighborhood relations between the stem cells within a crypt and the spatial arrangement of the crypts in which the stem cells reside. This serves as motivation for the formulation of a novel spatial MORAN model. We introduce a crypt-like structure which represents cell neighborhood relations that can be regarded as intermediate cell arrangement between a space-free and one-dimensional choice. We derive exact results for the absorption probabilities and absorption times of a two-type MORAN model without mutations and without selection of this model and compare these results to corresponding quantities on regular structures.

MODEL OF A CRYPT-LIKE STRUCTURE. The crypt-like structure is described by a graph $G = (S, E)$. The set of nodes S can be interpreted as locations of cells and the set of edges E as adjacency relation between these locations, i.e. two positions $x, y \in S$ are neighbors if $\{x, y\} \in E$. The definition of the K -crypt structure is as follows.

2.17 Definition (K -crypt structure) Let $N, K \in \mathbb{N}$ such that

$$\kappa := \frac{N}{K} \in \mathbb{N}.$$

For $i = 1, \dots, K$, define

$$S_i = \{(i-1)\kappa + 1, (i-1)\kappa + 2, \dots, i\kappa\}$$

and

$$E_i = \{\{x, y\} | x, y \in S_i, x \neq y\}.$$

The complete graph $G_i = (S_i, E_i)$ is called *i-th crypt* with κ nodes.

Furthermore, let

$$S_{Cr(K)} = \{S_1, S_2, \dots, S_K\} \text{ and } E_{Cr(K)} = \{\{S_i, S_{i+1}\} \mid i = 1, 2, \dots, K-1\}.$$

We call

$$G_{Cr(K)} = (S_{Cr(K)}, E_{Cr(K)})$$

K-crypt structure, see Figure 2 for a visualization.

Hence, each crypt consists of κ nodes which are all connected with each other. Moreover, the crypts themselves form a one-dimensional lattice of length K .

STATE SPACE. Each node within the crypts of the K -crypt structure is either in state 0 or in state 1. These states can be interpreted as two distinct individuals. The intra-crypt state space of crypt i is described as

$$X_i = \{0, 1\}^{S_i}, \quad i = 1, 2, \dots, K.$$

Taking all K crypts into account, the overall state space can be described by

$$X_{\text{intra}} = \{0, 1\}^S \text{ where } S = \bigcup_{i=1}^K S_i.$$

The inter-crypt state space is given by

$$X_{Cr(K)} = \{0, 1, 2, \dots, \kappa\}^K.$$

Hence, inter-crypt states are given by a vector with length K and components ranging from 0 to κ . The inter-crypt state of the model depends on the intra-crypt state by the following mapping $C : X_{\text{intra}} \rightarrow X_{Cr(K)}$ given by

$$C(\xi) = (\omega_i(\xi))_{i=1, \dots, K}$$

where

$$\omega_i : X_i \rightarrow \{0, 1, 2, \dots, \kappa\} : \omega_i(\xi) = \sum_{x \in S_i} \xi(x).$$

Here, $C(\xi)$ provides the number of nodes in state 1 in each of the crypts. In the following, we will use the formulation *crypt i is in state j* for the situation where crypt i contains j nodes in state 1. Figure 2 provides a visualization of the state space of the model and the interplay between the intra-crypt state and the inter-crypt state.

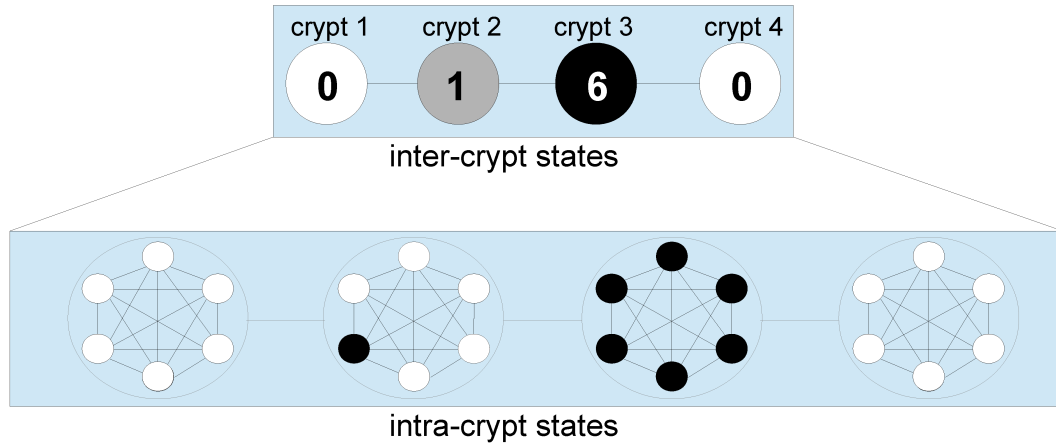


Figure 2: State space of the MORAN model on a crypt-like structure. An example of the K -crypt structure with $K = 4$ crypts, $N = 24$ nodes and $\kappa = \frac{N}{K} = 6$ nodes per crypt. The lower part of the figure illustrates the intra-crypt states. White nodes represent state 0 and black nodes represent state 1. In the upper part, the K -crypt structure is illustrated. Here, the crypts are linearly aligned in a one-dimensional lattice and the states provide the number of nodes in state 1 in the corresponding crypt, e.g. crypt 3 is in state 6. The darker the color, the more black nodes accumulated in the corresponding crypt.

2.3.2 Model dynamics & analysis

INTER-CRYPT AND INTRA-CRYPT DYNAMICS. We assume a classical MORAN model in continuous time for the evolution of the intra-crypt states in which individuals cannot both give birth and die at the same time, see Section 2.2.

Moreover, short-range dynamics can lead to induction of state 1 in neighboring crypts. These short-range dynamics are incorporated as follows. As soon as a crypt is in state κ , i.e. all nodes within the crypt are in state 1, a node in neighboring crypts in state 0 changes to state 1 with rate α , see also Figure 3.

INITIAL STATE OF THE MODEL. We assume that initially one node in crypt 1 is in state 1 and all other nodes are in state 0. Formally, this initial state can be described by state $\zeta_0 \in X_{Cr(K)}$ which is given by

$$\zeta_0 = (1, 0, 0, \dots, 0).$$

In the following, we will refer to the introduced dynamics with initial state ζ_0 as *Moran model on a crypt-like structure*.

CORRESPONDING BIRTH-AND-DEATH PROCESS. The MORAN model on a crypt-like structure can be regarded as a birth-and-death process in the following way. For crypts in state 0, the state can only change as consequence of short-range dynamics. Therefore, crypt 1 needs to be in state κ in order to induce

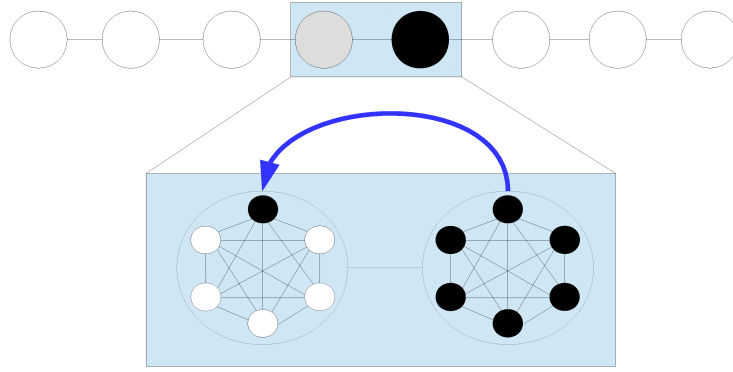


Figure 3: Intra-crypt short-range dynamics. If a crypt is in state κ , i.e. all nodes are in state 1, an arbitrary node in neighboring crypts flips from state 0 to state 1 with rate α .

state 1 in crypt 2 and, subsequently, crypt 2 needs to reach state κ first in order to subsequently induce state 1 in crypt 3 and so on. It is therefore sufficient to describe the state of the model by the number of crypts in state κ and the state of the neighboring crypt to the right. Taking these considerations into account, the dynamics can be described in a reduced way as a birth-and-death process $(B_t)_{t \geq 0}$ on $\{0, 1, 2, \dots, N\}$ since the overall number of nodes in state 1 contains all relevant information for the further evolution of the model.

If the state of the model at time t is given by some inter-crypt state $\zeta \in X_{Cr(K)}$, the state of $(B_t)_{t \geq 0}$ is the overall number of nodes in state 1, i.e.

$$B_t := \sum_{i=1}^K \zeta(i).$$

The birth rates of B_t are given by

$$c^+(i) = \begin{cases} \frac{(i \bmod \kappa)(\kappa - (i \bmod \kappa))}{\kappa - 1}, & 0 < i < N, i \bmod \kappa \neq 0, \\ \alpha, & 0 < i < N, i \bmod \kappa = 0, \\ 0, & \text{else.} \end{cases}$$

and the death rates are

$$c^-(i) = \begin{cases} \frac{(i \bmod \kappa)(\kappa - (i \bmod \kappa))}{\kappa - 1}, & 0 < i < N, i \bmod \kappa \neq 0, \\ 0, & \text{else.} \end{cases}$$

with initial state $B_0 = 1$. These rates can be explained as follows. If one crypt contains both nodes in state 0 and 1, the intra-crypt MORAN dynamics of cell death and cell proliferation determine the dynamics, see Section 2.2. Note that

we utilize a continuous time version of the classical MORAN model and assume that an individual cannot both give birth and die at the same time.

The number of nodes in state 1 in the corresponding crypt is increased or decreased until the corresponding crypt is either solely occupied by nodes in state 0 or 1. Inter-crypt short-range interactions apply if two neighboring crypts are in states 0 and κ , respectively. Then, an induction of state 1 in the neighboring crypt in state 0 occurs with rate α . Note that the process $(B_t)_{t \geq 0}$ has two absorbing states 0 and N . We are interested in the absorption of $(B_t)_{t \geq 0}$ in state N . In the following, we refer to $(B_t)_{t \geq 0}$ as *corresponding b-d process* of the MORAN model on a crypt-like structure and derive the absorption probabilities and times of the underlying stochastic process.

2.3.3 Derivation of the absorption probabilities and absorption times

The absorption probability in state N of the corresponding b-d process $(B_t)_{t \geq 0}$ starting in state i is defined as

$$\alpha^N(i) := \mathbb{P}(B_t = N \text{ for some } t \geq 0 | B_0 = i)$$

and can be derived by using a specific formula for birth-and-death processes [133]

$$\alpha^N(i) = \frac{1 + \sum_{k=1}^{i-1} \prod_{j=1}^k \frac{c^-(j)}{c^+(j)}}{1 + \sum_{k=1}^{N-1} \prod_{j=1}^k \frac{c^-(j)}{c^+(j)}}. \quad (2.7)$$

The absorption time of $(B_t)_{t \geq 0}$ conditioned on being absorbed in state N and starting in state 1 is denoted by $\tau^N(1)$ and can be derived by

$$\tau^N(1) = \sum_{k=1}^{N-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)}, \quad (2.8)$$

see [133].

2.18 Definition For $n \in \mathbb{N}$, the n -th harmonic number is defined by

$$H_n := \sum_{k=1}^n \frac{1}{k}.$$

We apply formulas (2.7) and (2.8) to derive the absorption probability and time of the MORAN model on a crypt-like structure. The following considerations are helpful for this derivation.

2.19 Lemma The following equalities hold for $(B_t)_{t \geq 0}$, the corresponding b-d process of the MORAN model on a crypt-like structure.

1.
$$\prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} = \begin{cases} 0, \exists n \in \{l+1, \dots, k\} : n \bmod \kappa = 0, \\ 1, \text{ else,} \end{cases} \quad , k = 1, \dots, N.$$
2.
$$\prod_{m=l+1}^{j\kappa} \frac{c^-(m)}{c^+(m)} = \begin{cases} 1, l \geq j\kappa, \\ 0, \text{ else,} \end{cases} \quad , j = 1, \dots, K.$$
3.
$$\sum_{k=j\kappa+1}^{(j+1)\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} = \sum_{k=j\kappa+1}^{(j+1)\kappa-1} \sum_{l=j\kappa}^k \frac{\alpha^N(l)}{c^+(l)}, \quad j = 1, \dots, K-1.$$

Proof. 1. This equality follows from the fact that $c^-(m) = 0$ for $m \bmod \kappa = 0$.

2. This product is a special case of the first equality and is only 1 if the product is empty.

3. This statement is a direct consequence of the first two equalities. \square

2.20 Theorem The absorption probability of the corresponding b-d process $(B_t)_{t \geq 0}$ is

$$\alpha^N(i) = \begin{cases} \frac{i}{\kappa}, & i < \kappa, \\ 1, & i \geq \kappa, \end{cases} \quad (2.9)$$

and the conditional absorption time conditioned on absorption in state N is given by

$$\tau^K(1) = (\kappa - 1) \left((K-1)H_\kappa + 1 - \frac{K}{\kappa} \right) + \frac{(K-1)\kappa}{\alpha} \quad (2.10)$$

or, in dependency of N ,

$$\tau^N(1) = \frac{N-K}{K} \left((K-1)H_{\frac{N}{K}} + \frac{N-K^2}{N} \right) + \frac{(K-1)N}{K\alpha}.$$

Proof. Lemma 2.19 implies that

$$1 + \sum_{k=1}^{N-1} \prod_{j=1}^k \frac{c^-(j)}{c^+(j)} = 1 + \sum_{k=1}^{\kappa-1} \prod_{j=1}^k \frac{c^-(j)}{c^+(j)} = \kappa.$$

since $c^-(j) = c^+(j)$ for $j < \kappa$.

Furthermore,

$$1 + \sum_{k=1}^{i-1} \prod_{j=1}^k \frac{c^-(j)}{c^+(j)} = \begin{cases} 1 + (i-1) = i, & i < \kappa, \\ 1 + \sum_{k=1}^{\kappa-1} \prod_{j=1}^k \frac{c^-(j)}{c^+(j)} = \kappa, & i \geq \kappa. \end{cases}$$

Hence, equation (2.7) implies that

$$\alpha^N(i) = \begin{cases} \frac{i}{\kappa}, & i < \kappa, \\ 1, & i \geq \kappa. \end{cases} \quad (2.11)$$

The absorption time can be calculated with equation (2.8) as follows. First, the sum is split such that Lemma 2.19 can be utilized.

$$\begin{aligned} \tau^N(1) &= \sum_{k=1}^{N-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} \\ &= \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} + \sum_{j=1}^{K-1} \sum_{k=j\kappa}^{(j+1)\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} \\ &= \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} + \\ &\quad \sum_{j=1}^{K-1} \left(\sum_{k=j\kappa+1}^{(j+1)\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^k \frac{c^-(m)}{c^+(m)} + \sum_{l=1}^{j\kappa} \frac{\alpha^N(l)}{c^+(l)} \prod_{m=l+1}^{j\kappa} \frac{c^-(m)}{c^+(m)} \right). \end{aligned}$$

By Lemma 2.19 and by splitting the sum in the brackets we obtain from the last equality above that

$$\begin{aligned} \tau^N(1) &= \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} \cdot 1 + \sum_{j=1}^{K-1} \left(\sum_{k=j\kappa+1}^{(j+1)\kappa-1} \sum_{l=j\kappa}^k \frac{\alpha^N(l)}{c^+(l)} + \frac{\alpha^N(j\kappa)}{c^+(j\kappa)} \right) \\ &= \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} + \sum_{j=1}^{K-1} \left(\sum_{k=j\kappa+1}^{(j+1)\kappa-1} \sum_{l=j\kappa+1}^k \frac{\alpha^N(l)}{c^+(l)} + \sum_{k=j\kappa+1}^{(j+1)\kappa-1} \frac{\alpha^N(j\kappa)}{c^+(j\kappa)} + \frac{\alpha^N(j\kappa)}{c^+(j\kappa)} \right). \end{aligned}$$

Summarizing the last terms in brackets and rearranging the sums leads to

$$\tau^N(1) = \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} + \sum_{j=1}^{K-1} \sum_{k=j\kappa+1}^{(j+1)\kappa-1} \sum_{l=j\kappa+1}^k \frac{\alpha^N(l)}{c^+(l)} + \kappa \sum_{j=1}^{K-1} \frac{\alpha^N(j\kappa)}{c^+(j\kappa)}.$$

In the next step, the indices of the second term are adjusted, i.e. $k \rightarrow k - j\kappa$ and $l \rightarrow l - j\kappa$, which yields

$$\tau^N(1) = \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l)}{c^+(l)} + \sum_{j=1}^{K-1} \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\alpha^N(l+j\kappa)}{c^+(l+j\kappa)} + \kappa \sum_{j=1}^{K-1} \frac{\alpha^N(j\kappa)}{c^+(j\kappa)}.$$

From this, we obtain by plugging in the absorption probability and birth and death rates that

$$\begin{aligned}\tau^N(1) &= \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{l(\kappa-1)}{\kappa l(\kappa-l)} + \sum_{j=1}^{K-1} \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{\kappa-1}{l(\kappa-l)} + \frac{\kappa(K-1)}{\alpha} \\ &= \frac{\kappa-1}{\kappa} \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{1}{\kappa-l} + (\kappa-1) \sum_{j=1}^{K-1} \sum_{k=1}^{\kappa-1} \sum_{l=1}^k \frac{1}{l(\kappa-l)} + \frac{\kappa(K-1)}{\alpha}\end{aligned}$$

We evaluate the sums and utilize that $H_{\kappa-1} = H_{\kappa} - \frac{1}{\kappa}$ and summarize to obtain

$$\begin{aligned}\tau^N(1) &= \frac{\kappa-1}{\kappa}(\kappa-1) + (\kappa-1)(K-1)H_{\kappa-1} + \frac{\kappa(K-1)}{\alpha} \\ &= (\kappa-1) \left(\frac{\kappa-1}{\kappa} + (K-1)H_{\kappa-1} \right) + \frac{\kappa(K-1)}{\alpha} \\ &= (\kappa-1) \left((K-1)H_{\kappa} + 1 - \frac{K}{\kappa} \right) + \frac{\kappa(K-1)}{\alpha}.\end{aligned}$$

Substituting $\kappa = \frac{N}{K}$ yields the absorption time in dependency of N

$$\tau^N(1) = \frac{N-K}{K} \left((K-1)H_{\frac{N}{K}} + \frac{N-K^2}{N} \right) + \frac{(K-1)N}{K\alpha}. \quad (2.12)$$

□

2.21 Proposition For the conditional absorption time in state N of the MORAN model on a crypt-like structure it holds that

$$\tau^N(1) \in O(N \log N).$$

Proof. The absorption time (2.12) is equivalent to

$$\tau^N(1) = \frac{(N-K)(K-1)H_{\frac{N}{K}}}{K} + \frac{(N-K^2)(N-K)}{NK} + \frac{(K-1)N}{K\alpha}.$$

Since $H_N \in O(\log N)$ it holds that $\tau^N(1) \in O(N \log N)$. □

2.3.4 Comparison with MORAN models on lattice structures

We compare the absorption probabilities and absorption times of the MORAN model on a crypt-like structure with the known absorption times of the neutral two-type MORAN model without mutations on the complete graph and the line graph with N nodes each. The results of this comparison are summarized in Table 2 and the following Corollary.

crypts	structure	$\tau^N(1)$	$\alpha^N(1)$
-	line graph (MORAN only)	$\frac{N^2-1}{6}$ [4]	$\frac{1}{N}$ [82]
$K = 1$	complete graph	$\frac{(N-1)^2}{N}$ [126]	$\frac{1}{N}$ [82]
$1 < K < N$	K -crypt graph	$\frac{N-K}{K} \left((K-1)H_{\frac{N}{K}} + \frac{N-K^2}{N} \right) + \frac{(K-1)N}{K\alpha}$	$\frac{K}{N} = \frac{1}{\kappa}$
$K = N$	line graph (short-range only)	$\frac{N-1}{\alpha}$	1

Table 2: The absorption probabilities and times of MORAN models on different structures. The table contains the absorption probabilities and times of the MORAN model on regular structures and of the MORAN model on a crypt-like structure. Note that the MORAN model on a crypt-like structure reflects MORAN dynamics on the complete graph for the special case $K = 1$ but not on the line graph for the special case $K = N$. In the latter case there are no MORAN dynamics since each crypt consists only of one cell.

2.22 Corollary Let $\alpha_G^N(1)$ and $\tau_G^N(1)$ denote the absorption probability and conditional absorption time in state N of the MORAN model on structure G where G_l denotes the line graph, G_C the complete graph and $G_{Cr(K)}$ the K -crypt structure with N nodes each. For $K = 1, 2, \dots, N - 1$ it holds for the absorption probabilities that

$$\alpha_{G_C}^N(1) = \alpha_{G_l}^N(1) \leq \alpha_{G_{Cr(K)}}^N(1).$$

For the conditional absorption times it holds for N sufficiently large, $\alpha > 0$ and $1 < K < N$ that

$$\tau_{G_C}^N(1) < \tau_{G_{Cr(K)}}^N(1) < \tau_{G_l}^N(1).$$

Proof. The observation for the probabilities holds due to the summarized results in Table 2. These results also imply that $\tau_{G_C}^N(1) \in O(N)$ and $\tau_{G_l}^N(1) \in O(N^2)$ which yields together with Proposition 2.21 the given order of the absorption times. \square

It turns out that the absorption probability is higher for the MORAN model on a crypt-like structure compared to the complete and the line graph. This observation can be explained by the twofold structure of the K -crypt graph. If every node in a crypt is in state 1, the state of this crypt will not change anymore and state 1 will be induced to the neighboring crypt by short-range dynamics. The higher absorption probability on the K -crypt structure is in accordance with results showing that individuals have a reasonable chance to overtake small compartments even without fitness advantage only by genetic drift [94].

The asymptotic absorption time of the MORAN model on a crypt-like structure is different from those of the MORAN dynamics on the complete and the line graph. The crypt structure delays the absorption time compared to the complete graph but decreases the absorption time compared to the line graph. This

can be explained by the fact that the K -crypt graph can be seen as an intermediate structure between the complete graph and the line graph. However, this finding is not trivial since also counter-intuitive results of absorption quantities on different structures can be constructed [51]. Moreover, our theoretical derivations provide the exact absorption time of the MORAN model on a crypt-like structure and allow a quantitative comparison to the absorption behavior on other structures.

Note that the introduced model reduces to the non-spatial MORAN model on the complete graph if $K = 1$. For this special case, the well-known absorption time of the MORAN model is recovered by our results [126]. This holds also for the other special case $K = N$ for which no MORAN dynamics is possible since each crypt consists only of one individual and therefore only short-range dynamics determine the evolution of the process.

Overall, the presented results provide an extension of previous works investigating MORAN dynamics on non-regular structures [3, 51, 126] by deriving exact solutions for the absorption probabilities and times for the MORAN model on a crypt-like structure. Importantly, the insights of this section will justify the choice of the underlying spatial structure in the applications of MORAN models in the following part of this thesis.

Part I

MORAN MODELS FOR THE INFERENCE OF GENETIC CELL CHANGES ON TUMOR DEVELOPMENT

In this part we introduce and analyze extended MORAN models to investigate the following two questions:

- In which way do *cellular* processes like competition, *genetic* alterations and proliferation and death determine whether tumors on the *tissue scale* grow, progress or even spontaneously regress?
- At which point of the *cellular* multistep process of cancer and the clonal expansion of *genetically* altered cells is the fate of tumor development decided?

From a mathematical point of view, the main results include the derivation of the absorption probabilities of the extended non-spatial and spatial three-type MORAN models with mutations which are derived in Theorems [3.4](#) and [4.3](#).

3

MODEL-BASED EVALUATION OF SPONTANEOUS TUMOR REGRESSION IN PILOCYTIC ASTROCYTOMA*

Contents

3.1	Clinical characteristics of pilocytic astrocytoma	39
3.2	A MORAN model for pilocytic astrocytoma growth and progression	41
3.2.1	Definition of a PA growth and progression model	41
3.2.2	Analysis of the TGP process	44
3.2.3	Derivation of tumor regression functions for the TGP process	53
3.3	Results & clinical implications	57
3.3.1	Derivation of the PA-regression-function	57
3.3.2	Linear dependency between residual tumor fraction and regression probability of PA	58
3.3.3	Quantitative prediction of the regression probability for benign PA	58
3.3.4	Non-existence of an extent of resection (EOR) threshold	58
3.4	Discussion & outlook	59

3.1 CLINICAL CHARACTERISTICS OF PILOCYTIC ASTROCYTOMA

Pilocytic astrocytoma (PA) is the most common pediatric brain tumor and the second most frequent tumor in childhood [43]. Three of four cases are diagnosed up to an age of 20 years with the highest age incidence between 5 and

* This chapter includes text and figures from the publication and supplementary material of: Thomas Buder, Andreas Deutsch, Barbara Klink, and Anja Voss-Böhme, *PLoS Computational Biology*, 11(12):e1004662, 2015. Author contribution: Thomas Buder wrote the paper together with Andreas Deutsch, Barbara Klink and Anja Voss-Böhme. Andreas Deutsch, Barbara Klink and Anja Voss-Böhme supervised the study and gave substantial input to the manuscript. Thomas Buder conceived and designed the model. Thomas Buder analyzed the model together with Anja Voss-Böhme.

15 years. PA is usually benign, often follows an indolent course and is mostly slow-growing [88]. In children, PA most frequently occurs in the cerebellum but can develop in the entire neuroaxis. Surgery is the treatment of choice [20]. If total excision is achieved, the prognosis is favorable with more than 90% of patients being cured [85]. However, in many cases tumor location in critical or deep areas (such as brain stem, optic pathway, or hypothalamus) restricts resection options and alternative management options are required [60, 115]. Patients with only partial resection have a worse and highly unpredictable prognosis [60, 85]. Tumors can regrow or even progress to a more aggressive tumor [20, 105, 114, 116, 120, 123] but spontaneous tumor regression of PA has also been observed [44, 85, 107, 125, 128] and is a common phenomenon. A review in [107] estimates a fraction of 14% of all residual cerebellar astrocytoma that regress spontaneously. Other studies claim an even higher portion [23]. While regression of PA after partial resection is reported in many case series [23, 44, 107, 125, 128], the influence of the residual tumor size has not been evaluated yet. Moreover, the management for patients in whom complete resection cannot be achieved is still unclear. Due to the chance of regression and the indolent nature of PA some authors propose a wait and see strategy in order to avoid potential risks induced by further therapies [85, 107, 123]. Other authors favor an aggressive surgical resection in combination with additional treatment strategies like radiation and chemotherapy to control tumor growth [52, 124, 125].

On the molecular level it has been shown that activation of the mitogen-activated protein kinase (MAPK) pathway is sufficient to induce the development of PA. This leads to the hypothesis that PA is a single-pathway disease [62, 63]. Furthermore, PA usually harbor only one alteration within the MAPK pathway. The majority of mutations are activating changes in the BRAF gene, the most common is the KIAA1549-BRAF fusion, but also other activating mutations have been described.

A more aggressive behavior of PA is observed if additional genetic alterations occur, e.g. loss of tumor suppressor gene CDKN2A [53, 114]. Furthermore, alterations in the PI3K/AKT pathway [66] have been associated with aggressive forms of PA [116]. One proposed mechanism for the often observed slow growth of the tumors is oncogene-induced senescence which is a mechanism limiting neoplastic growth by inducing cellular senescence. The MAPK activation might initially promote growth as well as induce senescence. Oncogene-induced senescence has also been observed in melanocytic nevi and melanoma [114]. Several mechanisms for tumor regression have been suggested, e.g. immunologic mechanisms, hormonal factors, induction of differentiation or apoptosis [128]. However, the reason why regression in PA occurs is not understood yet [85].

We formulate a mathematical model for growth, progression and regression of PA based on the above described clinical and molecular biological observations. We study the effects of competition between tumor and wild-type cells on the chance for regression. We distinguish two types of PA. Benign cases are

classified as *PA-I tumors* and assumed to be caused by alteration of a single pathway. Tumors in which an additional alteration occurs are categorized as *PA-II tumors* representing the more aggressive subset of *PA*. We introduce a stochastic tumor growth and progression model which is a MORAN model [98] with mutations. We choose this model class since astrocyte proliferation and diversification mainly happen during late embryogenesis and the first three weeks after birth. These processes are largely complete by early postnatal stages while early and late postnatal development is mainly characterized by maturation processes (like continuing elaboration of astrocyte processes and building of synaptic/vascular connections) [38, 97]. Since *PA* are usually diagnosed between 5 and 15 years the normal astrocyte population is not proliferating at this time anymore. Therefore, it is reasonable to assume an approximately homeostatic tissue. In such a tissue MORAN dynamics provide a natural and established framework for modeling competition between tumor and wild-type cells. In contrast to the introduced MORAN models in Section 2.2, we utilize a continuous time version and assume that cells cannot both give birth and die at the same time.

In our model, we derive the *PA*-regression-function describing the probability for regression in dependency of the residual tumor size after partial resection of benign *PA*. The accumulation of mutations in a tissue has been modeled and investigated by several authors by using a MORAN model. Work by IWASA, MICHOR, KOMAROVA and NOWAK [59, 100] has been extended by SCHWEINSBERG [121] and DURRETT, SCHMIDT and SCHWEINSBERG [29] to the case of m mutations. These models analyze tumor growth and progression [10, 58, 71, 96, 133] with a focus on theoretical results regarding the waiting time until a cell has accumulated a certain number of mutations. Our approach is motivated by a concrete clinical question which is the regression probability of a benign *PA* tumor in dependency of the residual tumor size. We modify the model introduced in [29]. In particular, we consider MORAN dynamics with two mutations but two absorbing states and investigate the precise relation of the two absorption probabilities which allows the incorporation of epidemiological data to calibrate the model. From the mathematical point of view, the relation of the two absorption probabilities can be connected to the portion of stochastic tunneling events in the model presented in [29].

3.2 A MORAN MODEL FOR PILOCYTIC ASTROCYTOMA GROWTH AND PROGRESSION

3.2.1 Definition of a *PA* growth and progression model

STATE SPACE AND REPRESENTATION OF *PA* SUBTYPES. We use a MORAN model with mutations to model tumor growth, regression and progression of *PA*. The model incorporates three cell types, *wild-type cells*, *type-I cells* and *type-II cells*. Wild-type cells have no genetic alteration. Type-I cells are characterized

by a MAPK pathway alteration. We assume that type-I cells proliferate without fitness advantage. This is motivated by the observation that benign PA grow very slowly, can be stable in size over a long time period or even regress. Type-II cells have acquired a second genetic alteration disabling oncogene-induced senescence, for example by loss of CDKN2A. Due to their aggressive behavior, we assume a very large fitness advantage of type-II cells. The parameter N in our model represents a critical tumor size in the sense that a PA which reaches size N cannot spontaneously regress. Our assumption that such a critical size exists is founded on clinical data and observations which are explained later. Here, we focus on the chance of tumor regression of a partially resected benign PA in dependency of the residual tumor size. Therefore, we do not consider growth of tumors beyond the critical size N since spontaneous regression would not be possible anymore in such a case. Spatial aspects are neglected so that the number of cells of each cell type is sufficient to describe the states of our model. Therefore, the state space can be described by $S = \{0, 1, 2, \dots, N, E\}$. Here, states 0 to N represent the occurrence of the respective number of type-I cells and no type-II cell. The additional state E indicates the presence of a type-II cell. States N and E are absorbing states of the model and represent the occurrence of a benign PA which we call *PA-I tumor* and an aggressive form of PA named *PA-II tumor*, respectively. Hence, the occurrence of a PA-I tumor is represented by the accumulation of N type-I cells. As soon as a single type-II cell appears, we identify this state as occurrence of a PA-II tumor which is modeled as absorption in state E . Please note that no transition from state N to state E is possible in the model. Figure 4 illustrates the three cell types and the representation of both types of PA tumors in the model which we call *tumor growth and progression (TGP)* process.

DYNAMICS IN THE MODEL. The dynamics of our model incorporates cell death, proliferation and genetic alterations. Cell death and proliferation are modeled according to the MORAN dynamics [98] as follows. Two cells are chosen randomly. One of these cells undergoes cell death and the other cell proliferates. The offspring of the proliferating cell substitutes the cell chosen for death. Since we neglect spatial aspects, the offspring of a cell can replace any other cell. During proliferation, a mutation of the new-born cell can occur. Wild-type cells mutate to type-I cells with probability u and these mutate to type-II cells with probability v . MORAN dynamics is defined with respect to a *relevant cell number* which describes the number of cells that potentially compete with each other. This relevant cell number can be assumed to be approximately equal to the critical tumor size N by the following arguments. Since PA tumors grow as a solid, well-circumscribed mass within the normal brain, new tumor cells are placed in the vicinity of already mutated cells. Similarly, only wild-type cells in the vicinity of mutated cells can potentially compete with tumor cells. Thus, the actual relevant cell number for the MORAN dynamics is somewhere between the

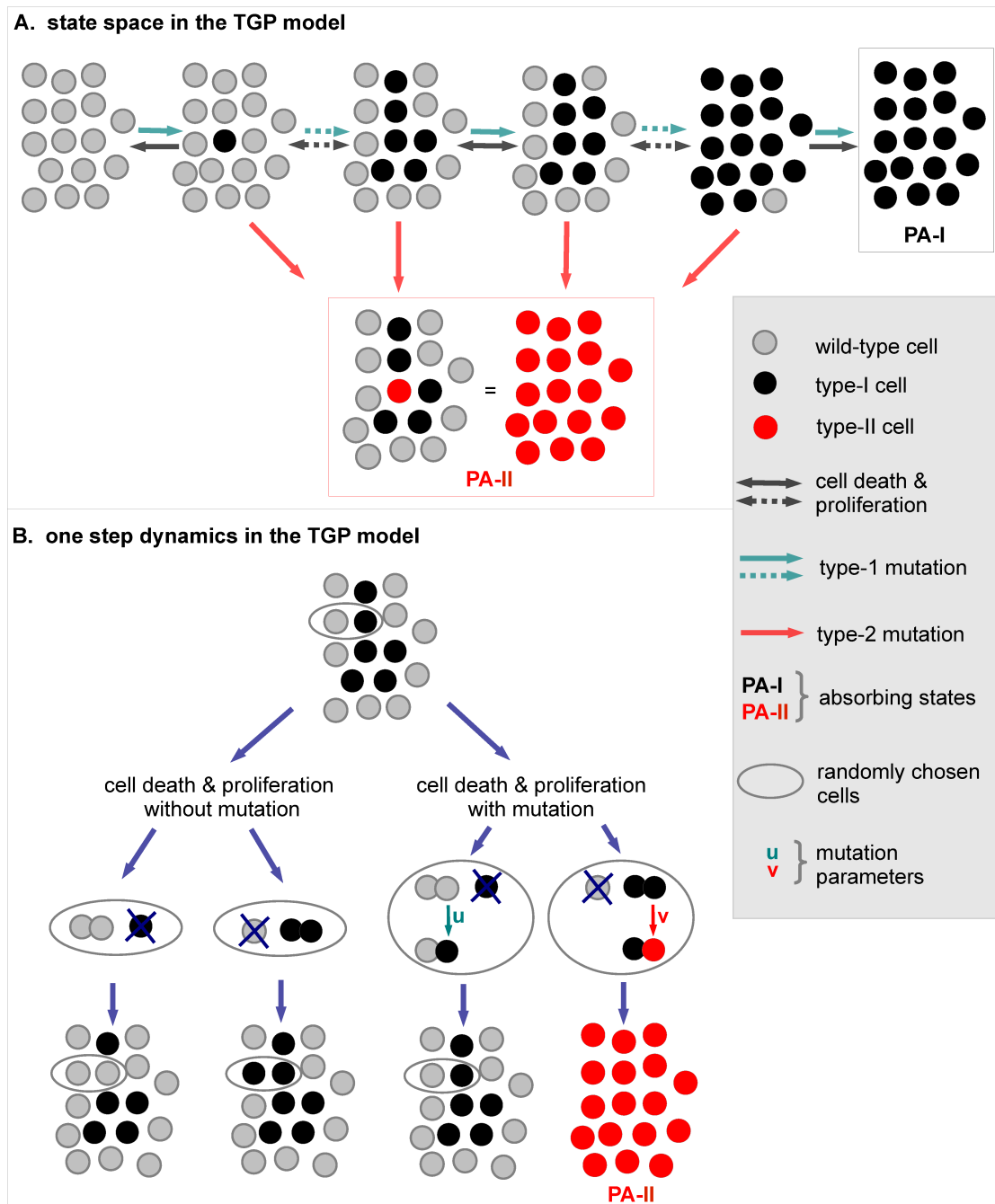


Figure 4: States and dynamics in the TGP process. A. The initial state of the TGP process is *all-cells-wild-type*. Type-I cells (black cells) can arise due to type-I mutations. If the number of type-I cells reaches at least the critical tumor size N , a PA-I tumor developed which cannot regress anymore (black box). We assume that a PA-II tumor occurs as soon as the first type-II cell appears in the system (red box). Solid arrows indicate a single transition whereas dotted arrows indicate several transitions. B. Wild-type cells mutate to type-I cells with probability u and type-I cells mutate to type-II cells with probability v during proliferation. Cell death and proliferation are included as follows. A cell is randomly chosen for cell death and replaced by the offspring of another randomly chosen cell. Spatial aspects are neglected in the model.

critical tumor size N and the total number of astrocytes in the brain but clearly much closer to N . Since there is no detailed experimental estimate of this number so far, we assume in our model that the relevant cell number for MORAN dynamics is equal to N . Although we do not explicitly incorporate spatial aspects, this assumption implicitly incorporates spatial aspects by implying that tumor cells cannot place its offspring too far away. We assume that initially all cells are wild-type cells. Hence, the process starts in state 0. The number of type-I cells changes according to the above described dynamics which are formally defined in the following way.

3.1 Definition (TGP process) The tumor growth and progression (TGP) process is a continuous time MARKOV chain $(X_t)_{t \geq 0}$ on the state space $S = \{0, 1, 2, \dots, N, E\}$ with rate matrix $Q = (q(k, l))_{k, l \in S}$ which is given by

$$q(k, l) = \begin{cases} Nu, & k = 0, l = 1, \\ \frac{(N-k)k}{N}(1-v) + \frac{(N-k)(N-k-1)u}{N}, & 1 \leq k \leq N-1, l = k+1, \\ \frac{(N-k)k(1-u)}{N}, & 1 \leq k \leq N-1, l = k-1, \\ kv, & 1 \leq k \leq N-1, l = E, \\ -\sum_{\substack{m \in S \\ m \neq k}} q(k, m), & l = k, \\ 0, & \text{else.} \end{cases} \quad (3.1)$$

The dynamics are illustrated in Figure 4. The TGP process is a MARKOV process with two absorbing states N and E representing a PA-I tumor and a PA-II tumor, respectively. Hence, the absorption probabilities of the TGP process in both states correspond to the clinically observed fraction of PA-I and PA-II tumors. Therefore, we will derive these absorption probabilities and analyze in which way they depend on the model parameters N , u and v . Furthermore, we assume that tumor regression is characterized by the vanishing of all tumor cells. Hence, tumor regression corresponds to reaching state 0 in the TGP process which will be described by a tumor regression function in the following.

3.2.2 Analysis of the TGP process

The behavior of the TGP process depends on its three parameters, the critical tumor size N , the mutation probability from wild-type cells to type-I cells u and the mutation probability from type-I cells to type-II cells v . For the analysis, we assume the following parameter regime.

3.2 Assumption The parameter regime for the analysis of the TGP process is chosen such that

$$u \ll \frac{1}{N} \quad (3.2)$$

and

$$N\sqrt{v} > 0. \quad (3.3)$$

In the following we explain this choice. It turns out that the parameters N and v play an important role in the analysis via equation (3.3) which motivates the following definition.

3.3 Definition The *risk coefficient* γ of the TGP process is defined as

$$\gamma := N\sqrt{v}.$$

DECOMPOSITION OF THE TGP PROCESS INTO TWO SUB-PROCESSES. Assumption (3.2) implies that type-I mutations are rare. Typically, an emerging type-I lineage either goes extinct or leads to absorption of the TGP process before another type-I mutation occurs. Hence, each newly arising type-I mutant can be investigated independently. This idea has been introduced in [29]. Therefore, we set $u = 0$ as soon as type-I cells are present, i.e. if the TGP process is in state k , $1 \leq k \leq N$. Since absorption is inevitable in the TGP process, a *successful* type-I mutant that leads to absorption in state N or state E must eventually occur. Before the occurrence of this specific type-I mutant, *unsuccessful* mutants arise and go extinct driving the process back to state 0. Hence, assumption (3.2) implies that a PA tumor develops from a single mutated cell. See also Figure 5 for an illustration of this decomposition of the TGP process.

DERIVATION OF THE ABSORPTION PROBABILITY OF THE TGP PROCESS. We are interested in the absorption probability of the TGP process in state N when the process is started with a single type-I cell. This quantity is denoted by $\alpha^N(1, v)$ and is derived in the following.

3.4 Theorem (Absorption probabilities of the TGP process) Consider the TGP process with rate matrix (3.1) under assumptions (3.2) and (3.3). Then, the absorption probability in state N , $N \in \mathbb{N}$, starting the process in state 1 is given by

$$\alpha^N(1, v) = \frac{1}{P_{N-1} \left(\frac{v+1}{1-v} \right)}, \quad (3.4)$$

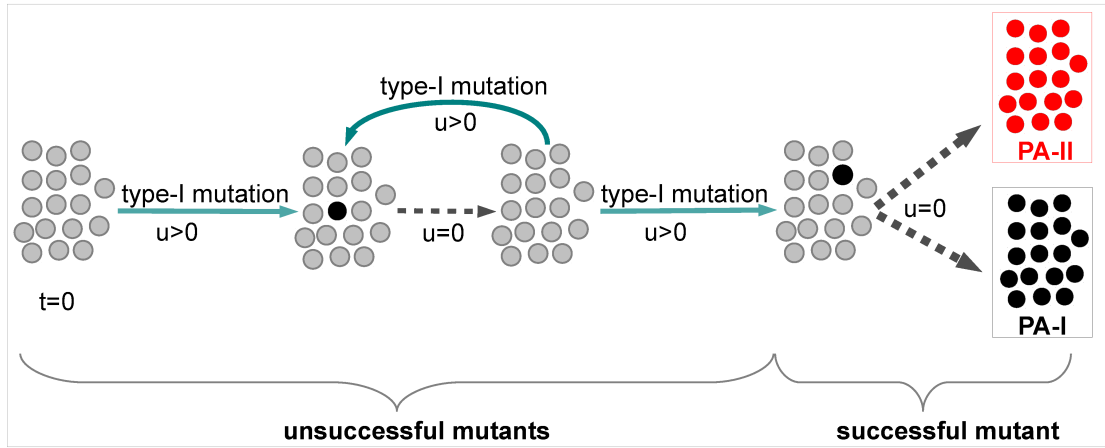


Figure 5: Decomposition of the TGP process. Assumption (3.2) implies that no other PA-I mutation occurs if type-I cells are already present in the system. Therefore, u can be set to zero when a single type-I cell emerged. This allows to decompose the process into two sub-processes. First, occurrence of unsuccessful mutants which go extinct and, second, the occurrence of a successful mutant which leads to absorption in one of the PA states.

where P_N denotes the LEGENDRE polynomials. Moreover, the asymptotic absorption probability in dependency of the risk coefficient γ as the system size N tends to infinity is

$$\alpha(\gamma) = \lim_{N \rightarrow \infty} \alpha^N(1, v) = \frac{1}{I_0(2\gamma)}, \quad (3.5)$$

where I_0 denotes the BESSEL function of the first kind.

Proof. As Figure 5 illustrates, the absorption probabilities of the TGP process agree with those of the sub-process starting with emergence of the successful mutant. This sub-process is given by the state space $\tilde{S} = \{1, 2, 3, \dots, N, E\}$ and rate matrix $\tilde{Q} = (\tilde{q}(k, l))_{k, l \in \tilde{S}}$ which is obtained from the original rates (3.1) by eliminating state 0 and setting $u = 0$ such that

$$\tilde{q}(k, l) = \begin{cases} \frac{(N-k)k}{N}(1-v), & 1 \leq k \leq N-1, l = k+1, \\ \frac{(N-k)k}{N}, & 2 \leq k \leq N-1, l = k-1, \\ kv, & 1 \leq k \leq N-1, l = E, \\ -\sum_{\substack{m \in \tilde{S} \\ m \neq k}} q(k, m), & l = k, \\ 0, & \text{else.} \end{cases} \quad (3.6)$$

By defining $\tilde{q}(k) := -\tilde{q}(k, k)$, we get

$$\tilde{q}(1) = \tilde{q}(1, E) + \tilde{q}(1, 2) = \frac{N + v - 1}{N},$$

$$\tilde{q}(N) = \tilde{q}(E) = 0,$$

$$\tilde{q}(k) = \tilde{q}(k, k + 1) + \tilde{q}(k, k - 1) + \tilde{q}(k, E) = \frac{2(N - k)k + k^2v}{N}, \quad 2 \leq k \leq N - 1.$$

We further regard the embedded MARKOV chain with transition probabilities

$$p(i, j) = \begin{cases} \frac{\tilde{q}(i, j)}{\tilde{q}(i)}, & i \neq j, \\ 1, & i = j = E, \\ 1, & i = j = N, \\ 0, & \text{else,} \end{cases}$$

in which the entries unequal to 0 look as follows

$$\begin{aligned} p(1, E) &= \frac{Nv}{N + v - 1}, \\ p(1, 2) &= \frac{(N - 1)(1 - v)}{N + v - 1}, \\ p(k, E) &= \frac{Nv}{2(N - k) + kv}, \quad 2 \leq k \leq N - 1, \\ p(k, k + 1) &= \frac{(N - k)(1 - v)}{2(N - k) + kv}, \quad 2 \leq k \leq N - 1, \\ p(k, k - 1) &= \frac{(N - k)}{2(N - k) + kv}, \quad 2 \leq k \leq N - 1, \\ p(N, N) &= p(E, E) = 1. \end{aligned}$$

The absorption probabilities for the underlying stochastic process with transition matrix $P = (p_{i, j})_{i, j \in \tilde{S}}$ is obtained as follows. Denote by $\alpha^N = (\alpha^N(i, v))_{i \in \tilde{S}}$ the absorption probabilities where $\alpha^N(i, v)$ describes the absorption probability in state N starting from state i . First step analysis yields

$$\alpha^N(i, v) = \sum_{j \in \tilde{S}} p(i, j) \alpha^N(j, v), \quad i \in \tilde{S}.$$

It holds that $\alpha^N(E, v) = 0$, $\alpha^N(N, v) = 1$ and therefore

$$\begin{aligned} \alpha^N(i, v) &= \sum_{j=1}^N p(i, j) \alpha^N(j, v) = \sum_{j=1}^{N-1} p(i, j) \alpha^N(j, v) + p(i, N) \\ &= p(i, i-1) \alpha^N(i-1, v) + p(i, i+1) \alpha^N(i+1, v) + p(i, N) \\ &= \begin{cases} \frac{(N-1)(1-v)}{N+v-1} \alpha^N(2, v), & i = 1, \\ \frac{(N-i)}{2(N-i)+iv} \alpha^N(i-1, v) + \frac{(N-i)(1-v)}{2(N-i)+iv} \alpha^N(i+1, v), & 2 \leq i \leq N-2, \\ \frac{1}{2+(N-1)v} \alpha^N(N-2, v) + \frac{1-v}{2+(N-1)v}, & i = N-1. \end{cases} \end{aligned}$$

Hence,

$$\begin{aligned} -\alpha^N(1, v) + \frac{(N-1)(1-v)}{N+v-1} \alpha^N(2, v) &= 0 \\ \frac{(N-i)}{2(N-i)+iv} \alpha^N(i-1, v) - \alpha^N(i, v) + \frac{(N-i)(1-v)}{2(N-i)+iv} \alpha^N(i+1, v) &= 0, \quad 2 \leq i \leq N-2, \\ -\alpha^N(N-1, v) + \frac{1}{2+(N-1)v} \alpha^N(N-2, v) &= -\frac{1-v}{2+(N-1)v}. \end{aligned}$$

By multiplying each equation with the corresponding denominator, one gets an equivalent system of equations $P' \tilde{\alpha}^N = b$ with a $(N-1) \times (N-1)$ matrix P' and $\tilde{\alpha}^N := (\alpha^N(i, v))_{i=1, \dots, N-1}$. This linear system of equations looks in tableau form as follows.

	$\alpha^N(1, v)$	$\alpha^N(2, v)$...	$\alpha^N(N-1, v)$	1
1	$-(N+v-1)$	$(N-1)(1-v)$	\dots	0	0
2	$(N-2)$	$-2(N-2)-2v$	\ddots	0	0
3	0	$(N-3)$	\ddots	\vdots	\vdots
\vdots	\vdots	\ddots	\ddots	0	\vdots
\vdots	\vdots	\ddots	\ddots	$2(1-v)$	0
$N-1$	0	\dots	1	$-2-(N-1)v$	$-(1-v)$

We are interested in the absorption probability $\alpha^N(1, v)$, i.e. the probability of getting absorbed in state N when the process is started with a single type-I cell. We use CRAMER'S rule and calculate the absolute value of the determinants to facilitate the calculations since the resulting probability is non-negative. In detail, we derive

$$\alpha^N(1, v) = \frac{|\det P'_1|}{|\det P'|},$$

where P'_1 is the matrix formed by replacing the first column of P' by the column vector b .

We calculate $|\det P'|$ first. By induction over N the general structure can be inferred. For $N = 4$ it holds that

$$\begin{aligned} |\det P'| &= \left| \det \begin{pmatrix} -(3+v) & 3(1-v) & 0 \\ 2 & -4-2v & 2(1-v) \\ 0 & 1 & -2-3v \end{pmatrix} \right| \\ &= 6(v^3 + 9v^2 + 9v + 1) \\ &= 3!(v^3 + 3^2v^2 + 3^2v + 1), \end{aligned}$$

and for $N = 5$

$$\begin{aligned} |\det P'| &= \left| \det \begin{pmatrix} -(4+v) & 4(1-v) & 0 & 0 \\ 3 & -6-2v & 3(1-v) & 0 \\ 0 & 2 & -4-3v & 2(1-v) \\ 0 & 0 & 1 & -2-4v \end{pmatrix} \right| \\ &= 24(v^4 + 16v^3 + 36v^2 + 16v + 1) \\ &= 4!(v^4 + 4^2v^3 + 6^2v^2 + 4^2v + 1). \end{aligned}$$

Furthermore,

$$\begin{aligned} |\det P'| &= 120(v^5 + 25v^4 + 100v^3 + 100v^2 + 25v + 1) \\ &= 5!(v^5 + 5^2v^4 + 10^2v^3 + 10^2v^2 + 5^2v + 1) \text{ for } N = 6. \end{aligned}$$

Therefore, we conclude that the general form of $|\det P'|$ is given by

$$\begin{aligned} |\det P'| &= (N-1)! \left(\binom{N-1}{N-1}^2 v^{N-1} + \binom{N-1}{N-2}^2 v^{N-2} + \dots + \binom{N-1}{1}^2 v^1 + \binom{N-1}{0}^2 v^0 \right) \\ &= (N-1)! \sum_{i=0}^{N-1} \binom{N-1}{i}^2 v^i \\ &= (N-1)! P_{N-1} \left(\frac{v+1}{1-v} \right) (1-v)^{N-1}, \quad N \in \mathbb{N}. \end{aligned} \tag{3.7}$$

Here, P_N denotes the LEGENDRE polynomials [1] which are the particular solutions to the LEGENDRE differential equation

$$(1-x^2) f''(x) - 2x f'(x) + N(N+1) f(x) = 0, \quad N \in \mathbb{N}_0.$$

The determinant of the other matrix P'_1 given by

$$\begin{pmatrix} 0 & (N-1)(1-v) & 0 & \cdots & \cdots & 0 \\ 0 & -2(N-2)-2v & (N-2)(1-v) & 0 & \ddots & 0 \\ 0 & (N-3) & -2(N-3)-3v & (N-3)(1-v) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 2(1-v) \\ -(1-v) & 0 & \cdots & \cdots & 1 & -2-(N-1)v \end{pmatrix}$$

can be calculated by applying LAPLACE expansion along the first column and evaluating the determinant of the remaining triangular matrix, i.e.

$$\begin{aligned} |\det P'_1| &= (1-v)(N-1)(1-v)(N-2)(1-v)(N-3)(1-v)\dots 2(1-v) \\ &= (N-1)!(1-v)^{N-1}. \end{aligned} \quad (3.8)$$

Using the calculated determinants from equations (3.7) and (3.8) allows the calculation of the absorption probability $\alpha^N(1, v)$ with CRAMER'S rule which yields

$$\alpha^N(1, v) = \frac{|\det P'_1|}{|\det P'|} = \frac{1}{P_{N-1}\left(\frac{v+1}{1-v}\right)}. \quad (3.9)$$

Next, we derive the absorption probability in dependency of the risk coefficient γ as the system size N tends to infinity. The definition of the risk coefficient (3.3) implies $v = \frac{\gamma^2}{N^2}$. Substitution in equation (3.9) yields

$$\alpha^N(\gamma) := \alpha^N(1, \gamma) = \frac{1}{P_{N-1}\left(\frac{v+1}{1-v}\right)} = \frac{1}{P_{N-1}\left(\frac{\frac{\gamma^2}{N^2}+1}{1-\frac{\gamma^2}{N^2}}\right)} = \frac{1}{P_{N-1}\left(\frac{N^2+\gamma^2}{N^2-\gamma^2}\right)}. \quad (3.10)$$

One important property of the LEGENDRE polynomials is the integral representation (see [1])

$$P_N(x) = \frac{1}{\pi} \int_0^\pi \left[x + \sqrt{x^2 - 1} \cos \varphi \right]^N d\varphi, \quad x \in \mathbb{R} \setminus \{-1, 1\}.$$

This integral representation is used in order to calculate the limit for the denominator in (3.10) as N goes to infinity.

$$\begin{aligned}
P_{N-1} \left(\frac{N^2 + \gamma^2}{N^2 - \gamma^2} \right) &= \frac{1}{\pi} \int_0^\pi \left[\frac{N^2 + \gamma^2}{N^2 - \gamma^2} + \sqrt{\left(\frac{N^2 + \gamma^2}{N^2 - \gamma^2} \right)^2 - 1 \cos \varphi} \right]^{N-1} d\varphi \\
&= \frac{1}{\pi} \int_0^\pi \left[\frac{N^2 + \gamma^2}{N^2 - \gamma^2} + \frac{\sqrt{(N^2 + \gamma^2)^2 - (N^2 - \gamma^2)^2 \cos \varphi}}{N^2 - \gamma^2} \right]^{N-1} d\varphi \\
&= \frac{1}{\pi} \int_0^\pi \left[\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right]^{N-1} d\varphi. \tag{3.11}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\lim_{N \rightarrow \infty} P_{N-1} \left(\frac{N^2 + \gamma^2}{N^2 - \gamma^2} \right) &= \lim_{N \rightarrow \infty} \frac{1}{\pi} \int_0^\pi \left[\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right]^{N-1} d\varphi \\
&= \frac{1}{\pi} \int_0^\pi \lim_{N \rightarrow \infty} \left[\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right]^{N-1} d\varphi. \tag{3.12}
\end{aligned}$$

The exchange of the limit and the integral is justified by using LEBESGUE'S dominated convergence theorem since there is an integrable majorant which can be derived as follows.

$$\begin{aligned}
\left| \frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right|^{N-1} &\leq \left| \frac{N^2 + \gamma^2 + 2N\gamma}{N^2 - \gamma^2} \right|^{N-1} \\
&= \left| \frac{(N + \gamma)^2}{(N - \gamma)(N + \gamma)} \right|^{N-1} \\
&= \left| \frac{N + \gamma}{N - \gamma} \right|^{N-1}
\end{aligned}$$

Moreover, it holds that

$$\begin{aligned}
\left| \frac{N + \gamma}{N - \gamma} \right|^{N-1} &= \left| \left(1 + \frac{2\gamma}{N - \gamma} \right) \right|^{N-1} \\
&= \left| \left(1 + \frac{2\gamma}{N - \gamma} \right)^{N-\gamma} \left(1 + \frac{2\gamma}{N - \gamma} \right)^{\gamma-1} \right| \\
&\leq \exp(2\gamma) \cdot 2
\end{aligned}$$

for N sufficiently large and since the sequence $\left(1 + \frac{1}{N}\right)^N$ is monotonously increasing. The limit in (3.12) can be calculated as follows.

$$\begin{aligned} \lim_{N \rightarrow \infty} \left[\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right]^{N-1} &= \lim_{N \rightarrow \infty} \exp \left((N-1) \ln \left(\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right) \right) \\ &= \exp \left(\lim_{N \rightarrow \infty} \frac{\ln \left(\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right)}{\frac{1}{N-1}} \right) \end{aligned}$$

Here, we can apply L'HOSPITAL'S rule for the calculation of the limit, i.e.

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{\ln \left(\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right)}{\frac{1}{N-1}} &= \lim_{N \rightarrow \infty} \frac{\frac{2N + 2\gamma \cos \varphi}{N^2 + \gamma^2 + 2\gamma N \cos \varphi} - \frac{2N}{N^2 - \gamma^2}}{-\frac{1}{(N-1)^2}} \\ &= \lim_{N \rightarrow \infty} - (N-1)^2 \frac{(2N + 2\gamma \cos \varphi)(N^2 - \gamma^2) - 2N(N^2 + \gamma^2 + 2\gamma N \cos \varphi)}{(N^2 + \gamma^2 + 2\gamma N \cos \varphi)(N^2 - \gamma^2)} \\ &= \lim_{N \rightarrow \infty} - \frac{(N-1)^2 - 2\gamma N^2 \cos \varphi - 4\gamma^2 N - 2\gamma^3 \cos \varphi}{N^2 - \gamma^2} \frac{1}{N^2 + \gamma^2 + 2\gamma N \cos \varphi} \end{aligned}$$

This representation allows to determine the limits of both factors.

$$\begin{aligned} \lim_{N \rightarrow \infty} - \frac{(N-1)^2 - 2\gamma N^2 \cos \varphi - 4\gamma^2 N - 2\gamma^3 \cos \varphi}{N^2 - \gamma^2} \frac{1}{N^2 + \gamma^2 + 2\gamma N \cos \varphi} &= (-1)(-2\gamma \cos \varphi) \\ &= 2\gamma \cos \varphi. \end{aligned}$$

Hence, we obtain for (3.12)

$$\begin{aligned} \frac{1}{\pi} \int_0^\pi \lim_{N \rightarrow \infty} \left[\frac{N^2 + \gamma^2 + 2N\gamma \cos \varphi}{N^2 - \gamma^2} \right]^{N-1} d\varphi &= \frac{1}{\pi} \int_0^\pi \exp(2\gamma \cos \varphi) d\varphi \\ &= I_0(2\gamma), \end{aligned}$$

where I_n denotes the modified BESSEL function of the first kind. This function can be expressed by the series

$$I_n(x) = \sum_{m=0}^{\infty} \frac{1}{m! \Gamma(m+n+1)} \left(\frac{x}{2}\right)^{2m+n} \quad (3.13)$$

where $\Gamma(x)$ denotes the Gamma function, see [1] for details. Equation (3.11) implies that

$$\alpha(\gamma) = \lim_{N \rightarrow \infty} \alpha^N(1, \gamma) = \lim_{N \rightarrow \infty} \frac{1}{P_{N-1} \left(\frac{N^2 + \gamma^2}{N^2 - \gamma^2} \right)} = \frac{1}{I_0(2\gamma)}. \quad (3.14)$$

□

N	u	v	γ	$\alpha^N(1, v)$	$\alpha(\gamma)$	fraction of absorption
10	10^{-4}	2500^{-1}	0.2	0.9649	0.96117	0.9632
100	10^{-4}	2500^{-1}	2	0.09	0.0885	0.0921
10	10^{-4}	10000^{-1}	0.1	0.9911	0.9901	0.9912
100	10^{-4}	10000^{-1}	1	0.4417	0.4387	0.4419
10	10^{-4}	1.52×10^{-3}	0.39	0.8759	0.8636	0.882
100	10^{-4}	1.52×10^{-5}	0.39	0.8649	0.8636	0.8665

Table 3: Comparison of absorption probabilities (3.9) and the results of 10000 simulations of trajectories of the TGP process.

A plot of the asymptotic absorption probability (3.5) is provided in Figure 6. A comparison between simulation results produced by sampling trajectories of the TGP process, absorption probabilities obtained by using the exact formula (3.4) and the asymptotic equation (3.5) is provided in Table 3.

3.2.3 Derivation of tumor regression functions for the TGP process

DESCRIPTION OF TUMOR REGRESSION. We are interested in the regression probability of a partially resected PA-I tumor in dependency of the remaining tumor size and assume that regression of a residual tumor is achieved if no tumor cells are present anymore. All suggested mechanisms of tumor regression influence the ratio of tumor and wild-type cell birth and death rates. Therefore, we assume that competition between tumor and wild-type cells leads to tumor regression which is incorporated by MORAN dynamics with relevant cell number equal to N again, see also Figure 7. Furthermore, we assume that the partial resection reduces the residual number of PA-I cells below the critical tumor size N .

Figure 7 provides a graphical representation of tumor regression in the TGP process. The tumor regression function describes the extinction probability of tumor cells, i.e. the probability to reach state 0 when starting the TGP process in some state k with $1 \leq k \leq N - 1$. For $v = 0$, our TGP process simplifies to a neutral two-type MORAN process in which the extinction probability is an established result and equals $1 - \frac{k}{N}$ [71]. Here, we derive this extinction probability for our TGP process with three cell types. For the mathematical analysis, it is convenient to express this function in terms of $\rho = \frac{k}{N}$. The fraction ρ describes the ratio between the residual number of PA-I cells after partial resection k , $1 \leq k \leq N - 1$, and the critical tumor size N . Formally, the tumor regression function of the TGP process is defined as follows.

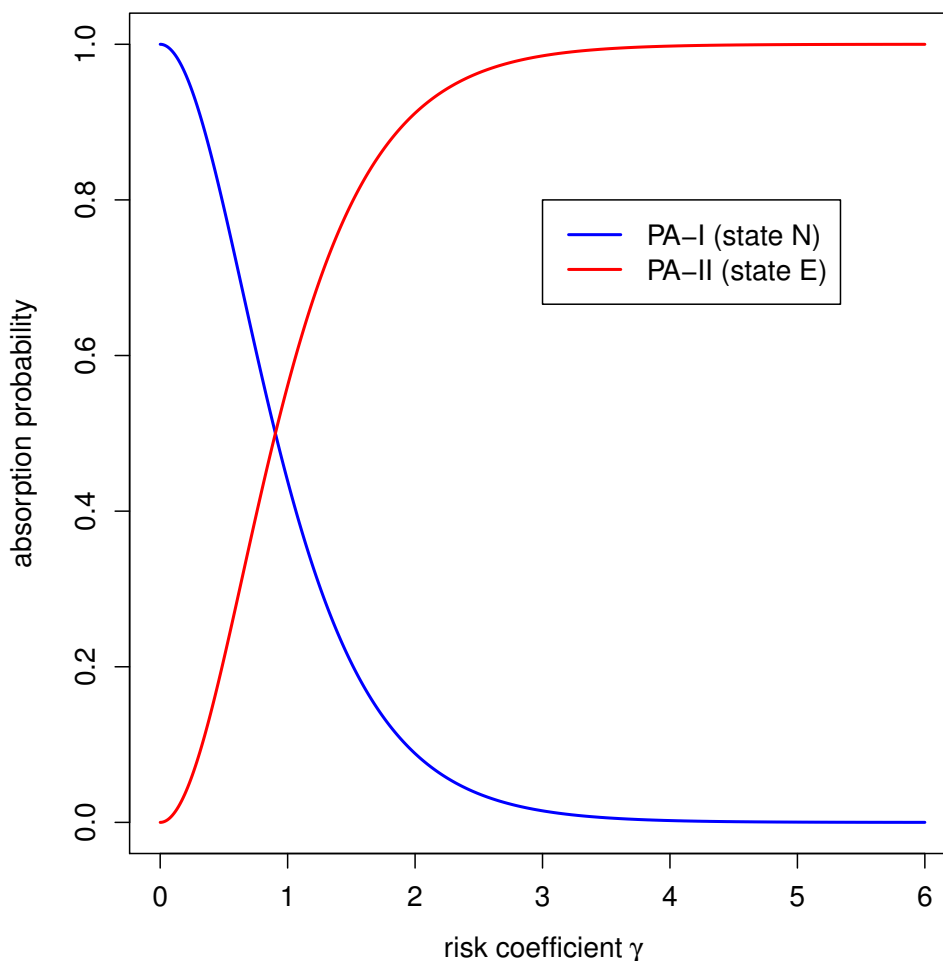


Figure 6: The absorption probabilities in the absorbing states N and E of the **TGP** process in dependency of the risk coefficient γ .

3.5 Definition (Tumor regression function) The tumor regression function β_γ^N describing the probability of tumor regression in the **TGP** process in dependency of the residual critical tumor fraction ρ is defined as

$$\beta_\gamma^N(\rho) := \mathbb{P}(X_t = 0 \text{ for some } t \geq 0 | X_0 = N\rho), \quad \rho \in [0, 1]. \quad (3.15)$$

APPROXIMATING THE TUMOR REGRESSION FUNCTION. The tumor regression function $\beta_\gamma^N(\rho)$ can be estimated by a diffusion approximation of the **TGP** process. In order to achieve this a derivation in [29] is used and extended. There, it is shown that it suffices to investigate a modified process $(Y_t)_{t \geq 0}$. This process is determined by the original rates given by equation (3.1) with the following modification. The rate for a type-I mutation equals $u = 0$ in $q(k, l)$ if $k > 0$. This means that type-I mutations are not allowed if type-I cells are already present in the system. The modification can be justified by the assumption $u \ll \frac{1}{N}$ which

allows to treat each mutant lineage independently. See also the details of the decomposition of the TGP process into two sub-processes, Figure 5. Note that this decomposition is also used in the calculation of the absorption probabilities. There, the state space has been reduced by state 0 since the occurrence of the successful mutant is assumed at the very beginning already. For the process $(Y_t)_{t \geq 0}$, this reduction of the state space is not sensible since we want to investigate the regression probability, i.e. the probability of reaching state 0. Hence, both modified processes only differ in the possibility of reaching state 0.

Under appropriate time and space scaling $(Y_t)_{t \geq 0}$ can be asymptotically approximated for $N \rightarrow \infty$ by the WRIGHT-FISHER diffusion process Z_t on $[0, 1]$. The details of this construction can be found in [29]. The important connection between the processes $(Y_t)_{t \geq 0}$ and $(Z_t)_{t \geq 0}$ is that $Z_t = 0$ implies $X_t = 0$.

Therefore, we approximate $\beta_\gamma^N(\rho) \approx \beta_\gamma(\rho)$ where $\beta_\gamma(\rho)$ is the probability that Z_t reaches 0 when starting in ρ ,

$$\beta_\gamma(\rho) := \mathbb{P}(Z_t = 0 \text{ for some } t > 0 | Z_0 = \rho), \quad 0 \leq \rho \leq 1.$$

It holds that

$$\beta_\gamma(\rho) = c \sum_{k=1}^{\infty} \frac{\gamma^{2k}}{k!(k-1)!} (1-\rho)^k, \quad 0 \leq \rho \leq 1, \quad (3.16)$$

where the constant c is determined by the condition $\beta_\gamma(0) = 1$. See [29, Lemma 6.9] for details and a rigorous derivation of this approximation. Importantly, note that in [29] the parameter γ is defined as the square of our definition and we will therefore replace γ in [29] by γ^2 in the following.

Next, we express the series in equation (3.16) in terms of BESSEL functions and derive the constant c as follows. In the first step, the indices of the sum are adjusted. In the second step, we apply $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{\gamma^{2k}}{k!(k-1)!} (1-\rho)^k &= \sum_{k=0}^{\infty} \frac{1}{(k+1)!k!} (\gamma^2(1-\rho))^{k+1} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+2)} (\gamma^2(1-\rho))^{k+1} \\ &= \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+2)} (\gamma\sqrt{1-\rho})^{2k+2} \end{aligned}$$

Moreover, the definition of the modified BESSEL function of the first kind (3.13) with $x = 2\gamma\sqrt{1-\rho}$ and $n = 1$ leads to

$$\begin{aligned} \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+2)} (\gamma\sqrt{1-\rho})^{2k+2} &= \gamma\sqrt{1-\rho} \sum_{k=0}^{\infty} \frac{1}{k!\Gamma(k+2)} (\gamma\sqrt{1-\rho})^{2k+1} \\ &= \gamma\sqrt{1-\rho} I_1(2\gamma\sqrt{1-\rho}). \end{aligned}$$

Hence, equation (3.16) yields

$$\beta_\gamma(\rho) = c\gamma\sqrt{1-\rho}I_1(2\gamma\sqrt{1-\rho}).$$

Since $\beta_\gamma(0) = 1$, one can conclude that the constant c equals

$$c = \frac{1}{\gamma I_1(2\gamma)}$$

and therefore

$$\beta_\gamma(\rho) = \frac{\sqrt{1-\rho}I_1(2\gamma\sqrt{1-\rho})}{I_1(2\gamma)} \quad (3.17)$$

for $0 \leq \rho \leq 1$. The graph of β_γ is plotted in Figure 8C for different values of the risk coefficient γ .

TAYLOR EXPANSION OF THE TUMOR REGRESSION FUNCTION. In order to estimate the deviation of the specific tumor regression function for PA from a linear function, we derive the first order TAYLOR polynomial $T_1(\rho)$ at $\rho = 0.5$ of the regression function (3.17) and an estimation of the remainder term $R_1(\rho)$.

The first two derivatives are given by

$$\beta'_\gamma(\rho) = -\frac{I_1(2\gamma\sqrt{1-\rho}) + \gamma\sqrt{1-\rho}(I_0(2\gamma\sqrt{1-\rho}) + I_2(2\gamma\sqrt{1-\rho}))}{2\sqrt{1-\rho}I_1(2\gamma)},$$

$$\beta''_\gamma(\rho) = \gamma^2 \frac{I_1(2\gamma\sqrt{1-\rho})}{I_1(2\gamma)\sqrt{1-\rho}}.$$

Here, we utilized that

$$\begin{aligned} \frac{d}{dx}I_0(x) &= I_1(x) \quad \text{and} \\ \frac{d}{dx}I_m(x) &= \frac{I_{m-1}(x) + I_{m+1}(x)}{2} = I_{m-1}(x) - \frac{mI_m(x)}{x}, \quad m \in \mathbb{N}, \end{aligned}$$

see [1]. Therefore, for $\rho_0 = 0.5$, it holds that

$$T_1(\rho) = \frac{I_1(\sqrt{2}\gamma)}{\sqrt{2}I_1(2\gamma)} - \frac{\sqrt{2}I_1(\sqrt{2}\gamma) + \gamma(I_0(\sqrt{2}\gamma) + I_2(\sqrt{2}\gamma))}{2I_1(2\gamma)}(\rho - 0.5). \quad (3.18)$$

For the remainder term it holds for some ξ between 0.5 and ρ that

$$\begin{aligned} |R_1(\rho)| &= \left| \frac{\beta''_\gamma(\xi)}{2}(\rho - 0.5)^2 \right| = \left| \frac{\gamma^2 I_1(2\gamma\sqrt{1-\xi})}{I_1(2\gamma)\sqrt{1-\xi}} \frac{(\rho - 0.5)^2}{2} \right| \\ &\leq \gamma^2 \frac{(0 - 0.5)^2}{2} \\ &= \frac{\gamma^2}{8}. \end{aligned} \quad (3.19)$$

for all $0 \leq \rho \leq 1$ since $\beta''_\gamma(\xi)$ is monotonously decreasing.

3.3 RESULTS & CLINICAL IMPLICATIONS

3.3.1 Derivation of the PA-regression-function

The regression function (3.17) depends on the parameters of the TGP process via the risk coefficient γ , see equation (3.3). This parameter is estimated such that the clinically observed fraction of PA-I tumors, denoted by \hat{p} , equals the theoretically obtained fraction $\alpha(\gamma)$ of absorption in state N in the TGP process. Subsequently, the derived risk coefficient is substituted into the regression function given by equation (3.17) in order to obtain the specific PA-regression-function. Figure 8 summarizes the overall strategy of this approach.

We estimate the clinically observed fraction of PA-I tumors on the basis of data reported in [114]. The authors analyzed 66 PAs with respect to their genetic profile and classified 57 cases as benign PA-I tumors and 9 cases as more aggressive PA-II tumors. This leads to

$$\hat{p} = \frac{57}{66} = 0.8636.$$

In the TGP process, this clinically observed fraction corresponds to the absorption probability in state N , given by equation (3.14). Therefore, we set

$$\alpha(\hat{\gamma}) = \hat{p} = 0.8636.$$

This equation allows to calculate the risk coefficient $\hat{\gamma}$ which yields

$$\hat{\gamma} = 0.39.$$

Substituting $\hat{\gamma} = 0.39$ into the regression function given by equation (3.17) allows to derive the PA-regression-function given by

$$\beta_{0.39}(\rho) = 2.3795\sqrt{1-\rho}I_1\left(0.78\sqrt{1-\rho}\right), \quad 0 \leq \rho \leq 1. \quad (3.20)$$

A plot of this function is provided in Figure 8C. This figure shows that the regression function is very robust to small alterations with respect to \hat{p} . Note that the actual risk coefficient may be smaller than the estimated value $\hat{\gamma} = 0.39$ due to the following considerations. The parameter N in our model represents a critical tumor size above which tumor regression cannot be expected anymore. However, the number of mutated cells in a diagnosed PA-I tumor may be larger than N because tumors could grow beyond this critical size without symptoms or due to a diagnostic gap between first symptoms and diagnosis. Therefore, a PA-I tumor can consist of more than N type-I cells and should have been more susceptible for progression to PA-II than accounted for in our TGP process. Hence, the risk of progression in our TGP process and therefore $\hat{\gamma}$ might be overestimated. However, this would not change the linear dependency between residual tumor size and regression probability which is discussed in the following.

3.3.2 *Linear dependency between residual tumor fraction and regression probability of PA*

We can show that the PA-regression-function (3.20) is approximately linear by utilizing a TAYLOR expansion using equation (3.18). Substituting the estimated risk coefficient of the PA-regression-function $\hat{\gamma} = 0.39$ into equation (3.18) leads to

$$T_1(\rho) = 0.9817 - \rho, \quad \rho \in [0, 1]. \quad (3.21)$$

This is a very good approximation since the remainder term can be estimated by

$$|R_1(\rho)| \leq \frac{\gamma^2}{8} = \frac{0.152}{8} = 0.0185 \quad (3.22)$$

for $\rho \in [0, 1]$. Hence, the deviation of the PA-regression-function from the linear function $T_1(\rho)$ is very small. Moreover, if the risk coefficient was overestimated, an even smaller deviation would be observed as equation (3.22) implies.

3.3.3 *Quantitative prediction of the regression probability for benign PA*

In order to provide a quantitative prediction of the regression probability given the absolute residual tumor size, we estimate the critical tumor size N in our model. Since the total cell number corresponds to the the tumor volume, we can interpret N also as minimum absolute tumor volume above which tumor regression cannot be expected anymore. The existence of this critical tumor size and its estimate of a cell number corresponding to a volume of 9 cm³ is justified in the following way. First, an extensive literature research indicated that tumor regression for residual cerebellar PA over 9 cm³ has not been reported yet, see Table 15 in the Appendix. Second, the prediction for patients with 78 cerebellar astrocytoma, including 62 PAs, has been investigated in [125]. Fig. 6 in [125] implies that the theoretical proportion of progression-free patients based on a Cox regression analysis with a residual tumor of 9 cm³ is estimated to be zero in the long-term. Finally, in [124], the role of the extent of resection in the long-term outcome of low-grade gliomas is investigated including 93 PAs. It is stated that "*the predicted outcome for patients is negatively influenced by even residual tumor volumes on the order of 10 cm³*" [124]. Incorporating the estimation for the critical tumor size of 9 cm³ into the PA-regression-function (3.20) allows to quantify our predictions, indicating that any volume reduction of one cm³ below the critical size will add 10% to the chance for regression (see also Figure 9 and Table 4).

3.3.4 *Non-existence of an extent of resection (EOR) threshold*

In malignant brain tumors it has been shown that there is an extent of resection (EOR) threshold below which no survival advantage is provided, e.g. in

RESIDUAL TUMOR SIZE (CM ³)	TUMOR REGRESSION PROBABILITY (IN %)
0.1	98.91
0.5	94.06
1	88.16
2	76.50
3	65.03
4	53.75
5	42.64
6	31.71
7	20.47
8	10.39

Table 4: Predicted regression probability for cerebellar PA based on the absolute residual tumor size.

glioblastoma this threshold is 78% [119]. The existence of different tumor zones which basically reflect tumor heterogeneity is one proposed reason for such a threshold in malignant brain tumors [31]. In contrast, our results suggest the non-existence of such a threshold in PA. This is an immediate consequence of the linear dependency between residual tumor size and regression probability. If the residual tumor is smaller than the critical tumor size N , which marks the volume for which regression cannot be expected anymore, any reduction of the tumor volume will contribute to the regression probability. Importantly, this behavior stands in contrast to a non-linear dependency which would have been obtained in our model for a higher estimated risk coefficient γ , see Figure 9

3.4 DISCUSSION & OUTLOOK

In order to gain insights into the regression behavior after partial resection of benign PA, we introduced a stochastic TGP process based on recent molecular findings, functional and clinical data. We derived a regression function that depends on the risk coefficient γ and quantifies the probability of regression in dependency of the residual tumor size. By incorporating epidemiological data on the clinically observed fractions of PA-I and PA-II cases we estimated γ and derived the specific PA-regression-function given by equation (3.20). The estimated PA-regression-function implies an approximately linear dependency between the residual critical tumor fraction and the regression probability as

illustrated in Figure 8C. This linear dependency is supported by a TAYLOR approximation and an estimation of the remainder term given by equations (3.21) and (3.22), respectively. Furthermore, we quantitatively predicted the chance for tumor regression for benign PA by estimating the critical tumor size N , see Table 4.

Our model incorporates assumptions based on clinical observations. It is observed in the clinics that PA-I tumors grow slowly, arrest in growth or even regress. Hence, type-I cells in our model proliferate without fitness advantage. Furthermore, we assume that the first type-II cell that occurs leads to an aggressive form of PA corresponding to malignant progression in PA. Alternatively, one could assign a success probability s to an emerging type-II cell which represents the probability that a single type-II cell leads to a PA-II tumor. However, it has been shown in [27] that this is equivalent to considering an analog process with type-II mutation probability sv instead of v . This alternative process would lead to the same estimated risk coefficient $\hat{\gamma}$. Therefore, the estimated PA-regression-function would not change since this function is determined only by $\hat{\gamma}$. Further, we use asymptotic results for $N \rightarrow \infty$ in order to calculate the theoretical portion of PA-I and PA-II cases in the model. This is justified by the fact that a tumor consists of billions of cells. Simulation results given in Table 3 support these asymptotic results. They show that excellent accordance with formulas for finite N is reached even for small values of N . Moreover, we could show that the model is robust against small changes in the proportion of PA-I versus PA-II tumors as shown in Figure 8C. This robustness is an important property of the model since the proportion of PA-I can vary between different studies, especially since the sample size is often very small [91, 122].

To our knowledge, the proposed model is the first theoretical attempt to predict the regression behavior of PA. In particular, we analyzed PA regression based on the population dynamics of tumor and wild-type cells. The ratio of tumor cell birth and death rates is influenced by immunologic mechanisms, hormonal factors, induction of differentiation, or apoptosis, which all could contribute to tumor regression [128]. Since PA-I tumors grow very slowly, we assumed identical birth and death rates of type-I cells in our model.

Our findings have clinically relevant implications. There is still controversy about the best treatment strategy for PA. Since PA is a slowly growing tumor and might even spontaneously regress, a wait and see strategy is an option besides more aggressive treatment strategies like radiation and chemotherapy. The decision for a more radical therapy would depend on the risk for recurrence (or even progression) and the chance of regression. However, long-term follow-up data about the probability of regression or progression after partial resection of PA is restricted and only retrospective studies with small case numbers are available [32, 91, 122, 129]. The linear dependency between residual tumor size and regression probability in our model implies that every resected percentage point of a PA-I tumor contributes equally to the regression probability. Hence, there is

no EOR threshold, but any small reduction in tumor mass provides an improvement in prognosis by increasing the probability for tumor regression. This prediction suggests a fundamentally different treatment strategy for PA compared to glioblastoma for which such a threshold has been determined [119]. Therefore, our results indicate that resection of a tumor should be aimed at even if a complete resection may not be possible. This is supported by studies showing that in patients with PA outcome depends on the extent of resection although these studies only differentiate between *biopsy*, *partial*, *subtotal* and *gross/total* resection and do not measure tumor volumes [20, 23, 125]. Moreover, if complete resection cannot be achieved our results predict that the outcome linearly depends on the residual tumor volume. If there is a reasonable chance for regression of the residual tumor it might be less justified to accept side effects by further therapies like radiation. This is an important result since the role of additional radiation therapy in treating children with tumors is highly controversial [105].

Unfortunately, as far as we know, there are no clinical studies on treatment of PA that take into account the influence of the residual tumor volume on patient outcome. We suggest that the residual tumor volume is an important prognostic marker and that a lack of sufficient volumetric data could be a reason for different results in clinical studies on additional treatment in PA. The results of this work should be further supported by future clinical studies that include volumetric data which will improve the quantitative prediction of our model and form a statistical basis for clinical decision rules.

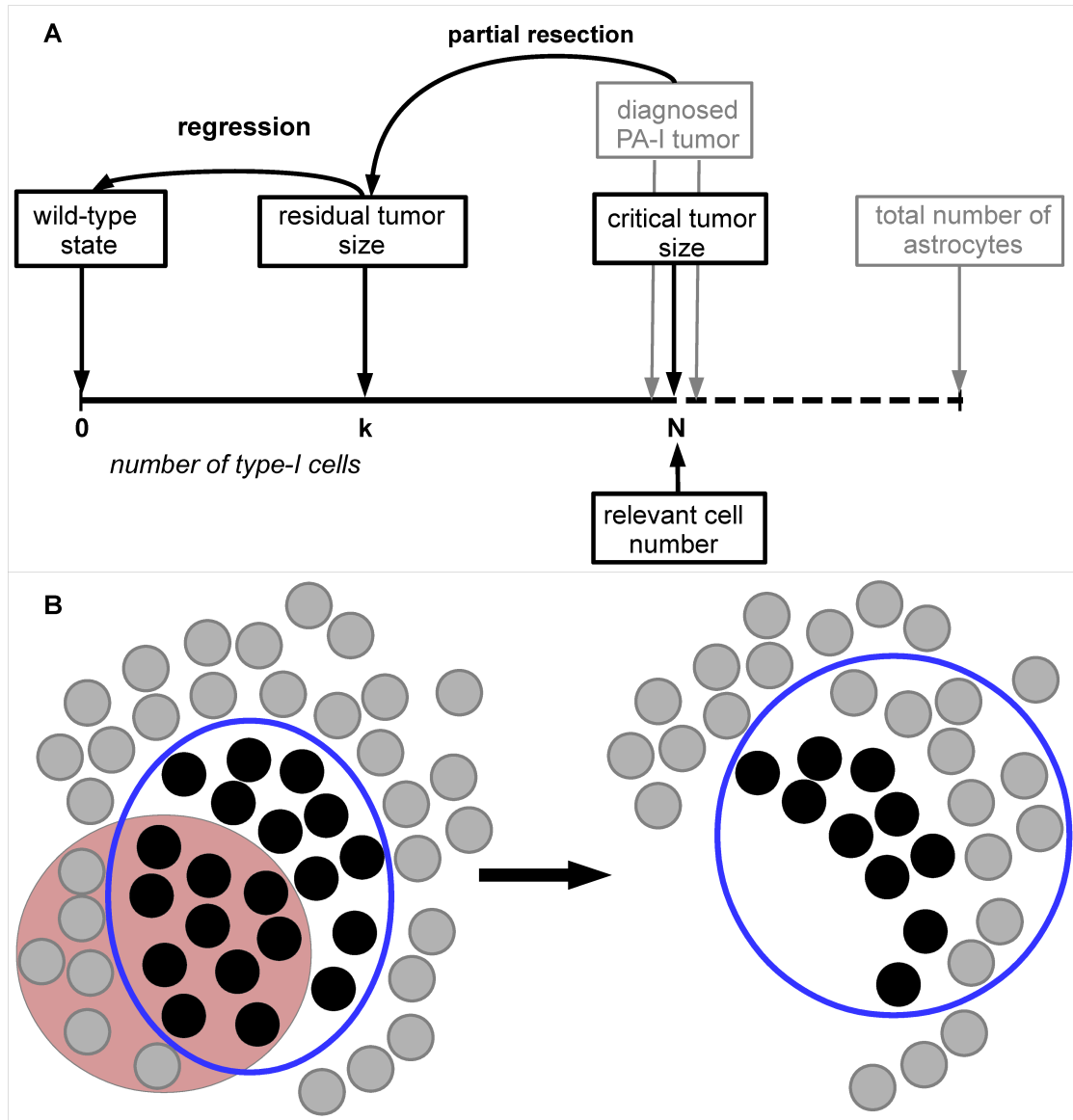


Figure 7: Tumor regression in the **TGP** process. **A**. Partial resection of a PA-I tumor reduces the number of tumor cells to size k which is assumed to be below the critical tumor size N . The residual tumor can regrow, progress or regress based on the same dynamics that led to the primary tumor. Hence, the **TGP** dynamics with relevant cell number N is utilized to describe the further development of the residual tumor. Regression is achieved if state 0 is reached, i.e. no tumor cells are present anymore. **B**. The red area indicates the resected part of the diagnosed PA-I tumor. This resection leads to removal of both tumor and wild-type cells. Subsequently, the residual number of tumor cells k competes with other wild-type cells which can lead to regrowth, regression or progression of the residual tumor. We assume that N is the relevant cell number for this competition as in the formation of the primary tumor. This relevant cell number is indicated by the blue circle.

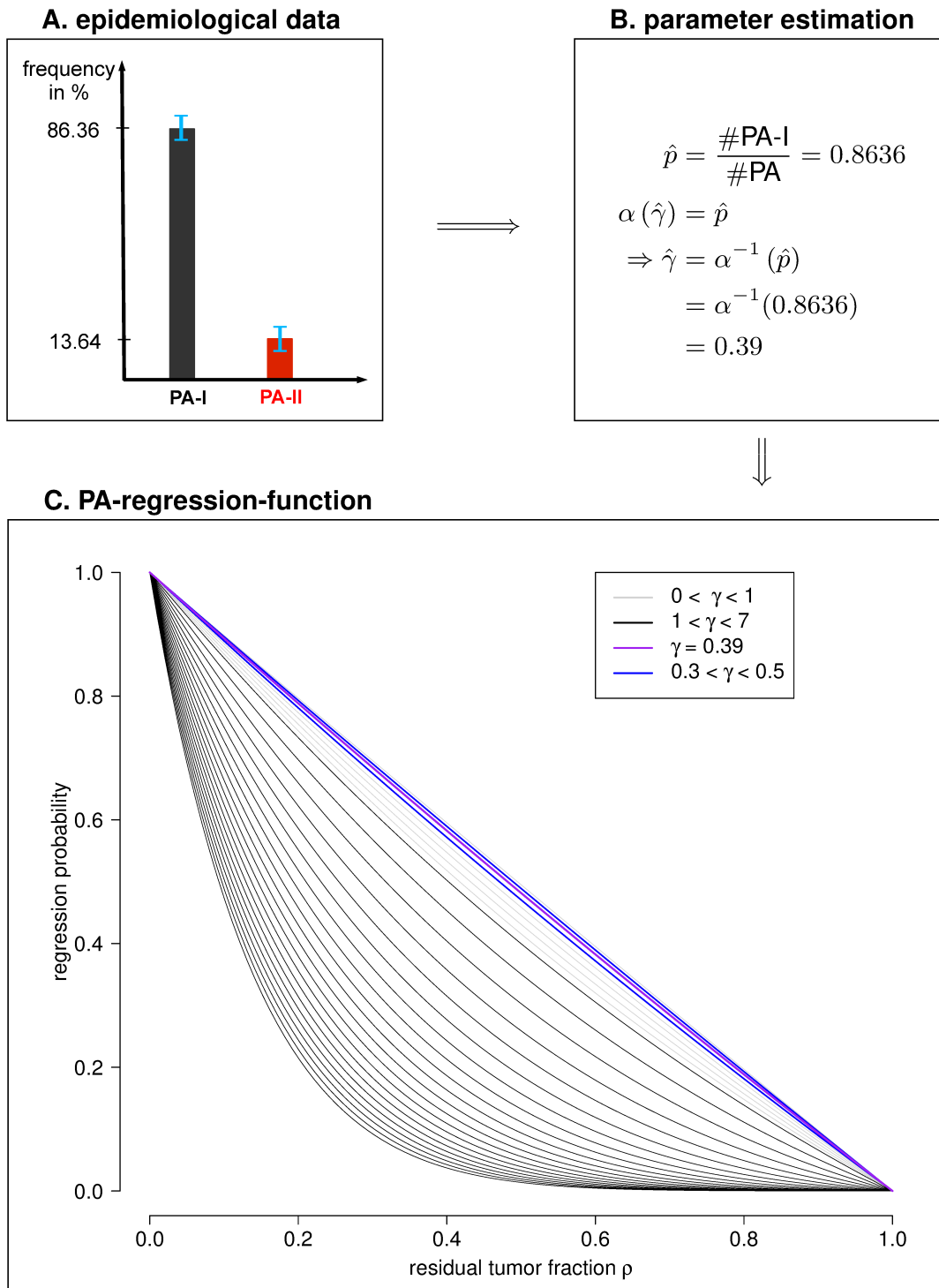


Figure 8: Parameter estimation from epidemiological data and derivation of the PA-regression-function. *A.* The clinically observed fraction of PA-I is obtained from the literature. *B.* Equation (3.14) allows to estimate the risk coefficient $\hat{\gamma}$. *C.* Substituting the estimated risk coefficient $\hat{\gamma} = 0.39$ in equation (3.17) determines the PA-regression-function $\beta_{0.39}(\rho)$ (purple). The blue plots indicate the corresponding regression functions for values of γ which are obtained within the standard deviation of \hat{p} .

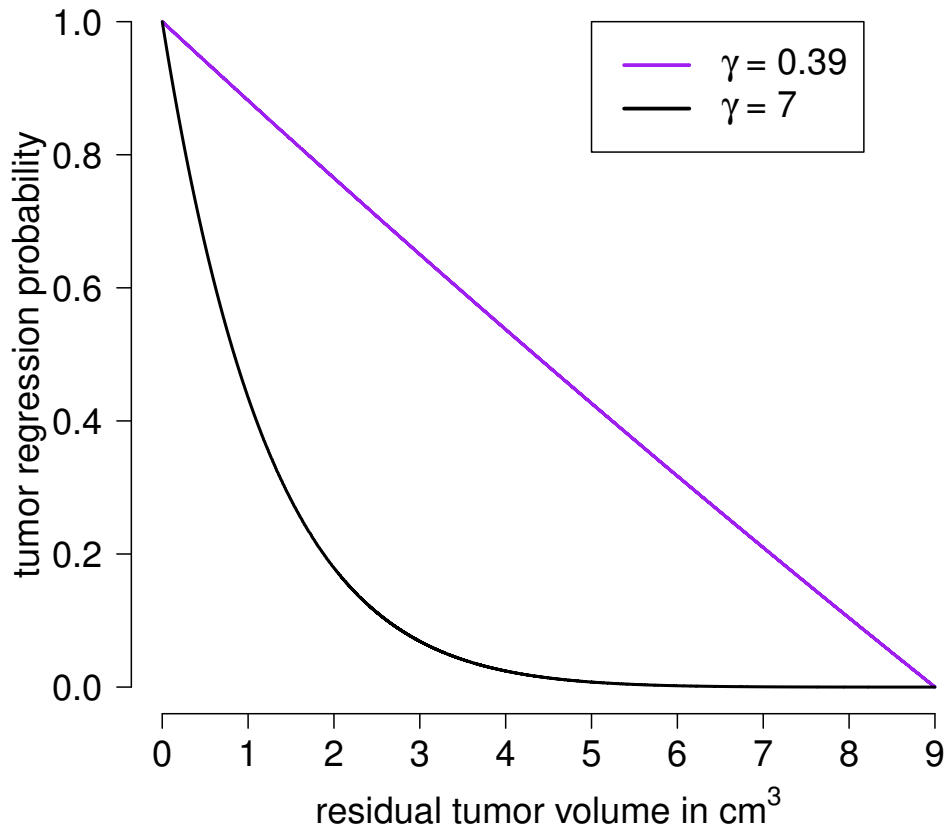


Figure 9: Non-existence of an extent of resection threshold in PA. The derived PA-regression-function (purple line) allows to quantitatively predict the regression probability based on the critical tumor size estimated as 9 cm³, see also Table 4. Roughly, one cm³ of resected tumor mass will elevate the chance of regression by further 10%. The direct consequence is the non-existence of an EOR threshold implying that any proportion of resected tumor mass will improve prognosis. This stands in contrast to the behavior of the regression function for a fictive high value of the risk coefficient of e.g. $\gamma = 7$ (black line).

4

PATTERNS OF TUMOR PROGRESSION PREDICT SMALL AND TISSUE-SPECIFIC TUMOR-ORIGINATING NICHE^{*}

4.1 TUMOR PROGRESSION ON THE CELLULAR AND THE TISSUE SCALE

Cancer development is a multistep process in which cells acquire a certain number of progressive epigenetic and genetic alterations [138]. This multistep process can be divided into a neutral and a selection phase. In the neutral phase, the epigenetic and genetic alterations do not confer a proliferative fitness advantage to the tumor precursor cells whereas cells gain such an advantage in the selection phase [144, 146]. A single genetically altered cell does not necessarily induce tumor formation but is rather exposed to competition with its corresponding wild-type cells [2]. The realization of this competition depends on the tissue and cell type. It can be direct due to replacement of a cell by the offspring of another proliferating cell [2] but also indirect, for example by symmetric and asymmetric division of cells [7]. Importantly, tissues are composed of different types of cells but only those which are capable to give rise to a progeny able to accumulate alterations can be tumor-originating cell types [84]. Tumor-originating cell refers to the wild-type cell of a certain type that acquires the first alteration in the multistep process of cancer development. Within the neutral phase, the progeny of the tumor-originating cell competes with wild-type cells within normal tissue homeostasis. Because this competition is controlled by the original tissue organization, the range of this competition is determined by the tissue structure which provides natural spatial boundaries for the spread of the progeny of the tumor-originating cell [76, 113].

In order to induce tumor formation, the progeny of the tumor-originating cell must not go extinct but has to establish within the tissue. This establishment is achieved by clonal expansion to a sufficiently large cell population [75]. For some tissues, there is experimental evidence that this establishment is characterized by an outcompetition of wild-type cells within the homeostatic

* This chapter is based on a manuscript submitted to *Nature Communications*: Thomas Buder, Andreas Deutsch, Barbara Klink, and Anja Voss-Böhme. Author contribution: Thomas Buder wrote the paper together with Andreas Deutsch, Barbara Klink and Anja Voss-Böhme. Andreas Deutsch, Barbara Klink and Anja Voss-Böhme supervised the study and gave substantial input to the manuscript. Thomas Buder conceived, designed and analyzed the model.

range of competition. For example, the tumor-originating cell within the human colon has been identified to be almost always a stem cell with a first hit in the APC gene, and a second hit in this gene is sufficient to induce adenoma formation, a benign precursor of malignant adenocarcinoma. These stem cells reside at the bottom of so-called niches within colonic crypts and are capable of self-renewal and multilineage differentiation [55]. It has been demonstrated that tumor-originating cells neutrally compete with wild-type stem cells for a position within the spatially restricted stem cell niche [7]. Either such an altered stem cell goes extinct due to this competition or eventually replaces all wild-type stem cells within the stem cell niche. This process has been termed monoclonal conversion and represents almost always the first step of tumor formation within the human colon [55]. Hence, the monoclonal conversion of the stem cell niche by the progeny of the tumor-originating cell with loss of the APC gene induces the establishment of an adenoma on the tissue scale. However, in other tissues the early phase of tumor development on the cellular scale is less understood. The main reason is a lack of knowledge regarding the tumor-originating cell type. Similar to the colon, it has been shown that stem cells within the hematopoietic system represent the tumor-originating cell type [41, 136]. In contrast, there is also evidence that non-stem cells can be the tumor-originating cell type, e.g. in oligodendroglioma [109]. Although the lineage in which cancer originates has been revealed for skin, pancreatic, brain and breast tumors, the tumor-originating cell type remains elusive in most cases [137]. Its identification may allow earlier detection of malignancies and may lead to preventive therapies for individuals at high risk of developing cancer [137].

On the tissue scale, one observes different types of tumor progression. Tumors can progress sequentially, i.e. with a clinically detectable benign precursor stage. Alternatively, they can also progress by tunneling without such a prior benign precursor stage. Epidemiological data allow to infer the progression patterns with respect to the ratios of tunneling versus sequential progression of different tumors. Interestingly, these progression patterns differ largely between tissues. Some tumors exhibit predominantly sequential progression, e.g. benign adenoma almost always develop prior to adenocarcinoma in the colon [55]. Similarly, multiple myeloma are in almost all cases preceded by a premalignant state called monoclonal gammopathy of undetermined significance (MGUS) [77]. In contrast, glioblastoma develops in 90% of all cases without evidence of a less malignant precursor lesion (primary glioblastoma) and progresses in 10% of all cases from low-grade tumors (secondary glioblastoma) [106]. In which way these progression patterns on the tissue scale emerge from the multistep process of cancer development on the cellular scale is difficult to infer since the early phase of this multistep process is hardly observable.

In this work, we use observables on the tissue scale to shed light on the early cellular processes of tumor development. We utilize a MORAN model with mutations [14, 26, 98] to describe cellular competition between wild-type cells and

tumor cells. Benign and malignant tumor subtypes on the tissue scale are represented by two absorbing states within the model. We incorporate epidemiological data on the progression patterns of cancers to calibrate the model. By analyzing the model dynamics with respect to different spatial cell arrangements, we obtain a lower and upper bound for the critical number of tumor cells needed for tumor development on the tissue scale. Interestingly, our estimates are considerably small, tissue-specific and far away from the overall number of cells in a clinically observable tumor. We therefore propose that the fate of tumor development is decided in tissue-specific tumor-originating niches. This proposal is supported by our estimate of the tumor-originating niche size for the human colon which agrees well with the size of the stem cell niche in colonic crypts. In particular, we propose that a tumor-originating niche size of 300 – 1900 cells within the human brain can explain the ratio of primary and secondary glioblastoma. Interestingly, our estimates also agree well with the minimal number of tumor cells needed for tumor formation in mice injection experiments and might allow to infer the tumor-originating cell type.

4.2 MATHEMATICAL MODEL

STATE SPACE AND REPRESENTATION OF BENIGN AND MALIGNANT TUMOR SUBTYPES. The multistep process in which cancer cells increase gradually in malignancy differs with respect to the number of steps, e.g. two steps in retinoblastoma [69] compared to seven steps in colon cancer [139]. In our cell-based model, we only regard the last step within the neutral phase and the first step within the selection phase such that we obtain a two-step process. This coarse-grained approach is appropriate for our purpose since we are only interested in modeling tumor progression patterns and not quantities which are largely influenced by the precise number of steps, e.g. the time-scale of tumor development or intra-tumor heterogeneity. In the cellular two-step process, genetic or epigenetic alterations can transform *wild-type cells* into *benign tumor cells* which can further progress to *malignant tumor cells*. We assume that the benign progeny of the tumor-originating cell competes with wild-type cells and can clonally expand within normal tissue homeostasis. The parameter N in our model describes the homeostatic range of this competition. We further assume that a benign tumor on the tissue scale will develop if monoclonal conversion of wild-type cells by benign tumor cells within the homeostatic range of competition N is achieved. This assumption is based on experimental observations for example within the colon where mutant cells either go extinct or fixate in the colonic stem cell niche [7]. Moreover, mice injection experiments indicate that a critical number of tumor cells is needed for tumor formation [8, 16, 42, 46, 54, 81, 104, 149] which suggests a point of no return on the cellular scale for tumor formation. If the first benign tumor cell progresses to a malignant tumor cell

we assume that a malignant tumor will inevitably develop on the tissue scale which reflects a high fitness advantage of malignant tumor cells.

The state space of the underlying stochastic process is $S = \{0, 1, 2, \dots, N, E\}$ where states 0 to N represent the occurrence of the respective number of benign cells without the occurrence of malignant cells. State E indicates the presence of a malignant cell. Therefore, states N and E correspond to emergence of benign and malignant tumor subtypes and therefore to sequential and tunneling tumor progression on the tissue scale, see also Figure 10. Both states N and E are absorbing states of the underlying process.

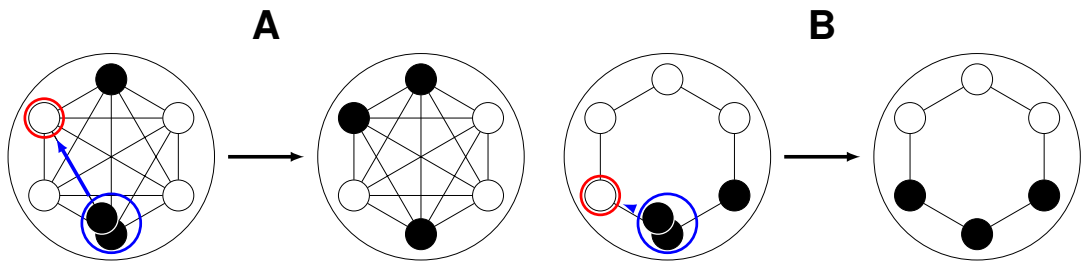


Figure 10: MORAN dynamics with different spatial cell arrangements. In the MORAN dynamics, a randomly chosen cell proliferates (blue circle) and replaces a neighboring cell which undergoes cell death (red circle). In A, the space-free dynamics is illustrated, i.e. each cell can be replaced by any other cell. In B, only neighboring cells can be replaced reflecting a one-dimensional cell arrangement.

DYNAMICS IN THE MODEL. In order to describe competition between cells and tumor cell progression, we use a MORAN model with mutations. This model class has often been used to investigate tumor growth and progression, e.g. in Chapter 3 of this thesis. MORAN models are characterized by a fixed population of size N which represents the homeostatic range of competition in our model. The dynamics is as follows. One cell is randomly chosen for proliferation and its offspring replaces a neighboring cell which undergoes cell death. During proliferation, a genetic or epigenetic alteration can lead to tumor cell progression. Wild-type cells can progress to benign cells with probability u and benign cells progress to malignant cells with probability v . For a comprehensive introduction to MORAN models, see Section 2.2. We assume that initially all cells are wild-type cells. Hence, the process starts in state 0.

4.2.1 Model analysis

CHOICE OF SPATIAL CELL ARRANGEMENT. As theoretical studies demonstrate, the interplay between tissue structure, the population size N and muta-

tion probabilities u and v in MORAN models are crucial for the dynamics of the model [25, 26, 30]. In particular, it has been shown that the absorption probability in state N on regular structures is the highest if all cells can potentially compete with each other and the lowest for a one-dimensional cell arrangement [26]. Since the tumor-originating cell type is unknown for most cancers also the spatial cell arrangement and realization of competition is unknown [2, 61]. Therefore, we consider a space-free and a one-dimensional cell arrangement in order account for this uncertainty by deriving a lower and an upper bound for the absorption probabilities.

Figure 10 illustrates the MORAN dynamics on these two different structures. The space-free model has already been introduced and analyzed in Chapter 3. Here, we introduce the 1D model within a biologically plausible parameter regime.

DEFINITION OF THE 1D MODEL. The one-dimensional model is a spatial three-type MORAN model with mutation probabilities u and v in which each cell can only be replaced by the offspring of the two neighboring cells, see Figure 10. For the analysis, we assume that

$$Nu \ll \sqrt[3]{v}. \quad (4.1)$$

Analogously to the analysis of the space-free MORAN model in Chapter 3, assumption (4.1) guarantees that mutations to benign cells are rare such that each arising benign cell mutant can be investigated independently, see [26]. Hence, for the analysis we can neglect mutations to benign cells if the system is already in a state $k > 0$. We only consider the last benign mutant which eventually leads to absorption in states N or E . Therefore, we study a modified process with $u = 0$ conditioned that this benign clone will not go extinct. Note that if there are any benign cells in the system, then there is always exactly one connected benign cell population, see Figure 10B. Formally, this modified one-dimensional model can be defined as follows.

4.1 Definition The modified 1D MORAN model under assumption (4.1) is a continuous time MARKOV chain $(X_t)_{t \geq 0}$ with $X_0 = 1$ on $S = \{1, 2, 3, \dots, N, E\}$, where states $1, \dots, N$ represent the number of benign cells in the population when no

malignant cell is present and state E represents the existence of a malignant tumor cell, and rate matrix $Q = (q(k, l))_{k, l \in S}$ which reads

$$q(k, l) = \begin{cases} 1 - v, & 1 \leq k \leq N - 1, l = k + 1, \\ 1, & 2 \leq k \leq N - 1, l = k - 1, \\ kv, & 1 \leq k \leq N - 1, l = E, \\ - \sum_{\substack{m \in \tilde{S} \\ m \neq k}} q(k, m), & l = k, \\ 0, & \text{else.} \end{cases} \quad (4.2)$$

The rate $q(k, k + 1)$ for an increase of the state is composed as follows. For an increase of the state, the offspring of a benign cell must replace a wild-type cell.

- If $k = 1$, i.e. there is only one benign cell in the system, then exactly this cell has to be selected to reproduce. The neighboring cell that is replaced is always a wild-type cell. However, during reproduction, there must not be a mutation which implies a rate of $1 - v$.
- If $k > 1$, then only the two benign cells at the boundary of the connected benign cell population have to be selected for reproduction since otherwise the offspring would replace a benign cell. One of them is selected with rate 2. The chosen cell has two neighbors, a benign and a malignant cell. Thus, the wild-type cell is chosen to be replaced with probability $\frac{1}{2}$. Finally, the offspring of the benign cell must not undergo a mutation which holds with a probability of $1 - v$. Hence, the offspring of a benign cell replaces a wild-type cell with rate $2 \cdot \frac{1}{2} \cdot (1 - v) = 1 - v$.

The other rates in (4.2) are similarly obtained. The states N and E are absorbing states of the process since $q(N, l) = q(E, l) = 0$ for $l \in S$. Note that a discrete time version of this process has been first introduced in [70].

TUMOR PROGRESSION REGIMES IN THE MODEL. Three parameter regimes within the model can be distinguished with respect to the tumor progression patterns. Within the *sequential fixation* regime, the benign tumor cell population is primarily able to reach size N before a benign tumor cell progresses to a malignant tumor cell. This regime corresponds to primarily sequential progression on the tissue scale. In the *tunneling* regime [95] a malignant clone will occur before the benign population is able to reach size N which corresponds to primarily tunneling progression in the model. In the *borderline* regime [30] both sequential fixation and tunneling are possible corresponding to both progression types on the tissue scale. An asymptotic classification of the model behavior with respect to these parameter regimes for large N has been theoretically derived in

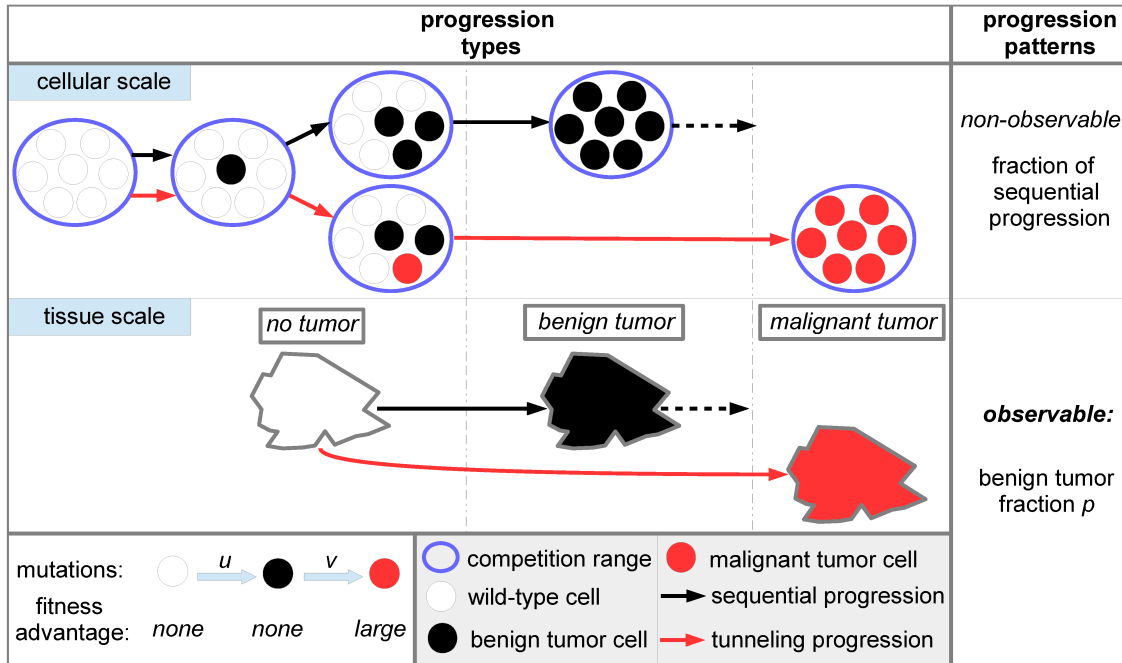


Figure 11: Tumor progression types and patterns. On the cellular scale, wild-type cells can progress to benign cells during proliferation with mutation probability u . Wild-type cells and benign cells neutrally compete with each other within the homeostatic range of competition which is modeled by MORAN dynamics, see Figure 10. A benign tumor will manifest if all cells within the homeostatic range of competition are converted to benign cells. Furthermore, benign cells can progress to malignant cells during proliferation with probability v . Then, a malignant tumor will develop. These cellular dynamics lead to two distinct progression types at the tissue scale, namely sequential progression and tunneling progression. The benign tumor fraction p determines the progression pattern. We do not model the transition from benign tumors to malignant tumors since we are only interested in the type of progression (dotted arrow).

a space-free model [101] and in lattice-like cell arrangements [25]. In [25], it has been shown that this classification for the 1D MORAN model crucially depends on the parameters N and v via the product of $N\sqrt[3]{v}$ which motivates the following definition.

4.2 Definition The 1D risk coefficient γ_{1D} is defined as

$$\gamma_{1D} := N\sqrt[3]{v}.$$

DERIVATION OF THE ABSORPTION PROBABILITIES. In Chapter 3 we already derived the absorption probability in state N of the space-free model in dependency of the corresponding risk coefficient $\gamma = N\sqrt[3]{v}$. Here, we derive the absorption probability α_{1D}^N in state N of the 1D model.

4.3 Theorem (Absorption probabilities of the 1D MORAN model) The absorption probability in state N of the 1D MORAN model introduced in Definition 4.1 starting the process in state 1 can be approximated by

$$\alpha_{1D}^N(1, v) = \frac{\pi^{-1}v(1-v)^{N-1}}{\left(J_{N+\frac{2}{v}}\left(\frac{2}{v}\right) \left((2v+1)Y_{1+\frac{2}{v}}\left(\frac{2}{v}\right) - Y_{2+\frac{2}{v}}\left(\frac{2}{v}\right) \right) + Y_{N+\frac{2}{v}}\left(\frac{2}{v}\right) \left(J_{2+\frac{2}{v}}\left(\frac{2}{v}\right) - (2v+1)J_{1+\frac{2}{v}}\left(\frac{2}{v}\right) \right) \right)} \quad (4.3)$$

where J denotes a BESSEL function of the first kind and Y denotes a BESSEL function of the second kind.

Proof. The first part of the proof is similar to the proof of Theorem 3.4.

For the rates (4.2) we set $q(k) := -q(k, k)$ and get

$$\begin{aligned} q(1) &= q(1, E) + q(1, 2) = 1, \\ q(N) &= q(E) = 0, \\ q(k) &= q(k, k+1) + q(k, k-1) + q(k, E) = 2 + v(k-1), \quad 2 \leq k \leq N-1. \end{aligned}$$

We further regard the embedded MARKOV chain with transition probabilities

$$p(i, j) = \begin{cases} \frac{q(i, j)}{q(i)}, & i \neq j, \\ 1, & i = j = E, \\ 1, & i = j = N, \\ 0, & \text{else,} \end{cases}$$

in which the entries unequal to 0 look as follows

$$\begin{aligned}
p(1, E) &= v, \\
p(1, 2) &= 1 - v, \\
p(k, E) &= \frac{kv}{2 + v(k-1)}, \quad 2 \leq k \leq N-1, \\
p(k, k+1) &= \frac{1-v}{2 + v(k-1)}, \quad 2 \leq k \leq N-1, \\
p(k, k-1) &= \frac{1}{2 + v(k-1)}, \quad 2 \leq k \leq N-1, \\
p(N, N) &= p(E, E) = 1.
\end{aligned}$$

The absorption probabilities for the underlying stochastic process with transition matrix $P = (p_{i,j})_{i,j \in S}$ is obtained as follows. Denote by $\alpha_{1D}^N = (\alpha_{1D}^N(i, v))_{i \in S}$ the absorption probabilities where $\alpha_{1D}^N(i, v)$ describes the absorption probability in state N starting from state i .

First step analysis yields

$$\alpha_{1D}^N(i, v) = \sum_{j \in S} p(i, j) \alpha_{1D}^N(j, v), \quad i \in S.$$

It holds that $\alpha_{1D}^N(E, v) = 0$, $\alpha_{1D}^N(N, v) = 1$ and therefore

$$\begin{aligned}
\alpha_{1D}^N(i, v) &= \sum_{j=1}^N p(i, j) \alpha_{1D}^N(j, v) \\
&= \sum_{j=1}^{N-1} p(i, j) \alpha_{1D}^N(j, v) + p(i, N) \\
&= \begin{cases} (1-v) \alpha_{1D}^N(2, v), & i = 1 \\ \frac{1}{2+v(i-1)} \alpha_{1D}^N(i-1, v) + \frac{1-v}{2+v(i-1)} \alpha_{1D}^N(i+1, v), & 2 \leq i \leq N-2, \\ \frac{1}{2+v(N-2)} \alpha_{1D}^N(N-2, v) + \frac{1-v}{2+v(N-2)}, & i = N-1. \end{cases}
\end{aligned}$$

Hence,

$$\begin{aligned}
-\alpha_{1D}^N(1, v) + (1-v) \alpha_{1D}^N(2, v) &= 0 \\
\frac{1}{2+v(i-1)} \alpha_{1D}^N(i-1, v) - \alpha_{1D}^N(i, v) + \frac{1-v}{2+v(i-1)} \alpha_{1D}^N(i+1, v) &= 0, \quad 2 \leq i \leq N-2, \\
\frac{1}{2+v(N-2)} \alpha_{1D}^N(N-2, v) - \alpha_{1D}^N(N-1, v) &= -\frac{1-v}{2+v(N-2)}.
\end{aligned}$$

By multiplying each equation with the corresponding denominator, one gets an equivalent system $P'_N \tilde{\alpha}_{1D}^N = b$ for a $(N-1) \times (N-1)$ matrix P'_N and $\tilde{\alpha}_{1D}^N := (\alpha_{1D}^N(i, v))_{i=1, \dots, N-1}$. This system reads in tableau form as follows.

$$\begin{array}{c|ccccccc|c}
 & \alpha_{1D}^N(1, v) & \alpha_{1D}^N(2, v) & \alpha_{1D}^N(3, v) & \dots & \dots & \alpha_{1D}^N(N-1, v) & 1 \\
 \hline
 1 & -1 & 1-v & 0 & 0 & \dots & 0 & 0 \\
 2 & 1 & -2-v & 1-v & 0 & \ddots & \vdots & \vdots \\
 3 & 0 & 1 & -2-2v & 1-v & \ddots & \vdots & \vdots \\
 \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 0 & \vdots \\
 \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & 1-v & 0 \\
 N-1 & 0 & \dots & \dots & 0 & 1 & -2-(N-2)v & -(1-v)
 \end{array} \tag{4.4}$$

CRAMER'S rule can be applied to derive the absorption probability in state N starting with a single type-1 cell, i.e.

$$\alpha_{1D}^N(1, v) = \frac{\det \tilde{P}'_N}{\det P'_N},$$

where \tilde{P}'_N is the matrix formed by replacing the first column of P'_N by the column vector b .

By applying LAPLACE expansion along the last column of P'_N we obtain

$$\begin{aligned}
 \det P'_N &= (-1)^{2N-2} (-2 - (N-2)v) \det P'_{N-1} \\
 &+ (-1)^{2N-3} (1-v) \begin{vmatrix} -1 & 1-v & 0 & 0 & \dots & 0 \\ 1 & -2-v & 1-v & 0 & \ddots & \vdots \\ 0 & 1 & -2-2v & 1-v & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 1 & -2-(N-4)v & 1-v \\ 0 & \dots & \dots & \dots & 0 & 1 \end{vmatrix} \\
 &= (-2 - (N-2)v) \det P'_{N-1} - (1-v) \cdot (-1)^{2N-4} \cdot 1 \cdot \det P'_{N-2} \\
 &= (-2 - (N-2)v) \det P'_{N-1} - (1-v) \det P'_{N-2},
 \end{aligned}$$

where the determinant in the first equality was evaluated by applying LAPLACE expansion along the last row. The result of these calculations is a second order difference equation with non-constant coefficients which reads as follows.

$$\begin{aligned} \det P'_N &= (-2 - (N - 2)v) \det P'_{N-1} - (1 - v) \det P'_{N-2}, \quad N \geq 4, \\ \det P'_3 &= 2v + 1, \\ \det P'_2 &= -1. \end{aligned} \tag{4.5}$$

Since the mutation rate v can be considered close to zero we can assume that $(1 - v)$ is close to one and obtain the modified equation

$$D_N = (-2 - (N - 2)v)D_{N-1} - D_{N-2}, \quad D_3 = 2v + 1, D_2 = -1, \quad N \geq 4. \tag{4.6}$$

This modified difference equation can be solved and the solution can be simplified with MATHEMATICA [56] which yields

$$\begin{aligned} D_N &= \frac{(-1)^{N+1}\pi}{v} \left(J_{N+\frac{2}{v}} \left(\frac{2}{v} \right) \left((2v + 1)Y_{1+\frac{2}{v}} \left(\frac{2}{v} \right) - Y_{2+\frac{2}{v}} \left(\frac{2}{v} \right) \right) \right. \\ &\quad \left. + Y_{N+\frac{2}{v}} \left(\frac{2}{v} \right) \left(J_{2+\frac{2}{v}} \left(\frac{2}{v} \right) - (2v + 1)J_{1+\frac{2}{v}} \left(\frac{2}{v} \right) \right) \right). \end{aligned} \tag{4.7}$$

Here, J denotes a BESSEL function of the first kind and Y denotes a BESSEL function of the second kind, see [1]. We will use the solution D_N of the modified difference equation as approximation for $\det P'_N$.

The determinant of \tilde{P}'_N is calculated as follows. The matrix \tilde{P}'_N is given by

$$\tilde{P}'_N = \begin{pmatrix} 0 & 1 - v & 0 & \dots & \dots & 0 \\ 0 & -2 - 2v & 1 - v & 0 & \ddots & 0 \\ 0 & 1 & -2 - 2v & 1 - v & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & 1 - v \\ -(1 - v) & 0 & \dots & \dots & 1 & -2 - (N - 2)v \end{pmatrix}.$$

Therefore, the determinant can be calculated by applying LAPLACE expansion along the first column and evaluating the determinant of the remaining triangular matrix, i.e.

$$\begin{aligned} \det \tilde{P}'_N &= (-1)^N \cdot (-1 - v) \begin{vmatrix} 1 - v & 0 & \dots & \dots & \dots & 0 \\ -2 - 2v & 1 - v & 0 & \ddots & \ddots & 0 \\ 1 & 2 - 2v & 1 - v & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & \ddots & \ddots & \ddots & 1 - v \end{vmatrix} \\ &= (-1)^{N+1} \cdot (1 - v)^{N-1}. \end{aligned} \tag{4.8}$$

Using this result and the approximation derived in equation (4.7) allows the approximation of the absorption probability $\alpha_{1D}^N(1, v)$ with CRAMER'S rule

$$\alpha_{1D}^N(1, v) \approx \frac{\det \tilde{P}_N}{D_N} = \frac{\pi^{-1}v(1-v)^{N-1}}{\left(J_{N+\frac{2}{v}}\left(\frac{2}{v}\right) \left((2v+1)Y_{1+\frac{2}{v}}\left(\frac{2}{v}\right) - Y_{2+\frac{2}{v}}\left(\frac{2}{v}\right) \right) + Y_{N+\frac{2}{v}}\left(\frac{2}{v}\right) \left(J_{2+\frac{2}{v}}\left(\frac{2}{v}\right) - (2v+1)J_{1+\frac{2}{v}}\left(\frac{2}{v}\right) \right) \right)}.$$

□

A comparison of this analytical approximation and simulation results from 10000 trajectories of the underlying stochastic process of the 1D MORAN model for the probability of absorption in state N is provided in Table 5.

4.3 RESULTS

THE ABSORPTION PROBABILITY IS A FUNCTION OF THE RISK COEFFICIENT γ_{1D} . Numerical analysis suggests that the absorption probability α_{1D}^N of the 1D MORAN model solely depends on the risk coefficient γ_{1D} even for small values of N . We numerically evaluated the absorption probability for fixed values of the risk coefficient for different combinations of N and v . The results are visualized in Figure 12 and indicate that α_{1D}^N can be considered as function of γ_{1D} for approximately $N \geq 40$.

THE RANGE OF COMPETITION DETERMINES TUMOR PROGRESSION PATTERNS. Our analysis allows to determine the progression patterns in both the space-free and the one-dimensional model in dependency of the competition range N . Interestingly, we find that a considerably small value of N corresponds to primarily tunneling progression in both the space-free and one-dimensional model. Moreover, the estimates of the parameter N largely depend on the considered underlying spatial cell arrangement. In particular, the smaller the number of neighboring cells, the smaller is the estimated competition range. The estimated values for a mutation probability $v = 10^{-6}$ per cell division [21] are summarized in Table 6 and visualized in Figure 13.

FATE OF TUMOR DEVELOPMENT IS DECIDED IN SMALL AND TISSUE-SPECIFIC TUMOR-ORIGINATING NICHES. Our model allows to estimate the range of cellular competition N in different human tissues. For these estimations, we calibrate the space-free and 1D model with epidemiological data on the diagnosed fraction of benign and malignant tumor subtypes. We performed an extensive literature research to obtain these data which allow to estimate the risk coefficients $\gamma = N\sqrt{v}$ and $\gamma_{1D} = N\sqrt[3]{v}$ in the following way. We equal the clinically diagnosed fraction of benign tumors p with the absorption probabilities in

N	v	simulated absorption prob.	$\alpha_{1D}^N(1, v)$ in (4.3)	deviation
10	0.0015	0.7931	0.8226	0.0295
10	0.001	0.8541	0.8761	0.0220
10	0.000001	0.9998	0.9999	0.0001
10	0.0000001	1	1	0
50	0.0015	0.0004	0.0004	0
50	0.001	0.0019	0.0022	0.0003
50	0.00001	0.8216	0.8299	0.0083
50	0.000001	0.9795	0.9806	0.0011
100	0.0015	0	0	0
100	0.001	0	0	0
100	0.00000005	0.9917	0.9920	0.0003
100	0.00000001	0.9983	0.9984	0.0001
500	0.001	0	0	0
500	0.000001	0.0019	0.0019	0
500	0.0000001	0.2429	0.2444	0.0015
500	0.00000005	0.4389	0.4407	0.0018
750	0.00001	0	0	0
750	0.0000001	0.0385	0.0388	0.0003
750	0.00000005	0.1276	0.1283	0.0007
750	0.00000001	0.5531	0.5542	0.0011
1000	0.00001	0	0	0
1000	0.000001	0	0	0
1000	0.0000001	0.0040	0.0041	0.0001
1000	0.00000001	0.3025	0.3034	0.0009

Table 5: Comparison of the analytical approximation given by equation (4.3) and simulation results from 10000 trajectories of the underlying stochastic process of the one-dimensional model for the probability of absorption in state N . The results are also visualized in Figure 12.

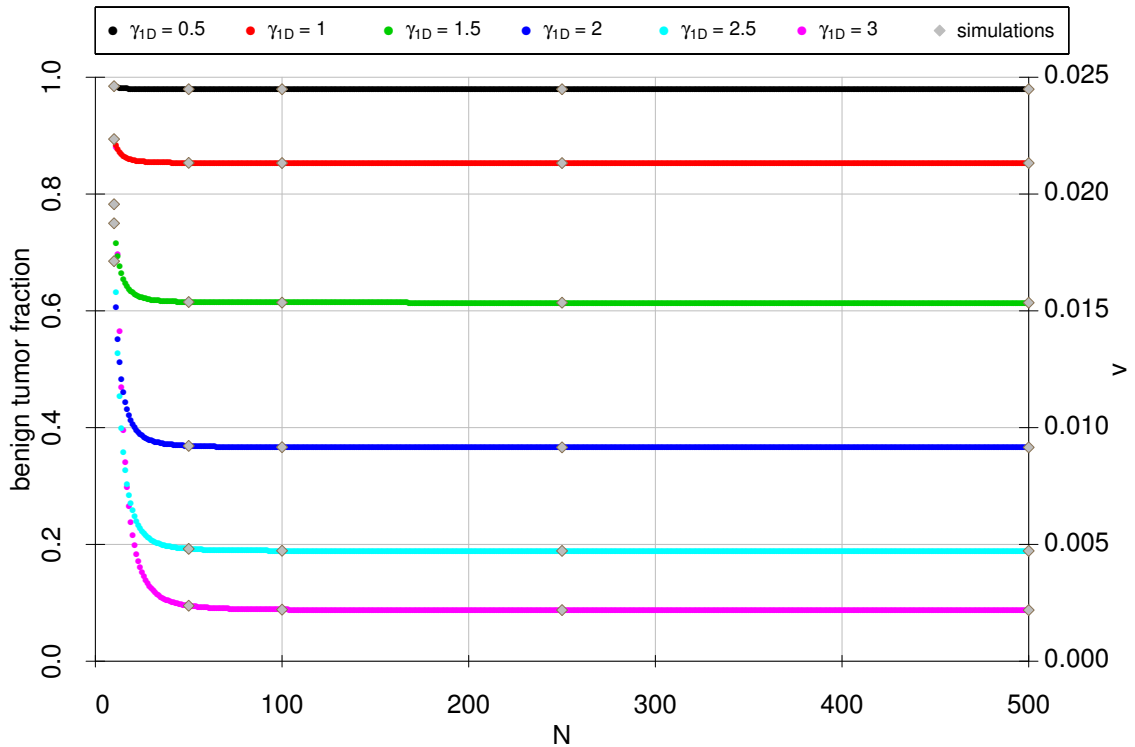


Figure 12: The absorption probability α_{1D}^N of the 1D MORAN model as function of the risk coefficient. We numerically approximated the absorption probability in state N for different values of N and v such that the risk coefficient γ_{1D} is constant. This analysis suggests that the absorption probability solely depends on the risk coefficient γ_{1D} for approximately $N \geq 40$. The squares indicate the results of simulation studies of the absorption probability in state N and therefore the benign tumor fraction in the model, see also Table 5.

progression patterns	1D model	space-free model
primarily sequential	$N \leq 17$	$N \leq 29$
both sequential and tunneling	$17 < N \leq 528$	$29 < N \leq 4530$
tunneling only	$N > 528$	$N > 4530$

Table 6: Tumor progression patterns in dependency of the spatial cell arrangement in the model. For these estimates, we have chosen $v = 10^{-6}$. Primarily sequential and primarily tunneling progression patterns refer to a fraction of 99.9% of sequential and tunneling progression, respectively.

state N . Subsequently, we numerically calculate the risk coefficients by evaluating the inverse of the absorption probability function at the diagnosed fraction of benign tumors p , i.e. $\gamma = \alpha^{-1}(p)$. The resulting estimates of the competition

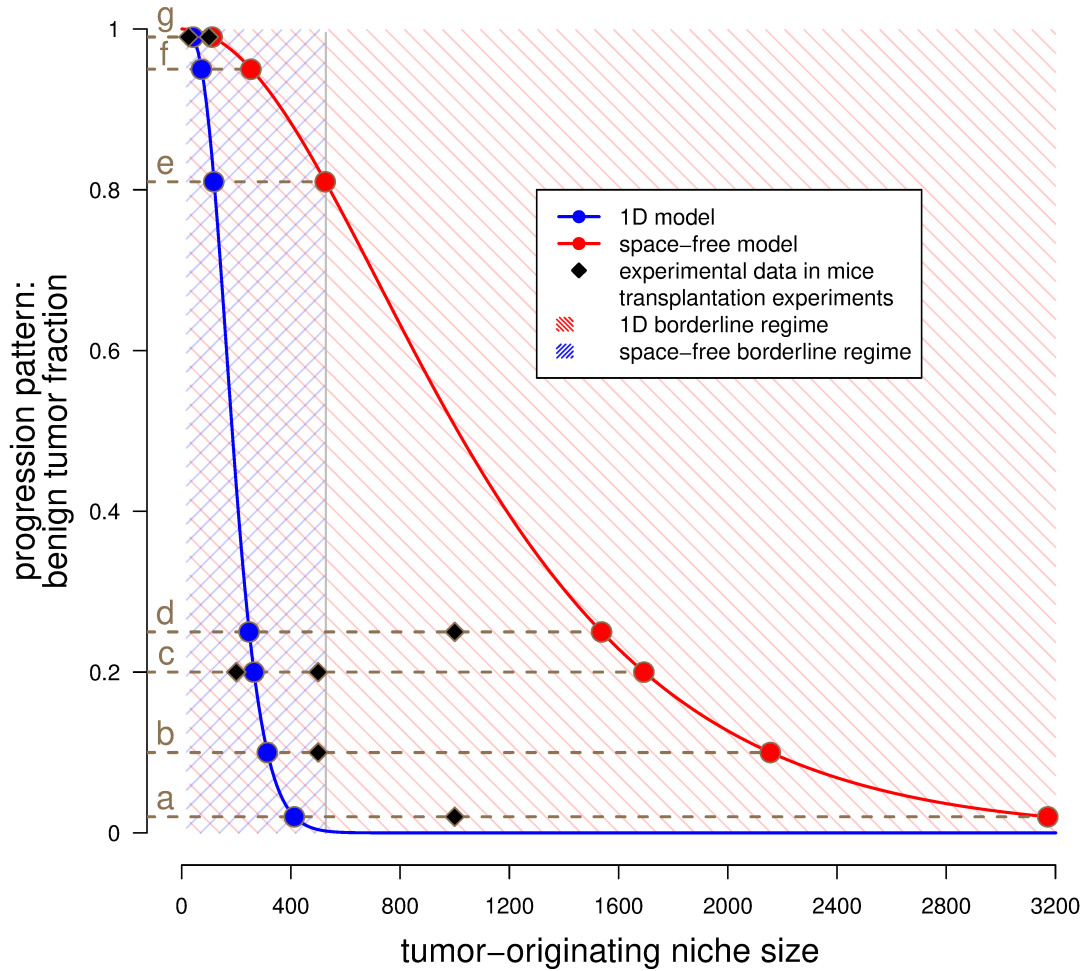


Figure 13: This plot shows the benign tumor fraction in the space-free (red) and one-dimensional (blue) model as function of the tumor-originating niche size. The blue curve has been numerically evaluated, see equation (4.3). The red curve represents the plot of equation (3.4). The shaded areas illustrate the borderline parameter regimes, i.e. in which both sequential and tunneling progression are possible for the space-free and the 1D model. The dots indicate the estimated tumor-originating niche size and the squares represent experimental data for different tissues, see Table 7 for the values.

ranges in various tissues are provided in Table 7 and visualized in Figure 13. Our model predicts that the range of competition is considerably small compared to the overall number of cells in a tumor. Note that we do not assume any upper bound for the parameter N in our model. Moreover, although the estimates are considerably small, the range of competition largely depends on the tissue. For example, the estimated competition range within the liver is 383 – 2837 cells whereas the estimates for the bone marrow are 18 – 31 cells. Importantly, the estimate of the tumor-originating niche size for the human colon assortments well with the stem cell niche size in colonic crypts of about 40 cells [64] but surely less than 100 cells [148]. Overall, these results can be interpreted as existence of a tissue-specific tumor-originating niche in which the fate of tumor development is decided long before a tumor becomes detectable. Based on our results, we propose that a tumor-originating niche size of 291 – 1928 cells within the human brain might be responsible for the clinically observed fraction of primary and secondary glioblastoma, see Table 7b.

PREDICTED TUMOR-ORIGINATING NICHE SIZES AGREE WELL WITH MICE DATA. Interestingly, the estimated tumor-originating niche sizes in Table 7 correspond very well to the necessary cell numbers for tumor induction in mice experiments [8, 16, 42, 46, 54, 81, 104, 149], see also Figure 13. This observation supports two of our model assumptions. First, the possibility of tumor cell extinction for too small populations in these experiments shows that tumor and wild-type cells compete with each other. Second, the experimental data justify our model assumption that a sufficient number of tumor cells is needed to induce tumor formation on the tissue scale.

4.4 DISCUSSION

On the tissue scale, one observes tumor progression types with and without detectable benign precursor stages. Data on the progression patterns with respect to the ratios of these progression types exhibit large differences between tissues. The underlying cellular processes causing these progression patterns are hardly observable and remain unclear. In this work, we shed light on the early phase of tumor development on the cellular scale with the help of a stochastic model. Our model is based on competition between wild-type and tumor cells and assumes that a sufficient amount of tumor cells is needed for tumor formation. We estimate this number by fitting the model to data on the diagnosed ratios of benign and malignant tumor subtypes. Our model predicts that this number is considerably small compared to the overall number of cells in a clinically detectable tumor and largely depends on the tissue which can be interpreted as existence of a tissue-specific tumor-originating niche. Hence, our results suggest that the fate of tumor development is decided long before a tumor becomes detectable.

tissue	benign precursor	malignant tumor	tu- benign fraction	$N_{\text{space-free}}$	N_{1D}	mice data
a) liver	hepatocellular adenoma	hepatocellular carcinoma	2% [135]	2837	383	1000 [149]
b) brain	low-grade astrocytoma	glioblastoma	10% [106]	1928	291	500 [8]
c) breast	ductal carcinoma in situ	invasive ductal carcinoma	20% [67]	1514	246	200[46]–500[42]
d) skin	nevus	melanoma	25%[17]	1375	230	1000 [16]
e) stomach	gastric adenomas	gastric cancer	81% [110]	471	111	200 [81]
f) meninges	benign meningioma	aggressive meningioma	95% [78]	227	67	NA
g) colon	colonic adenoma	adenocarcinoma	99% [22]	100	39	25[54]–100[104]
h) bone marrow	MGUS	myeloma	99.9*% [68]	31	18	NA

Table 7: Estimation of the homeostatic competition range N in different tissues. This table summarizes epidemiological data on benign and malignant tumor subtypes. We calibrate the model such that the absorption probabilities in state N given by equations (3.5) and (4.3) are equal to the diagnosed benign tumor fraction p . We assume a mutation probability of $v = 10^{-6}$ throughout these calculations. The last column contains the necessary number of injected cells to obtain a tumor in mice transplantation experiments. The estimates and mice data are also illustrated in Figure 13. * = estimated. MGUS = monoclonal gammopathy of undetermined significance.

Interestingly, our estimates of the tumor-originating niche size of about 39 cells for colon cancer agrees well with the number of stem cells found in one colonic crypt [64]. Indeed, it is the current understanding that colon adenomas and carcinomas develop within one colonic crypt with intestine stem cells likely to be the cell type of origin [150]. This demonstrates that our model might be utilized to predict tumor-originating niche sizes, thereby allowing to infer the potential cell type of tumor-origin for cancers from other tissues in which the origin is still elusive, e.g. for glioblastoma [83]. Glioblastoma can be divided into two classes dependent on the progression dynamics. In about 90% of cases, glioblastoma occur *de novo*, i.e. without evidence of a less malignant precursor lesion (primary glioblastoma) whereas 10% develop slowly by progression from low-grade gliomas (secondary glioblastoma). Using this data, our model predicts that the size of the tumor-originating niche from which glioblastoma

develop is about $291 - 1928$ cells. Neural stem cells (NSCs) of the subependymal zone (SEZ) have been suggested as a potential cell of origin for glioblastoma. Moreover, recent experimental evidence regarding NSCs in the SEZ of the adult brain suggests that the total number and fate of NSCs is regulated by a density-dependent mechanism [9]. Importantly, the finding in [9] that the fate of a NSC, e.g. activation or quiescence, is coupled to its neighbors perfectly fits to our hypothesis of cells competing within a certain range. Interestingly, the authors also suggest that the fate of active NSCs is coupled to the total number of neighboring NSCs in a shared locally restricted area which suggests that this area is a potential candidate for the tumor-initiating niche in the adult brain. It would be interesting to investigate if the range of coupled NSCs fits to our predicted size of the tumor-originating niche for glioblastoma.

From a modeling perspective, our analysis allows to distinguish different model regimes in MORAN models not only asymptotically as in [26], but also for finite values of the system size N . We find that the risk coefficient in both the space-free and the 1D model does not only distinguish the different model regimes, but quantifies the proportion of tunneling and sequential progression, see Table 6. Moreover, we show that the corresponding asymptotic results in dependency of the risk coefficient provide very good approximations even for small population sizes N . This finding obviates the need of ad-hoc rules like $a \ll b$ if $a/b \leq 1/10$ to choose the appropriate parameter regime [28].

The potential existence of tumor-originating niches in which tumor fate is decided at an early stage of the cellular multistep process supports the view that cancer development is an ecological process [65, 93]. Ecology studies the dynamics of communities of species and their interactions and describe the origin of new species. From this point of view, the size of the tumor-originating niche might represent a critical effective population size that has to be reached by the progeny of the tumor-originating cell type in order to establish a tumor on the tissue scale. A deeper understanding of the processes and the origin of the tumor-originating niche contributes to the understanding of the early phase of tumor development. Further modeling and experimental effort is needed to understand this early phase of tumor development on the cellular scale in a better way. In this work, we demonstrated how observable quantities on the tissue scale might be utilized to achieve this goal.

Part II

THE ROLE OF PHENOTYPIC CELL CHANGES IN TUMOR DEVELOPMENT

This part is devoted to the following two questions:

- What is the role of *phenotypic cell state changes* of cancer cells in the establishment of a *tissue-scale* equilibrium with respect to different cell phenotypes in tumors?
- To which extent is the observed phenotypic heterogeneity on the *tissue scale* caused by hierarchical phenotypic cell changes and what is the role of different environmental conditions in determining this heterogeneity?

5

CELLTRANS: AN R PACKAGE TO QUANTIFY STOCHASTIC CELL STATE TRANSITIONS*

Contents

5.1	Characteristic cell state compositions in cancer	86
5.2	Introducing CELLTRANS	87
5.2.1	Reference experiment	87
5.2.2	Detailed description and analysis of CELLTRANS	88
5.2.3	Implemented functions in CELLTRANS	92
5.2.4	Applications of CELLTRANS	93
5.3	Case studies demonstrating the applicability of CELLTRANS	94
5.3.1	Dynamics between cancer cell sub-populations in colon cancer	94
5.3.2	Dynamic switch between adhesive and suspended cell types in colon cancer	97
5.3.3	Proportions of stem-like, basal and luminal phenotypes in breast cancer	100
5.3.4	Epithelial-mesenchymal transition in breast cancer	102
5.3.5	Influence of the choice of the time step length τ	103
5.3.6	Validation of the predicted equilibrium of CELLTRANS	105
5.3.7	Influence of non-pure initial cell state compositions	105
5.3.8	A simulation study demonstrating matrix regularization	105
5.4	Discussion	109

* This chapter includes text and figures from the publication and supplementary material of: Thomas Buder, Andreas Deutsch, Michael Seifert, and Anja Voss-Böhme, *Bioinformatics and Biology Insights*, 11, 2017. Author contribution: Thomas Buder wrote the manuscript, designed, and analyzed the mathematical model and wrote the R package. Andreas Deutsch contributed to study design and writing the manuscript. Michael Seifert and Anja Voss-Böhme contributed to study design, writing the manuscript, and supervised the study.

5.1 CHARACTERISTIC CELL STATE COMPOSITIONS IN CANCER

Homeostasis with respect to the proportions of cells in different states is crucial for the functioning of multi-cellular organisms and its regulation enables organisms to stay in a healthy state. Different types of tissues and organs need to maintain a stable composition of different cell types regardless of external conditions, injuries and changing environmental conditions in order to function normally [12]. Hence, finding mechanisms of homeostasis regulation is a key aspect in understanding the emergence of diseases, such as cancer, that lead to disturbance and loss of cell state homeostasis [47].

Remarkably, diseases like cancer disturb healthy homeostatic states but can lead themselves to a characteristic composition with respect to the proportions of distinctive neoplastic cell states [86]. The establishment and maintenance of such a characteristic composition has been experimentally shown by utilizing fluorescence activated cell sorting (FACS) and flow cytometry experiments for many types of cancer, e.g. breast cancer [45] and colon cancer [40, 143, 147]. In these experiments, it has been observed that sub-populations of cells purified for a given cell state return to the composition of cell state proportions of the original tumor over time.

The mechanisms for the maintenance of these characteristic compositions are only poorly understood. Cell state proportions could be maintained by regulated cell-state-specific proliferation rates, e.g. due to inter-cellular signaling [45]. However, in many cases, this possibility can be experimentally excluded by showing that the proliferation rates of all involved cell types are equal and constant over time. There is evidence that cell types stochastically transition between different states and that the transition rates do not depend on the current tissue composition or on inter-cellular signaling [45], i.e. the chances to transition into other cell states only depend on the current state of the cell. Quantifying the probabilities for transitions from one cell state to another would allow to predict the evolution of cell state proportions. Such a quantification can potentially help to understand the differences in homeostasis regulation between healthy and diseased tissues.

One approach to model cell state transitions utilizes ordinary differential equations (ODEs). Typically, the dynamics between different cell states are described by formulating ODEs incorporating parameters which describe detailed cell properties like symmetrical / asymmetrical division rates and transition rates between cell states [40, 143].

Another possibility to model the evolution of cell state proportions are discrete time MARKOV models. Discrete time MARKOV models are particular stochastic processes which can be understood as sequences of random variables indexed by discrete time points, where the next state only depends on the current state of the process but is independent of earlier states [13]. For instance, in [45], a MARKOV model describing the evolution of cell state proportions has been in-

troduced and applied to breast cancer cell lines. However, a detailed discussion of how the transition probabilities are derived from the experimental data and of potential analytical challenges is missing. The quantification of cell state transitions by estimating transition probabilities would allow to better understand characteristics of cell state proportions in both healthy and disease-related tissues. To our knowledge, there is no tool available that allows to automatically estimate cell state transition rates from [FACS](#) and flow cytometry experiments.

We develop such a general tool to estimate the transition probabilities between different cell states from appropriately prepared data. The underlying model is based on a discrete time `MARKOV` model and allows to quantify cell state transitions from data on the temporal evolution of cell state proportions. We utilize a discrete time `MARKOV` model since it serves as a minimal model for the evolution of cell state proportions. In contrast, `ODE` models often require additional parameters which must be measured experimentally [143] or obtained by fitting [40]. Moreover, `MARKOV` models have already been successfully utilized to analyze dynamic cell compositions [14, 34, 45, 72]. Here, we generalize this approach and develop an automated tool for the analysis of cell state transitions. We demonstrate which analytical problems can occur in the estimation and in which way these problems are automatically solved by our tool. Furthermore, we provide a publicly available R package called `CELLTRANS` which can be directly utilized by experimentalists to analyze cell state proportion data from [FACS](#) and flow cytometry cell line experiments.

We illustrate potential applications of `CELLTRANS` by analyzing publicly available data on the evolution of cell state compositions in different cell lines. We show that the quantification of cell state transitions allows to predict the cell state composition at any time point of interest. In particular, our model is able to predict the long-term equilibrium composition of cell types. Furthermore, our model can reveal frequent and rare cell state transitions. Moreover, `CELLTRANS` can be utilized to estimate the time needed until perturbations of the characteristic cell state compositions level out. Such predictions have the potential to support experimentalists in planning the duration of [FACS](#) and flow cytometry cell line experiments.

5.2 INTRODUCING CELLTRANS

5.2.1 *Reference experiment*

`CELLTRANS` is able to analyze data recording changes of cell state proportions over time. The identification of individual cell states from mixed cell populations is mainly based on cell-type specific gene markers which allow to experimentally separate the different cell types, for instance by [FACS](#) techniques [49]. We assume that cell state proportions change in time due to stochastic transitions dependent only on the current state of the cell. A further prerequisite for

the application of our model is the equality of proliferation and death rates of all cell types.

According to the number of different cell types distinguished in the experiments, an arbitrarily large, but finite integer m is fixed, defining the number of distinct cell states in the model. Typically, all distinct cell states are purified and a large number of cells are separately cultured for each cell type. This experimental setup leads to m experiments whose evolution of cell state proportions are simultaneously monitored at t different time points n_1, n_2, \dots, n_t . The time points are multiples of a unit time step of length τ depending on the time-scale of the experiment, e.g. τ is 1 hour, 1 day or 1 week. The time points of measurement are not necessarily integer multiples of τ . The data on cell state proportions at each time point $n_j, j = 1, \dots, t$, are the basis of the analysis with CELLTRANS as described in the next section.

Note that CELLTRANS also allows to analyze experiments with non-pure initial cell state proportions. Importantly, the number of experiments has to be the same as the number of defined cell states m . Furthermore, the vectors describing the initial cell state proportions have to be linearly independent, i.e. they cannot be represented as linear combinations of each other, see also the next subsection.

5.2.2 Detailed description and analysis of CELLTRANS

We denote the cell states distinguished in the experiments by $1, \dots, m$. The experimental data on cell state proportions obtained for each time point $n_j, j = 1, \dots, t$, need to be arranged in matrices

$$W^{(n_j)} = \left(w_{kl}^{(n_j)} \right)_{k,l=1,\dots,m} .$$

Here, $w_{kl}^{(n_j)}$ describes the proportion of cell state k in the l -th experiment at time n_j . These matrices can be stored into text files and read into the CELLTRANS R package. Figure 14 illustrates the experimental setup and the construction of the data matrices.

We assume that each cell transitions from state i to state j during a time step of length τ with probability $p_{ij}, i, j = 1, 2, \dots, m$. Note that the choice of the time step length has no substantial influence on the predictions which we discuss later. Furthermore, it is assumed that the transition probabilities are constant over time and do only depend on the state of the cell [45]. These considerations lead to a discrete time MARKOV process with transition matrix $P = (p_{ij})_{i,j=1,\dots,m}$ for the random evolution of the state of individual cells. Our goal is to estimate P , i.e. all transition probabilities between the cell states, from the experimental data stored in the matrices $W^{(n_0)}, \dots, W^{(n_t)}$. The interaction graph of the underlying MARKOV model together with the transition probabilities are illustrated in Figure 18A.

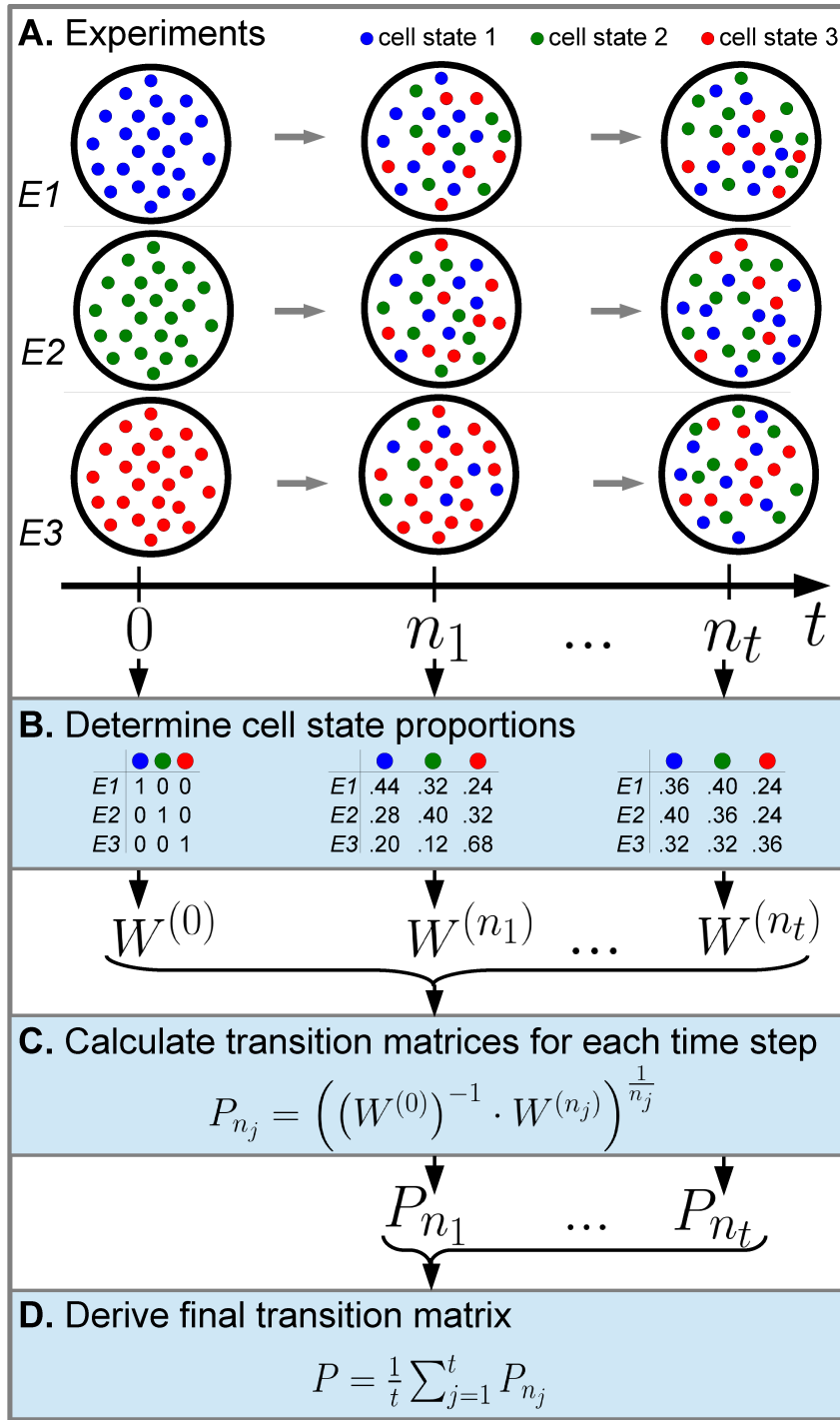


Figure 14: General CELLTRANS application workflow. In order to demonstrate the general workflow of CELLTRANS, we utilize a fictive experiment with three illustrative cell states 1 (blue), 2 (green) and 3 (red). *A.* Three different FACS experiments E1, E2 and E3 with pure initial cell states and fictional evolutions of these proportions are illustrated. *B.* The measured cell state compositions at times n_j are utilized to construct the cell state proportion matrix $W^{(n_j)}$, $j = 1, \dots, t$. Here, each row corresponds to the cell state proportions of one experiment. *C.* CELLTRANS estimates a transition matrix P_{n_j} based on the cell state proportion matrix $W^{(n_j)}$ for each time point of measurement $j = 1, \dots, t$. *D.* Finally, averaging the estimated transition matrices P_{n_j} leads to the final transition matrix P .

Construction of data matrices

Let $w_k^{(0)} = (w_{k1}^{(0)}, \dots, w_{km}^{(0)})$ denote the initial cell state proportions in the k -th experiment which is a row vector of length m with non-negative entries summing to one. As explained before, m of those initial cell state proportions $w_1^{(0)}, w_2^{(0)}, \dots, w_m^{(0)}$ are needed. The initial experimental matrix $W^{(0)}$ is row-wise constructed from these cell state proportions, i.e.

$$W^{(0)} := \begin{pmatrix} - & w_1^{(0)} & - \\ - & \vdots & - \\ - & w_m^{(0)} & - \end{pmatrix} = \begin{pmatrix} w_{11}^{(0)} & \dots & w_{1m}^{(0)} \\ \vdots & \ddots & \vdots \\ w_{m1}^{(0)} & \dots & w_{mm}^{(0)} \end{pmatrix}.$$

An analytical requirement for the applicability of CELLTRANS is the existence of the inverse of $W^{(0)}$. This can be assured by appropriate initial cell state proportions, e.g. by choosing purified initial cell cultures. Since this experimental setup is very common in FACS and flow cytometry experiments, it is implemented as default option in CELLTRANS. However, also an individually designed initial experimental matrix can be used.

In the experiments, the cell state proportions at each time point n_1, n_2, \dots, n_t have to be assessed. Let $w_k^{(n_j)}$ denote the cell state proportions of experiment $k = 1, 2, \dots, m$, after time n_j with $j = 1, 2, \dots, t$. For each of these time points, a cell state proportion matrix after time n_j is obtained by constructing the matrix

$$W^{(n_j)} = \begin{pmatrix} - & w_1^{(n_j)} & - \\ - & w_2^{(n_j)} & - \\ - & \vdots & - \\ - & w_m^{(n_j)} & - \end{pmatrix}$$

as described above for $W^{(0)}$. Since each row describes the cell state proportions in the corresponding experiment, all rows sum up to one with non-negative entries. In total, t of such matrices need to be constructed from the experiments, one for each time point of measurement.

Derivation of transition matrices P_{n_j}

For each time point n_j with $j = 1, 2, \dots, t$, CELLTRANS derives a transition matrix P_{n_j} as follows. This derivation is based on the theory of MARKOV models [13]. We utilize that the distribution of a MARKOV chain after n_j time steps can be obtained by multiplying the initial distribution with the transition matrix raised to the power of n_j , hence

$$W^{(0)} P^{n_j} = W^{(n_j)}.$$

Here, $W^{(0)}$ denotes the initial experimental matrix and $W^{(n_j)}$ the cell state proportion matrix after time n_j . Since we are interested in the underlying transition matrix P , we solve this equation for P by multiplying with the inverse of $W^{(0)}$ and computing the n_j -th matrix root, i.e.

$$\begin{aligned} P^{n_j} &= \left(W^{(0)}\right)^{-1} \cdot W^{(n_j)} \text{ and} \\ P_{n_j} &:= \left(\left(W^{(0)}\right)^{-1} \cdot W^{(n_j)}\right)^{\frac{1}{n_j}} \end{aligned} \quad (5.1)$$

for $j = 1, 2, \dots, t$. Here, we compute the so-called principal matrix root since matrix roots are not necessarily unique. For an overview of the existence and computation of matrix roots, see [50]. Note that the existence of $\left(W^{(0)}\right)^{-1}$ is assured by the appropriate choice of the initial cell state proportions described earlier. The matrix P_{n_j} is the estimated transition matrix derived from time point n_j . Although this derivation looks straightforward, there are potential analytical problems which are described in the following.

Regularizing matrix roots to stochastic matrix roots

Importantly, equation (5.1) should yield a transition matrix of a MARKOV chain, i.e. a stochastic matrix with non-negative entries and row sums equal to one. However, the root of a stochastic matrix is not necessarily stochastic again [50]. CELLTRANS verifies whether the matrix roots are stochastic or not. If not, the matrix roots are regularized to be stochastic with the QOM-Algorithm (Quasi-optimization of the root matrix) which is sketched in the following and described in detail in [74].

The QOM-Algorithm performs a row-wise Euclidean distance minimization by transforming each row of the matrix into a valid row of a transition matrix, i.e. a vector containing non-negative entries which sum to one. The result is a uniquely determined stochastic matrix which closely approximates the original matrix. In [74], the effect of QOM regularization on non-stochastic matrix roots is numerically investigated. The authors calculated the infinity matrix norm and also the mean absolute deviation of the difference between the QOM result and the original transition matrix for 32 examples of matrix regulation. This numerical comparison demonstrates a low approximation error of the QOM regularization.

Computation of the transition matrix P

CELLTRANS estimates the transition matrix P by averaging the transition matrices P_{n_j} for each time point $j = 1, 2, \dots, t$, i.e.

$$P := \frac{1}{t} \sum_{j=1}^t P_{n_j}. \quad (5.2)$$

This transition matrix is the final estimation of CELLTRANS quantifying the transition probabilities between all cell states. The overall workflow of CELLTRANS is summarized in Figure 14.

Note that the dynamics of the MARKOV chain model can also be described by a master equation, i.e. a set of first-order differential equations. The master equation reads

$$\frac{d}{dt} a_i(t) = \sum_{j \neq i} (p_{ji} a_j(t) - p_{ij} a_i(t)), \quad i = 1, \dots, m \quad (5.3)$$

with initial conditions $a_i(0) = c_i$ [39]. Here, $a_i(t)$ describes the proportion of cells in state i at time t and $P = (p_{ij})_{i,j=1,\dots,m}$ represents the estimated transition matrix from CELLTRANS. Hence, these differential equations describe the temporal evolution of the cell state proportions. We will utilize this master equation later to compare the results of CELLTRANS to those of ODE models.

5.2.3 *Implemented functions in CELLTRANS*

Here, we introduce the most important functions which are implemented in CELLTRANS. In the following section, we will demonstrate the usage of these functions in several case studies.

`readExperimentalData()`

This function reads all necessary data. First, it opens a dialog box which asks for the number of cell types, the names of the cell types, the time step length τ and the time points of measurement. Then, the files containing the cell state proportion matrices are read. First, the initial experimental setup matrix $W^{(0)}$ can either be chosen as identity matrix (for pure initial cell populations) or an individual initial matrix can be used. Then, the experimental cell proportion matrices are read for each time point of measurement. It is recommendable to save the input into a variable for further analysis, e.g. `input <- readExperimentalData()`.

`celltransitions(input)`

This functions derives and prints the estimated transition probabilities and the predicted equilibrium distribution. The variable `input` contains the read data from the function `readExperimentalData()`.

`celltransplot(input)`, `celltransplotPDF(input)`

These functions allow to create plots of the predictions of CELLTRANS and the experimental data. The variable `input` contains the read data from `readExperimentalData()`.

`timeToEquilibrium(input, initialDistribution, tol)`

This function estimates the time from any initial cell state proportions until the equilibrium proportions are reached. The variable `input` contains the read data from the function `readExperimentalData()`.

The variable `initialDistribution` is a vector of length m which describes the initial cell state proportions, for example `c(0.25, 0.25, 0.25, 0.25)` for equal proportions of $m = 4$ cell types. The third parameter `tol` gives a tolerance deviation between the cell state proportions of the equilibrium distribution and those of the predicted cell state proportions, since the exact equilibrium distribution is not reached, in general. For the parameter `tol`, we recommend values between 0.01 and 0.02.

For a comprehensive introduction demonstrating the application of these functions, see the detailed vignette provided in the Appendix, Section [A.2](#).

5.2.4 Applications of CELLTRANS

CELLTRANS can be applied to analyze cell state proportion data from [FACS](#) and flow cytometry experiments with respect to several questions. The applications are based on the estimation of the transition probabilities of the underlying model as described above.

1. The estimated probabilities quantify the frequencies of state transitions and can be used to detect frequent, rare or almost never occurring transitions. Such a prediction allows to hypothesize about biological mechanisms which are responsible for the observed transition structure, e.g. an underlying transition hierarchy.
2. Another application is the prediction of cell line compositions at any time point. Such an estimate can be utilized to predict cell line compositions even beyond the time periods of experiments which we will demonstrate in the Results section.
3. CELLTRANS can be used to estimate the equilibrium cell state proportions. This information can support experimentalists to decide whether experimentally observed cell line compositions already reached equilibrium.
4. CELLTRANS allows the prediction of the time needed to reach equilibrium proportions from any initial cell state composition. The choice of the time

period of [FACS](#) experiments and the time points at which cell state compositions are measured is often difficult. Here, the estimate of the time needed to reach equilibrium can be useful.

Table 8 summarizes the main applications and the corresponding functions in CELLTRANS.

estimation of	application / interpretation	implemented function(s)
transition probabilities	quantification of cell state transitions	
	detect transition hierarchies detect rare and frequent transitions & conclude responsible mechanisms	celltransitions()
cell state compositions	predictions beyond time period of experiments	celltransplot()
	validate experimental results	celltransplotPDF()
equilibrium compositions	predict equilibrium compositions	
	validate experimental results	celltransitions()
time to equilibrium	plan time periods of experiments choice of time points of measurement	timeToEquilibrium()

Table 8: Main applications of CELLTRANS and corresponding implemented functions.

5.3 CASE STUDIES DEMONSTRATING THE APPLICABILITY OF CELLTRANS

In this section, we apply CELLTRANS to publicly available data on the evolution of cell state proportions obtained from [FACS](#) and flow cytometry experiments. We point out possible conclusions that can be drawn from the application of CELLTRANS.

5.3.1 *Dynamics between cancer cell sub-populations in colon cancer*

BACKGROUND There is evidence that $CD133^+$ cells represent a cancer stem cell (CSC) sub-population within SW620 human colon cells [147]. In [147], the

dynamics between CSCs and non-stem cancer cells (NSCC) has been experimentally investigated. In detail, purified NSCCs and CSCs sorted from the SW620 cell line by FACS were cultured for 26 days, and the composition of these cultures was measured every second day in both experiments. Here, we analyze the data from these experiments with CELLTRANS and compare the resulting predictions to those obtained from an ordinary differential equations (ODE) model which has been analyzed in [143].

APPLICATION OF CELLTRANS. The experimental setup in [147] can be formulated within the framework of CELLTRANS as follows. There are $m = 2$ subpopulations considered in SW620 human colon cancer cells which are called CSC and NSCC. Note that the assumption that both of these cell types do not differ in their proliferation and death rates is justified [143]. The initial experimental matrix representing purified initial cell cultures is

$$W^{(0)} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

where the first row corresponds to the experiment with sorted CSC and the second row with sorted NSCC. Hence, we choose the identity matrix in CELLTRANS as initial experimental matrix.

There are in total 12 time points of cell state proportion measurements which are given by 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22 and 24 days. Thus, it is sensible to choose a time step length τ of 1 day. The cell state proportion matrices for each of these time points can be derived from the experimental data provided in Table S2 and Table S3 in [143] which leads to 12 corresponding cell state proportion matrices. For example, the cell state proportion matrix after 8 days is

$$W^{(8)} = \begin{pmatrix} 0.7150 & 0.2850 \\ 0.5523 & 0.4477 \end{pmatrix},$$

i.e. the experiment starting with 100 % CSC evolved to 71.5% CSC and 28.5% NSCC and the experiment starting with pure NSCC evolved to 55.23% CSC and 44.77% after 8 days, respectively.

The function `celltransitions(input)` allows to derive the transition probabilities between the cell states and the predicted equilibrium distribution. CELLTRANS derives a transition matrix for each of the 12 cell state proportion matrices. Note that none of these matrices require regularization by the QOM algorithm, since the matrix roots are already stochastic matrices. For example, the transition matrix derived from the experimental data after 8 days $W^{(8)}$ is

$$P_8 = \begin{pmatrix} 0.9309 & 0.0691 \\ 0.1339 & 0.8661 \end{pmatrix}.$$

Finally, the transition matrix P is obtained by computing the average of the 12 transition matrices from each time point, i.e.

$$P = \frac{1}{12} \sum_{j=1}^{12} P_{n_j} = \begin{pmatrix} 0.9455 & 0.0545 \\ 0.1030 & 0.8970 \end{pmatrix},$$

which is an estimate of the transition probabilities between NSCC and CSC per day. For example, a CSC converts with probability 0.0545 to a NSCC or a NSCC converts with probability 0.1030 to a CSC within a day. The predicted long-term cell state proportions can be obtained by `CELLTRANS` from the steady state of the transition matrix P and are also provided by the function `celltransitions(input)` which yields

$$w^* = \begin{pmatrix} 0.654 \\ 0.346 \end{pmatrix},$$

i.e. `CELLTRANS` predicts a proportion of 65.4 % CSC and of 34.6 % NSCC after sufficient time. The time from 100% CSC to the equilibrium composition can be estimated with the command `timetoEquilibrium(input,c(1,0),0.01)` which yields 21 days. The time from 100% NSCC to the equilibrium composition can be estimated in an analogous way which yields 25 days. Therefore, the time of the experiment was sufficient to reach the equilibrium in this case.

MODEL COMPARISON. In [143], the dynamics between CSC and NSCC has been modeled by an ODE model with 8 parameters. Some of these parameters have been collected from *in situ* experiments (see Table 1 in [143]). The predictions from our model show a much better accordance to the predictions of the ODE model (Root mean square deviation (RMSD) `CELLTRANS` vs. ODE: 0.03737 vs. 0.09874 for the experiment starting with pure CSC cultures, RMSD `CELLTRANS` vs. ODE: 0.05326 vs. 0.10484 for the experiment starting with pure NSCC cultures). A comparison between the original data from [143] and the predictions of the derived MARKOV chain is shown in Figure 15A illustrating that the estimated MARKOV model well describes the experimental data.

We can utilize the master equation (5.3) to equivalently describe the cell state dynamics with an ODE system

$$\begin{aligned} \frac{d}{dt} \text{CSC}(t) &= -0.0545 \text{CSC}(t) + 0.1030 \text{NSCC}(t) \\ \frac{d}{dt} \text{NSCC}(t) &= 0.0545 \text{CSC}(t) - 0.1030 \text{NSCC}(t) \end{aligned}$$

with appropriate initial conditions, e.g. $CSC(0) = 1, NSCC(0) = 0$ for the experiment with pure CSC cells in the beginning. The solution

$$\begin{aligned} CSC(t) &= 0.346e^{-0.1575t} + 0.654 \\ NSCC(t) &= -0.346e^{-0.1575t} + 0.346 \end{aligned}$$

allows to obtain the steady state by letting $t \rightarrow \infty$.

SUMMARY. This case study demonstrates that the MARKOV model underlying CELLTRANS is potentially able to make better predictions than more complex ODE models with parameter calibration. The experimental data on the evolution of cell state proportions are sufficient to interpolate the data. Moreover, the equilibrium proportion is reliably predicted.

5.3.2 *Dynamic switch between adhesive and suspended cell types in colon cancer*

BACKGROUND. In [40], the dynamic switch between two different adhesion phenotypes in colorectal cancer cells has been analyzed. The involved cell states in this study are *adherent* and *suspended* cells. Both of these cell types have been separated and re-plated. Subsequently, both initially pure cell cultures have been monitored over 8 hours, and the cell state proportions have been measured after 0.5 and 1 hours and subsequently in 2 hour intervals.

APPLICATION OF CELLTRANS. The study can be integrated in the CELLTRANS framework in the following way. There are $m = 2$ cell states, adherent and suspended. Although there is no explicit statement regarding the proliferation rates in [40], we assume equal rates since the authors model the evolution of the cell state proportions without considering effects of proliferation. The initial experimental matrix is given by the identity matrix, since the experimental initial cell state proportions are pure cell cultures of both phenotypes. Here, the first row corresponds to the experiment with 100% adherent cells and the second row with 100% suspended cells.

Subsequently, the cell state proportion matrices for $t = 6$ time points (at 0.5h, 1h, 2h, 4h, 6h and 8h) can be deduced from the experiments. Here, the time step length τ is chosen to be 1 hour. Subsequently, CELLTRANS derives transition matrices for each of these cell state proportion matrices. This leads to 6 transition matrices $P_{0.5}, P_1, P_2, P_4, P_6$ and P_8 . All of these matrices do not require regularization in this case, since all of them are already stochastic matrices.

Finally, the transition matrix P is obtained as average

$$P = \frac{P_{0.5} + P_1 + P_2 + P_4 + P_6 + P_8}{6} = \begin{pmatrix} 0.8612 & 0.1388 \\ 0.2908 & 0.7092 \end{pmatrix}.$$

The predicted cell state proportions based on the estimated transition matrix and the experimental data is plotted in Figure 15B.

The steady state of the derived transition matrix P is given by

$$w^* = \begin{pmatrix} 0.6768 \\ 0.3232 \end{pmatrix},$$

i.e. the long-term cell state proportions of adherent and suspended cells is predicted as 67.68% and 32.32%, respectively. This prediction is in good accordance to the corresponding experimentally observed equilibrium proportions described in [40].

MODEL COMPARISON. The authors in [40] formulated a mathematical ODE model to describe the dynamics for the adherent and the suspended cells to reestablish the equilibrium ratio. Note that this approach requires to fit the analytical solution of the model to the experimental data. The fit in [40] yields the solution $\text{adh}(t) = 0.3e^{-0.5t} + 0.7$.

In contrast, we can derive an alternative ODE system with the same structure by utilizing the master equation (5.3) based on the predictions of CELLTRANS, i.e.

$$\begin{aligned} \frac{d}{dt}\text{adh}(t) &= -0.1388\text{adh}(t) + 0.2908\text{susp}(t) \\ \frac{d}{dt}\text{susp}(t) &= 0.1388\text{adh}(t) - 0.2908\text{susp}(t) \end{aligned}$$

with initial conditions $\text{adh}(0) = 1, \text{susp}(0) = 0$. The solution

$$\begin{aligned} \text{adh}(t) &= 0.323e^{-0.4296t} + 0.677 \\ \text{susp}(t) &= -0.323e^{-0.4296t} + 0.323 \end{aligned}$$

is close to the solution in [40] and allows to predict the equilibrium distribution by letting $t \rightarrow \infty$. This equilibrium solution well corresponds to the experimental findings.

The predictions of CELLTRANS and the ad hoc ODE in [40] are in good accordance with the original data (RMSD CELLTRANS vs. ODE: 0.0219 vs. 0.0278 for the experiment starting with pure adherent cultures, RMSD CELLTRANS vs. ODE: 0.02709 vs. 0.02407 for the experiment starting with pure suspended cultures), see also Figure 15B.

SUMMARY. It remains unclear in which way the *best-fit* solution in the introduced ODE approach in [40] has been obtained. Instead of fitting parameters, our approach offers a transparent estimation of the underlying transition matrix yielding good predictions. Moreover, especially from the point of view of an experimentalist, the automated estimation by CELLTRANS does not require a deeper engagement with mathematical modeling.

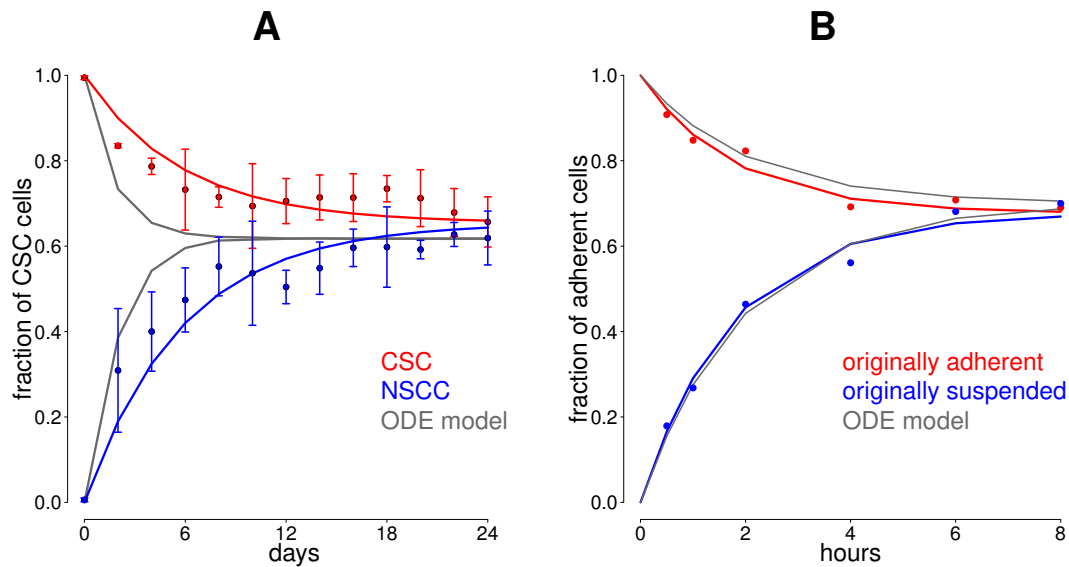


Figure 15: CELLTRANS model predictions for colon cancer cell lines. *A.* We utilized CELLTRANS to analyze data about the evolution of cancer cell line compositions from [147]. The involved cell states are cancer stem cells (CSC) and non-stem cancer cells (NSCC) and the original data are plotted as colored dots including the experimental standard deviation. The red curve is the prediction of the cell state compositions of the experiment starting with pure CSC and the blue curve represents the prediction starting with pure NSCC. The gray line corresponds to the predictions of ODE models which have been proposed in the original study [147]. *B.* Analysis of colon cancer cell line data from [40] with the cell states adherent and suspended. The red curve is the prediction of CELLTRANS for the experiment starting with adherent cells only and the blue curve is the corresponding prediction starting with suspended cells. The gray line is the prediction of the ODE model introduced in [40].

5.3.3 Proportions of stem-like, basal and luminal phenotypes in breast cancer

BACKGROUND. The dynamics of phenotypic proportions in human breast cancer cell lines is studied in [45]. In detail, the authors used FACS analysis to isolate three mammary epithelial cell states (stem-like, basal and luminal) from the SUM159 and SUM149 breast cancer cell lines. Pure sub-populations of the three cell states have been cultured for six days and cell state proportions have been measured at the end of the experiment.

APPLICATION OF CELLTRANS. We apply CELLTRANS to cell lines SUM149 and SUM159. The proliferation rates of the involved cell types are equal [45]. The initial experimental cell proportions in both cases can be described by

$$W^{(0)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

where the first line corresponds to the experiment with sorted stem-like cells, the second row to sorted basal cells and the third row to sorted luminal cells.

The proportions of cell states have been obtained at a single time point after six days. The time step length τ is one day in this case. For SUM149, the following cell state proportion matrix can be obtained from [45]

$$W^{(6)} = \begin{pmatrix} 0.053 & 0.168 & 0.78 \\ 0.026 & 0.504 & 0.47 \\ 0.013 & 0.007 & 0.98 \end{pmatrix},$$

where the first row contains the cell state proportions of the experiment with sorted stem-like cells, the second row with sorted basal cells and the third row with sorted luminal cells in the beginning, respectively.

By utilizing the function `celltransitions(input)`, CELLTRANS derives the following transition matrix from this time point

$$P_6 = \begin{pmatrix} 0.5814 & 0.1102 & 0.3084 \\ 0.0145 & 0.8887 & 0.0968 \\ 0.0058 & 0.0006 & 0.9936 \end{pmatrix},$$

which yields the final transition probabilities, since there is only one time point of measurement in this case. For example, the second row indicates that basal cells transition to stem-like cells within a day with a probability of 1.45%, do not change their state with a probability of 88.87% and convert to the luminal

state with a probability of 9.68%. A similar derivation leads to the transition probabilities for SUM159 which exhibits different transition dynamics.

The predicted equilibrium distribution of our model for SUM149 is

$$w^* = \begin{pmatrix} 0.014 & 0.019 & 0.967 \end{pmatrix}$$

and for the SUM159 cell line

$$w^* = \begin{pmatrix} 0.0235 & 0.9734 & 0.0031 \end{pmatrix}$$

(first entry stem-like, second basal, third luminal), see Figures 16A-F. These plots can be created in CELLTRANS with the command `celltransplot(input)`. The predictions are in good accordance with the original tumor compositions. In detail, the SUM149 tumor sample is composed of 3.9 % stem, 3.3% basal and 92.8 % luminal cells. The proportions within the SUM159 cell line is 1.9% stem, 97.3% basal and 0.62% luminal cells [45].

The time from 100% stem-like cells to the predicted equilibrium composition can be estimated by CELLTRANS with the command `timeToEquilibrium(input,c(1,0,0),0.01)` which yields 31 days. In contrast, the command `timeToEquilibrium(input,c(0,0,1),0.01)` gives an estimation of 8 days to reach the equilibrium from 100% luminal cells. These predictions reflect that purified luminal cells are much closer to the equilibrium composition than purified stem-like cells.

COMPARISON TO PREVIOUSLY USED MARKOV MODEL. The authors of [45] introduced a MARKOV model of cell state transitions to explain the observed equilibrium. The predictions are based on a single time point, and no regularization of the matrix root is required. CELLTRANS is able to recover the transition matrices for both cell lines SUM149 and SUM159 presented in [45], Table 1. Figures 16A-F illustrate the predicted evolution of cell fractions for both cell lines.

The master equation (5.3) can also be applied to derive an equivalent ODE system

$$\begin{aligned} \frac{d}{dt}\text{stem}(t) &= -0.4186\text{stem}(t) + 0.0145\text{bas}(t) + 0.0058\text{lum}(t) \\ \frac{d}{dt}\text{bas}(t) &= 0.1102\text{stem}(t) - 0.1113\text{bas}(t) + 0.0006\text{lum}(t) \\ \frac{d}{dt}\text{lum}(t) &= 0.3084\text{stem}(t) + 0.0968\text{bas}(t) - 0.0064\text{lum}(t) \end{aligned}$$

The solution for the initial conditions $\text{stem}(0) = 1, \text{bas}(0) = 0, \text{lum}(0) = 0$ reads

$$\begin{aligned} \text{stem}(t) &= 0.0141 + 0.9771e^{-0.4274t} + 0.0088e^{-0.1089t} \\ \text{bas}(t) &= 0.0191 - 0.3394e^{-0.4274t} + 0.3203e^{-0.1089t} \\ \text{lum}(t) &= 0.9668 - 0.6377e^{-0.4274t} - 0.3291e^{-0.1089t}. \end{aligned}$$

SUMMARY. Several applications of our model are demonstrated here. First, CELLTRANS is able to analyze arbitrarily many cell states, not only two. Second, the original tumor composition, which is not included in the analysis, is precisely predicted. Third, the estimation of the time to the equilibrium composition potentially helps experimentalists in planning the time periods of cell line experiments.

5.3.4 Epithelial-mesenchymal transition in breast cancer

BACKGROUND. In order to investigate the epithelial-mesenchymal transition (EMT) and its implication on the development and progression of breast cancer, the authors of [89] induced an EMT in non-tumorigenic, immortalized human mammary epithelial cells (HMLEs). Subsequently, they used flow cytometry analysis to sort the cells based on the expression of CD44 and CD24, two cell-surface markers whose expression in the $CD44^+/CD24^-$ configuration is associated with both human breast cancer stem cells and normal mammary epithelial stem cells. One of their aims was to determine whether the $CD44^+/CD24^-$ cells isolated from monolayer cultures of HMLE cells could generate $CD44^-/CD24^+$ cells *in vitro*.

To examine this question, the authors cultured purified cell phenotypes into monolayer cultures and assayed for the appearance of other cell phenotypes during time. The results of these experiments are summarized in Table S1 in [89].

APPLICATION OF CELLTRANS. This experimental setup can be formulated within the CELLTRANS framework as follows. There are $m = 2$ cell states which are named $CD44^+/CD24^-$ and $CD44^-/CD24^+$. The initial experimental matrix is given by the identity matrix. This matrix is constructed such that the first row represents the pure $CD44^+/CD24^-$ initial cell culture, and the second row corresponds to pure $CD44^-/CD24^+$ cell culture.

The cell state proportions have been experimentally determined for $t = 4$ time points (at 2, 4, 6 and 8 days). Here, the time step length τ is 1 day. The cell state proportions matrices $W^{(2)}$, $W^{(4)}$, $W^{(6)}$ and $W^{(8)}$ can be constructed from the data in [89], e.g.

$$W^{(4)} = \begin{pmatrix} 0.70 & 0.30 \\ 0.01 & 0.99 \end{pmatrix},$$

where the first row represents the experiments starting with pure cultures of cell phenotype $CD44^+/CD24^-$ and the second row with cell phenotype $CD44^-/CD24^+$, respectively.

CELLTRANS estimates transition matrices for each of the cell state proportion matrices. This approach leads to 4 matrices P_2, P_4, P_6 and P_8 . In this case, all of the derived matrices do not require regularization, since all matrices are already

stochastic matrices. The transition matrix P is obtained by averaging over the derived transition matrices which yields

$$P = \frac{P_2 + P_4 + P_6 + P_8}{4} = \begin{pmatrix} 0.8868 & 0.1132 \\ 0.0007 & 0.9993 \end{pmatrix}.$$

The predicted cell state proportions of this transition matrix with the initial states from the experiments and the experimental data is plotted in Figure 16G.

The solution of the master equation (5.3) with pure $CD44^+/CD24^-$ cells in the beginning is

$$\begin{aligned} CD44^+/CD24^-(t) &= 0.0006e^{-0.1133t} + 0.9994 \\ CD44^-/CD24^+(t) &= -0.0006e^{-0.1133t} + 0.0006. \end{aligned}$$

SUMMARY. This case study demonstrates two potential applications of CELLTRANS. First, CELLTRANS can reveal rare cell transitions which is indicated by the estimated probability to convert from $CD44^-/CD24^+$ to $CD44^+/CD24^-$ of 0.0007 per day. Hence, transitions from $CD44^-/CD24^+$ to $CD44^+/CD24^-$ cells almost never occur suggesting a potential cell transition hierarchy.

Second, the experiments only cover the first 8 days after culturing, but the time period until the equilibrium proportions are reached is not clear from the beginning. Here, the predicted equilibrium distribution is given by

$$w^* = \begin{pmatrix} 0.0063 \\ 0.9937 \end{pmatrix},$$

which corresponds to an equilibrium proportion of 0.63% of $CD44^+/CD24^-$ and 99.37% of $CD44^-/CD24^+$. CELLTRANS can estimate the expected time until this equilibrium is reached.

Applying `timeToEquilibrium(input, c(1, 0), 0.01)` estimates the expected time to the equilibrium starting with a pure $CD44^+/CD24^-$ cell line composition with a tolerance deviation of 1%. With these parameters, CELLTRANS predicts a time of 39 days until the equilibrium proportions are reached.

5.3.5 Influence of the choice of the time step length τ

In order to demonstrate that the predictions and results of CELLTRANS are independent of the choice of the time step length τ , we varied τ in the case studies with data from [143] and [40]. We predicted the equilibrium distribution, the expected time to equilibrium and the predicted cell state composition after a certain time for different time step lengths. The results are summarized in Tables 9 and 10 and indicate that the predictions are not influenced by the choice of τ .

steps per day	predicted equilibrium (CSC / NSCC)	estimated time to equilibrium (days)	predicted distribution after 5 days starting with 100% CSC (CSC / NSCC)
50	0.6538 / 0.3462	20.46	0.799 / 0.201
24	0.6539 / 0.3461	20.5	0.799 / 0.201
10	0.6539 / 0.3461	20.5	0.8 / 0.2
5	0.6539 / 0.3461	20.6	0.8 / 0.2
2	0.6539 / 0.3461	21	0.8 / 0.2
1	0.654 / 0.346	21	0.801 / 0.199
0.5	0.6542 / 0.3458	22	0.83 / 0.17

Table 9: Predictions of CELLTRANS for different choices of the time step length τ for data from the case study in [147]. The original time-step for the analysis was 1 day. CSC = cancer stem cell, NSCC = non-stem cancer cell.

steps per hour	predicted equilibrium (adh / susp)	estimated time to equilibrium (hours)	predicted distribution after 2 hours starting with 100% adh (adh / susp)
60	0.6804 / 0.3196	4.97	0.712 / 0.288
30	0.6804 / 0.3196	5	0.712 / 0.288
10	0.6803 / 0.3197	5.1	0.713 / 0.287
5	0.6801 / 0.3199	5.1	0.713 / 0.287
1	0.679 / 0.321	6	0.726 / 0.274
0.5	0.678 / 0.322	8	0.741 / 0.259

Table 10: Predictions of CELLTRANS for different choices of the time step length τ for data from the case study in [40]. The original time-step for the analysis was 1 hour. adh= adherent cell type, susp = suspended cell type.

5.3.6 *Validation of the predicted equilibrium of CELLTRANS*

One important application of CELLTRANS is the prediction of the equilibrium of cell state proportions and the time needed to reach this equilibrium. In some experiments, the equilibrium proportions are not reached at the end of the experiment, e.g. in the investigation of the epithelial-mesenchymal transition in breast cancer introduced above [89]. Here, we show that CELLTRANS is able to make predictions beyond the duration of the experiments. In detail, CELLTRANS is able to reliably predict both the equilibrium cell state proportions and the time needed to reach this proportion. To demonstrate this, we performed a validation analysis based on the two case studies dealing with colon cancer cells [40, 147]. We used only a subset of the available data points to create predictions of the cell state proportions over time with CELLTRANS. We excluded late data points to mimic an experimental situation in which the equilibrium is not reached yet. The results of this validation study are illustrated in Figures 16H-I. It turns out that only a few data points are sufficient to reliably predict the equilibrium cell state proportions. Moreover, the predicted time to reach equilibrium proportions is very robust with respect to the choice of available data points. This investigation indicates that CELLTRANS is able to make predictions even beyond the time period of experiments and might therefore also support the planning of experimental time periods.

5.3.7 *Influence of non-pure initial cell state compositions*

As our case studies demonstrate, most FACS and flow cytometry experiments are based on pure initial cell state compositions. However, CELLTRANS can also be utilized to analyze experimental data with non-pure initial compositions. Here, we demonstrate this possibility and investigate whether the predictions and results are influenced by such an initial composition. For this, we reused data from [147] and [40] but excluded several of the first data points such that the initial cell state compositions are non-pure. We then utilized CELLTRANS to estimate a transition matrix based on these remaining data and compare the corresponding predictions to the original estimates derived from all available data. It turns out that both the evolution of cell line compositions and the equilibria are reliably predicted with non-pure initial cell state compositions. The results of this analysis are illustrated in Figure 17.

5.3.8 *A simulation study demonstrating matrix regularization*

The presented case studies so far do not require matrix regularization, i.e. the matrix roots that CELLTRANS calculates are already stochastic matrices. In order to demonstrate that such a regularization might be necessary, we introduce a

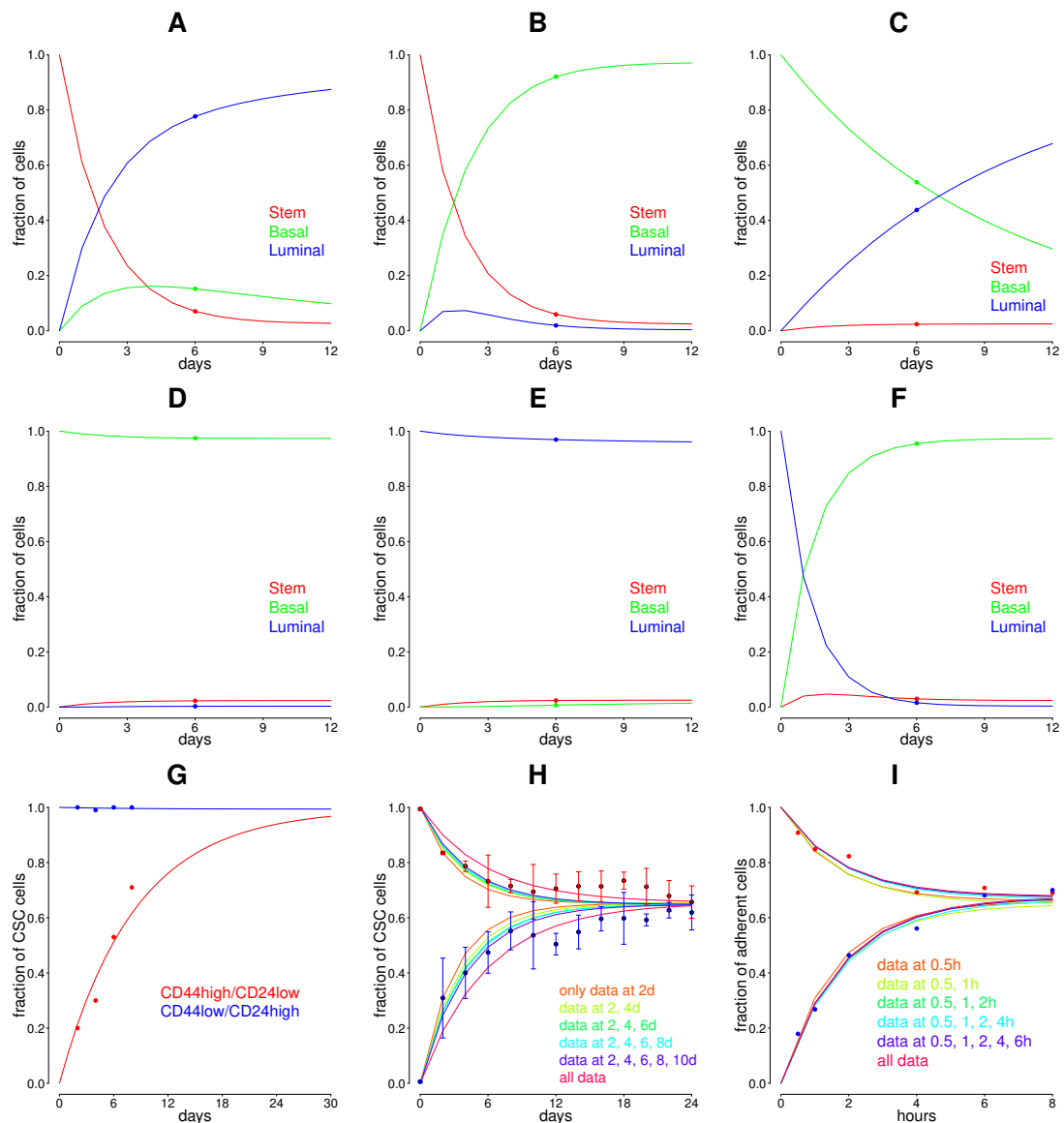


Figure 16: CELLTRANS model predictions and validations. A-G. We utilized CELLTRANS to analyze data of the evolution of cell state proportions of publicly available data. The data are plotted by colored dots. A-F. The data originates from [45] with three different cell states (stem, basal and luminal) from SUM149 and SUM159 breast cancer cell lines. The predictions of our analysis is plotted by colored curves. The color indicates the state of the cells at the beginning of the corresponding experiment. G. Analysis of composition data (dots) of HMLE cell lines [89] with cell states $CD44^{-}/CD24^{+}$ and $CD44^{+}/CD24^{-}$. The corresponding predictions are plotted as colored curves. H-I. We excluded several of the late data points from the data of the proportions of cancer stem cells (CSC) and non-stem cancer cells (NSCC) from [147] and the data with adherent and suspended cells from [40]. The predictions based on the remaining data is plotted, compare also with Figure 15. This investigation indicates that CELLTRANS is able to predict cell state proportions even beyond the duration of experiments and that only few data points are needed to reliably predict the equilibrium.

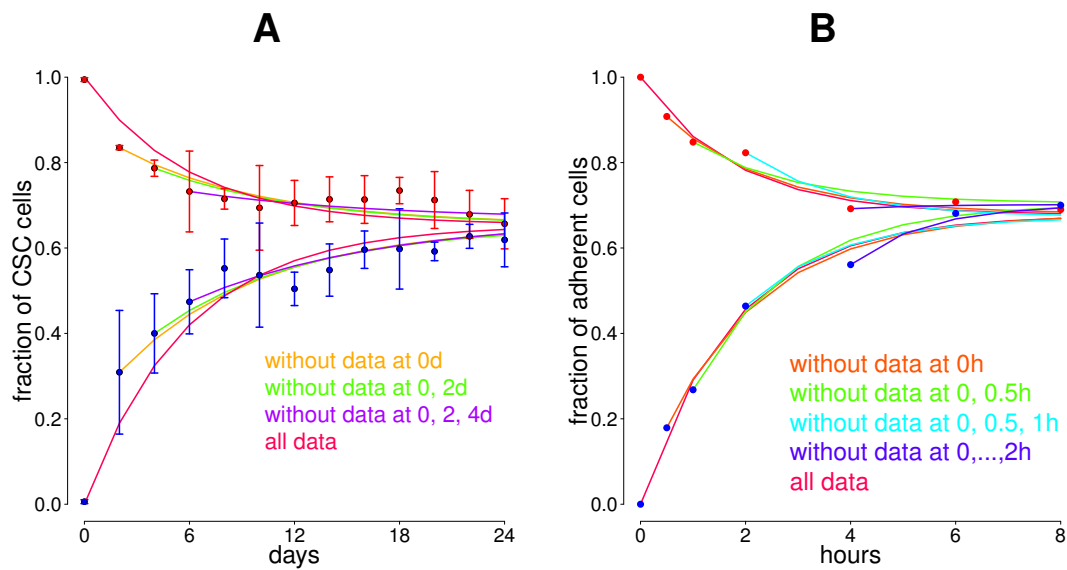


Figure 17: Influence of non-pure initial cell state compositions on the predictions of CELLTRANS. In order to obtain non-pure initial cell state proportions, we excluded several of the early data points, as indicated in the legends, from the case studies with data from [147] and [40]. The plotted predictions in both figures are based on the analysis of the remaining data points with CELLTRANS. The starting point of each curve indicates the initial cell state composition. *A.* Data from [147]. *B.* Data from [40].

simulation study and explain how the QOM-algorithm performs the necessary regularization.

In the simulation study we assume the existence of $m = 3$ cell states. We simulate three experiments starting with pure cultures of each cell state. Hence, the experimental initial matrix is

$$W^{(0)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

As time points we choose $t = 2\tau, 4\tau, 6\tau$. Here, τ is an arbitrary time step length.

We created an arbitrary transition matrix to describe the transition probabilities between the three cell states,

$$P_{sim} = \begin{pmatrix} 0.7739 & 0.1762 & 0.0499 \\ 0.0006 & 0.8210 & 0.1784 \\ 0.0937 & 0.0712 & 0.8351 \end{pmatrix}.$$

The MARKOV chain associated with this transition matrix has the steady state distribution

$$w_{sim}^* = (0.1872 \quad 0.3632 \quad 0.4496).$$

Then, we generated experimental data after times $2\tau, 4\tau$ and 6τ from the matrix powers P^2, P^4 and P^6 . We also added a normally distributed noise with mean zero and a standard deviation of 0.01 to reflect unprecise measurements. This led to the cell state proportion matrices

$$W^{(2)} = \begin{pmatrix} 0.6028 & 0.2769 & 0.1203 \\ 0.0087 & 0.7003 & 0.2910 \\ 0.1526 & 0.1224 & 0.7250 \end{pmatrix}, W^{(4)} = \begin{pmatrix} 0.3934 & 0.3819 & 0.2247 \\ 0.0492 & 0.5175 & 0.4333 \\ 0.1989 & 0.2465 & 0.5546 \end{pmatrix},$$

$$W^{(6)} = \begin{pmatrix} 0.2824 & 0.4050 & 0.3126 \\ 0.1044 & 0.4489 & 0.4467 \\ 0.2191 & 0.2862 & 0.4947 \end{pmatrix}.$$

Applying formula (5.1) to derive the matrix roots does not yield a stochastic matrix for all of these three matrices. The 4-th matrix root of $W^{(4)}$ is

$$\begin{pmatrix} 0.7822 & 0.1737 & 0.0441 \\ -0.0102 & 0.8205 & 0.1896 \\ 0.0943 & 0.0792 & 0.8265 \end{pmatrix},$$

which is not a stochastic matrix due to the negative entry in the second row. The QOM-algorithm transforms this matrix root into the stochastic matrix

$$P_4 = \begin{pmatrix} 0.7822 & 0.1737 & 0.0441 \\ 0.0000 & 0.8154 & 0.1846 \\ 0.0943 & 0.0792 & 0.8265 \end{pmatrix}.$$

The further derivation is continued with this regularized matrix root.

Finally, CELLTRANS derives the transition matrix

$$P = \begin{pmatrix} 0.7786 & 0.1708 & 0.0506 \\ 0.0000 & 0.8258 & 0.1742 \\ 0.0951 & 0.0694 & 0.8355 \end{pmatrix},$$

which is close to the originally generated P_{sim} . Furthermore, the steady state of the process associated with P is

$$w^* = \begin{pmatrix} 0.191 & 0.3644 & 0.4446 \end{pmatrix}.$$

The effect of the QOM regularization on the prediction of CELLTRANS in this case is visualized in Figures 18B-D.

5.4 DISCUSSION

Characteristic equilibrium proportions of distinct cell states are commonly observed *in vivo* and *in vitro*. Normal and cancerous cell lines exhibit and are able to maintain such an equilibrium [40, 45, 89, 143]. Understanding the mechanisms which are responsible for this observation is important to develop appropriate therapies against cancer. There is evidence that stochastic transitions between different cell states can lead to such an equilibrium. To infer the underlying transition dynamics and to quantify these transitions is a key to understand and control the origin of such an equilibrium.

We introduce CELLTRANS, an automated framework to deduce the transition probabilities between different cell states from FACS and flow cytometry experiments in which no differences in the proliferative properties between cell states are observed. The key assumption of the underlying mathematical model is that cells stochastically transition between different states and that the rates for these transitions only depend on the current state of the cell.

We point out that the transition probabilities can be derived on the basis of MARKOV chain theory by determining matrix roots from appropriately arranged data matrices and regularizing them to stochastic matrices if necessary. We discuss which mathematical challenges can occur and demonstrate how these challenges are handled by CELLTRANS. We utilize the QOM algorithm [74] to achieve

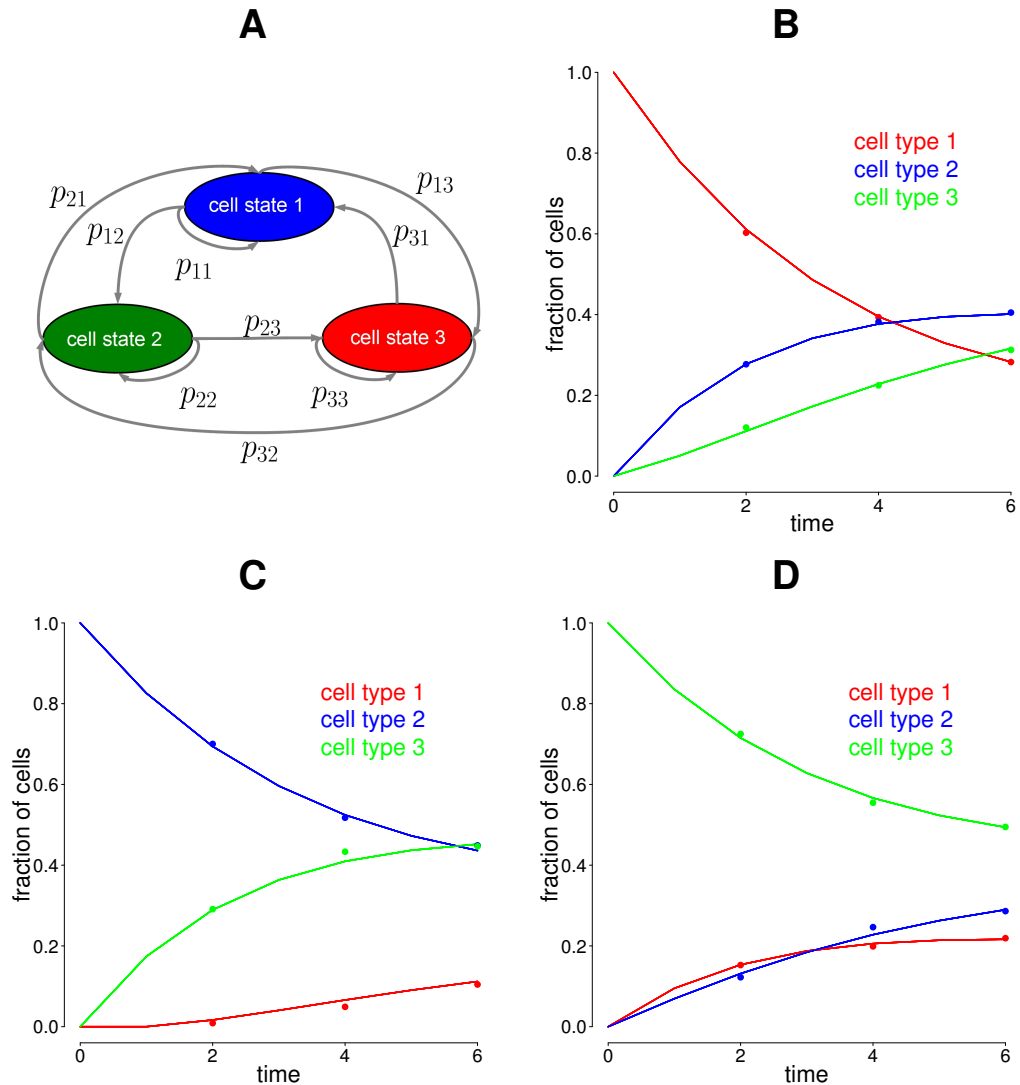


Figure 18: Cell state transitions and regularization of the matrix root. *A*. CELLTRANS estimates the transition matrix of the underlying discrete time MARKOV model. This matrix contains the transition probabilities between all cell states. Here, these transitions are illustrated for three fictive cell states. *B-D*. We simulated cell state proportion data for an experiment with three fictive cell states in which regularization of the matrix roots is necessary to obtain a stochastic matrix. The dots represent simulated data points, the solid curves the prediction of CELLTRANS. For details on the required matrix root regularization, see the main text.

matrix regularization and provide a simulation study which demonstrates that regularization might be necessary and how it can be achieved by the QOM algorithm.

In order to ensure a reliable estimation of cell state transitions, the cell state proportion data should be obtained on basis of a large number of cells. This ensures the validity of the estimation based on the law of large numbers [39]. We suggest to use at least 100 cells in each cell culture experiment to obtain the data for CELLTRANS.

Our stochastic approach on the basis of a MARKOV model can be translated into a system of first-order ODEs with the help of the master equation (5.3). The predictions of these ODEs are equivalent to those of the underlying MARKOV model of CELLTRANS. We demonstrate that CELLTRANS is able to predict the evolution of cell state proportions even more precisely compared to more complicated ODE models which utilize *in situ* parameter estimation, as for example in [143].

The analyzed case studies in this work demonstrate that CELLTRANS can be utilized to compare different cell types with respect to their ability to convert to other cell types, their frequency within equilibrium proportions and their position within an existing cell transition hierarchy. We showed that the main application of CELLTRANS is the quantitative inference of transitions between distinctive cell types. The resulting transition probabilities allow to estimate which transitions occur frequently, rarely or even almost never. Therefore, CELLTRANS might be able to reveal a hierarchy with respect to the importance of specific transitions for maintaining the observed equilibrium distributions.

The steady state derived by CELLTRANS can be interpreted as prediction of the equilibrium cell state proportions. Moreover, CELLTRANS is able to predict the duration from any initial experimental setup to such an equilibrium, see Figures 16H-I.

As demonstrated in Figures 16A-F, even patients classified to have the same type of tumor exhibit different equilibrium distributions. Hence, the dynamics of the cell state transitions are patient-specific. CELLTRANS is able to predict these patient-specific transitions and can therefore be utilized to reveal differences within the same tissue in different patients. These predictions might be a step towards individual therapies.

Although the presented case studies in this work deal with disease-related cell line experiments, CELLTRANS can also be utilized to analyze non-diseased cell lines, such as immunostained progenitor cells, as long as the underlying model prerequisites are fulfilled. Hence, CELLTRANS can be used to analyze all cell line experiments in which cell state transitions only depend on the current state of the cell and cell proliferation and death rates are equal.

We focus here on the case in which cell states exhibit similar proliferation and death rates. In principle, it would be possible to apply a similar approach also if this prerequisite is not fulfilled. This would first require to formulate

an extended and more complicated mathematical model which is a challenging task for future work.

Our case studies demonstrate that CELLTRANS is a valuable tool to model and quantify cell state transitions. Experimentalists only have to validate the model assumptions. Then, the whole process of mathematical modeling and estimation is automatized. CELLTRANS allows versatile experiments by being able to analyze also cell state data originating from non-pure initial cell state distributions. In summary, CELLTRANS is an automated tool that facilitates the analysis and interpretation of cell state proportion data from these experiments on basis of a MARKOV model for cell state transitions.

 QUANTIFICATION OF PHENOTYPIC CELL STATE TRANSITIONS BETWEEN GLIOBLASTOMA CELLS*

 Contents

6.1	Glioblastoma multiforme and the cancer stem cell hypothesis .	113
6.2	Material & methods	115
6.3	Results	117
6.4	Discussion	122

6.1 GLIOBLASTOMA MULTIFORME AND THE CANCER STEM CELL HYPOTHESIS

Glioblastoma multiforme (**GBM**) is the most malignant brain tumor and exhibits a large degree of heterogeneity at both the cellular and molecular level [57]. This degree of heterogeneity is a major challenge for effective therapies but the underlying mechanisms leading to its emergence are not understood yet. Progression of **GBM** is proposed to be caused by a sub-population of so-called cancer stem cells (**CSCs**) [80]. According to the classical **CSC** model, **CSCs** exhibit stem cell properties including self-renewal and multipotency which are demonstrated in experimental models in vivo in mice [134]. The **CSC** model proposes that tumor cells are hierarchically organized and follow a unidirectional differentiation process with **CSCs** at the top of the hierarchy. Only the **CSC** sub-population has the potential to induce tumor formation. Due to their unlimited self-renewal capacity and their ability to form all cells within a tumor, remaining **CSCs** after therapies like radiotherapy and chemotherapy are believed to be responsible for tumor recurrence [6]. Therefore, eradication of all **CSCs** within a tumor is considered as one approach to reach long-term disease-free survival. However,

* This chapter includes text and figures from a manuscript entitled *Stem cell-associated heterogeneity in Glioblastoma is a result of intrinsic tumor plasticity shaped by the microenvironment* submitted to *Cancer Cell*. Authors: Anne Dirkse, Anna Golebiewska, Thomas Buder, Petr V. Nazarov et. al. Author contribution: Thomas Buder conceived the model and performed the formal data analysis.

especially for GBM, it is unclear whether the classical CSC model is true. Recently, several studies suggest that tumor cells in GBM follow more likely a plasticity behavior in contrast to a strict hierarchical organization [118]. In this plasticity model, phenotypic heterogeneity is not a result of a one-way hierarchical structure but tumor cells can reversibly convert in a bidirectional manner between different phenotypic states. These studies suggest that CSCs do not constitute a defined cellular entity present across distinct patient tumors but rather represent a cellular state in response to micro-environmental conditions. The question whether the intratumoral heterogeneity in GBM is a result of a hierarchical differentiation process as proposed by the classical CSC model or a consequence of reversible phenotypic plasticity remains open.

One major challenge in investigating the CSC hypothesis in many tumors, also in GBM, is the experimental identification of CSCs. Several cell surface markers have been identified to characterize CSCs in GBM such as CD133, CD15/SSEA, CD44, A2B5 or CD49f [79]. However, surface marker expression may not be unique and there is currently no universal clear marker signature that reliably identifies the functional criteria of CSCs in GBM [11]. Moreover, tumors might contain more than one CSC clone which can be distinguished by different marker expression profiles and inter-patient heterogeneity might imply that CSCs cannot be defined universally in all GBMs.

In this chapter, we address the question if and to which extent tumor cells with stem cell properties in GBM follow a hierarchical organization or whether those states can be adopted by any cancer cell via reversible phenotypic state transitions. Moreover, we investigate the influence of micro-environmental conditions on this question. For this purpose, we utilize cell state proportion data from fluorescence activated cell sorting (FACS) experiments on the expression of the potential stem cell markers CD133, CD44, CD15 and A2B5. These data have been derived under normal and hypoxic conditions [19]. We utilize CELLTRANS (see Chapter 5) to quantify the cell state transitions between GBM cells under both environmental conditions. Furthermore, we interpret the estimated transition matrices with respect to potential underlying hierarchical transition structures.

It turns out that the estimated state transition probabilities and also the extent of the underlying transition hierarchy strongly depend on the environmental conditions. This finding is important for understanding GBM progression and for the successful design of novel therapies [19]. Moreover, we introduce and discuss additional features that could be implemented in CELLTRANS in the future.

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁	P ₁₂	P ₁₃	P ₁₄	P ₁₅	P ₁₆
CD ₁₃₃	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×	×	×
CD ₄₄	×	×	×	×	✓	✓	✓	✓	×	×	×	×	✓	✓	✓	✓
A ₂ B ₅	✓	✓	×	×	✓	✓	×	×	✓	✓	×	×	✓	✓	×	×
CD ₁₅	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓	×	✓

Table 11: Definition of 16 distinct GBM cell phenotypes depending on the marker expression of CD₁₃₃, CD₄₄, CD₁₅ and A₂B₅. Checkmarks indicate positive, x's low/negative expression of the corresponding marker.

6.2 MATERIAL & METHODS

Experimental setup and data

We utilize data on the evolution of cell state proportions of glioblastoma cell line experiments with CELLTRANS [15] in order to investigate the state transition behavior of these cells. For a description of CELLTRANS, see Chapter 5.

In [19], 16 distinct cell phenotypes named P₁, . . . , P₁₆ can be discriminated depending on the expression of the four potential stem cell markers CD₁₃₃, CD₄₄, CD₁₅ and A₂B₅, see Table 11. Multicolor flow cytometry experiments with cells from the in vitro glioblastoma cell culture NCH644 are performed under normal and hypoxic cell culture conditions. In detail, pure sub-populations of the 16 cell phenotypes have been cultured for 70 days and cell state proportions with respect to the 16 phenotypes have been measured after 20, 30 and 70 days in normoxia. In hypoxia, the purified sub-populations have been cultured for 60 days and the cell state proportions have been obtained at the end of the experiment.

Application of CELLTRANS

We apply CELLTRANS to analyze the phenotypic state proportion data under both normal and hypoxic conditions. The prerequisite for the application of CELLTRANS that the proliferation rates of the involved cell types are equal is fulfilled since differences in doubling times of the 16 cell phenotypes are not statistically significant [19]. The initial experimental cell state proportions under both environmental conditions can be described by the 16 × 16 identity matrix in which the *i*-th line corresponds to the experiment with pure cells of phenotype *i* in the beginning. See also Chapter 5 for technical details about the construction of data matrices.

We choose a time step length of 3 days for the analysis of the data which reflects the cell cycle time in normoxia [19] and round the time steps to the next integer. Hence, we obtain three cell phenotype matrices from the obtained data in normoxia: $W^{(7)}$ after 20 days, $W^{(10)}$ after 30 days and $W^{(23)}$ after 70 days. In hypoxia we construct only one matrix, namely $W^{(20)}$ for the measurements after 60 days.

By utilizing CELLTRANS, we estimate the transition probabilities between the 16 cell phenotypes. The interaction graph of the estimated transition matrix is visualized in Figure 19. These estimates allow to discriminate frequent and non-frequent state transitions and between hierarchical and non-hierarchical transition behavior.

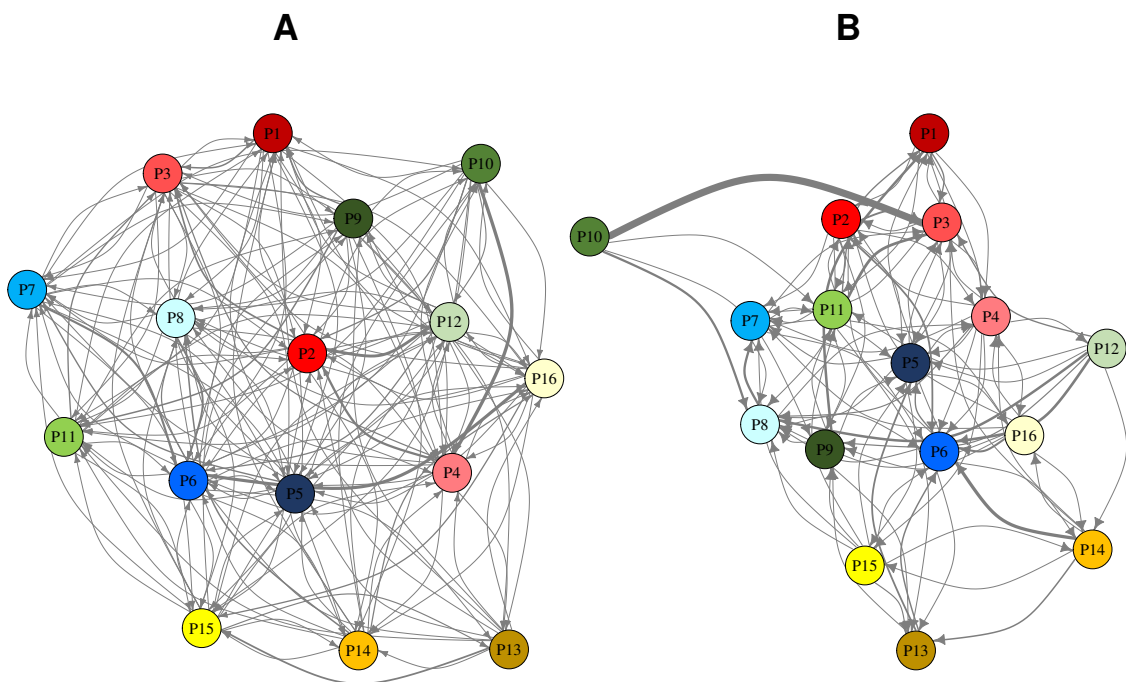


Figure 19: Interaction graphs of the estimated transition matrices of GBM cell state transitions. Arrows between sub-populations represent predicted direct state transitions between subpopulations, thickness of lines corresponds to transition probability. See also Tables 16 and 17 for the transition matrices. A. Normoxia, B. Hypoxia.

In detail, if the estimated MARKOV chain is irreducible, each state can transit directly or via intermediate states into any other state. This behavior implies that the state transitions are non-hierarchical and therefore reversible. In contrast, a tree structure corresponds to a perfect hierarchy. Intermediate structures are possible as well, which imply some degree of hierarchy between transient states and recurrent states. Moreover, on the basis of the estimated transition matrix it

is possible to predict the equilibrium cell composition and the time needed to reach this equilibrium with help of CELLTRANS.

Marker dependency calculations

In order to investigate whether the cell state transitions in normoxia and hypoxia are dependent on the current marker expression, transition probabilities considering only a single marker are calculated. From the measured composition data, the percentage of all 16 phenotypes which are positive or low/negative for a single marker expression are added. This is performed for each of the four markers and results in four individual 2×2 experimental setup matrices. Subsequently, CELLTRANS is applied to estimate four transition matrices between the corresponding two cell phenotypes. These transition matrices can be utilized to calculate a 16×16 transition matrix under the assumption of transition independence by multiplying the corresponding transitions probabilities for each of the four markers. The resulting transition matrices assuming marker-independent transitions are provided in Tables 16 and 17. Subsequently, the steady state of this transition matrix is calculated and the fractions of each subpopulation with the same expression of a single marker are added to obtain the steady state proportion for each individual marker assuming independent transitions. The estimated four 2×2 transition matrices considering a single marker and the steady state proportions for each individual marker assuming independent transitions are provided in Tables 12 and 13.

6.3 RESULTS

Glioblastoma sub-populations do not undergo hierarchical state transitions

The estimation of the transition matrix between the 16 defined cell phenotypes in normoxia suggests 175 of 240 possible direct state transitions leading to a change of the phenotype, see Figure 19. Importantly, the estimated transition matrix appeared irreducible with no obvious bottlenecks, i.e. each phenotype can transition to all other phenotypic states directly or through several transitions. Moreover, the transition matrix suggests that there are no hierarchies or sub-hierarchies between the phenotypes which is also reflected by a KRACKHARDT hierarchy score of 0 [73]. In particular, the quantification of the transition probabilities, see Appendix Table 16, allows to identify rare and frequent transitions.

The validity of the predictions is strongly supported by two observations. First, the theoretical equilibrium composition, i.e. the steady state of the underlying MARKOV chain, assort well with the original tumor composition which is not utilized in the estimation of the transition matrix, see Figure 20. Second, we have designed two heterogeneous compositions of the 16 phenotypes and

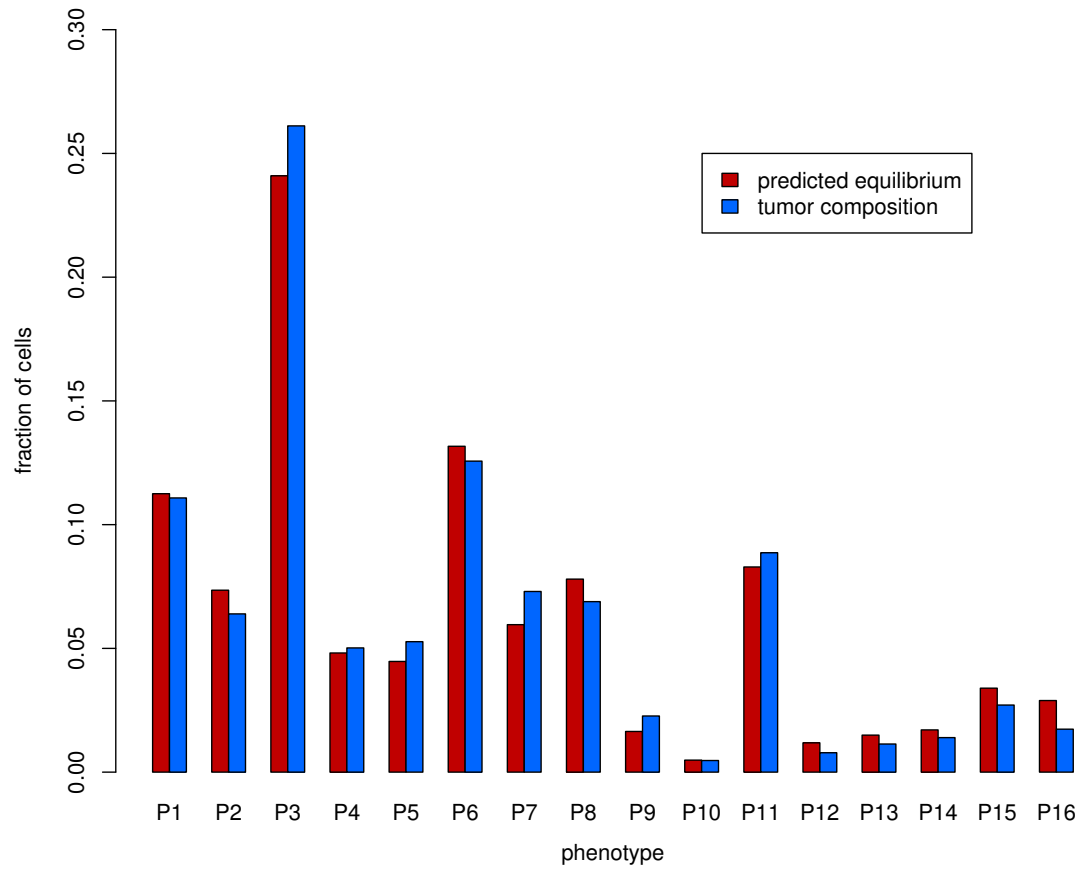


Figure 20: Predicted [GBM](#) equilibrium composition and the original tumor composition.

predicted equilibrium composition and the needed time to reach it as 39 days. Indeed, the two combinations evolved after 39 days in culture to a composition close to the predicted equilibrium, see Appendix Figure 23.

Glioblastoma cell state transitions in hypoxia

Our model analysis suggests that direct state transitions between the 16 GBM cell phenotypes are more restricted under hypoxic conditions by predicting 103 of possible 240 possible direct transitions, see also Figure 19. Moreover, the estimated transition matrix is reducible with P₁₀ (CD₁₃₃-CD₄₄-A2B5+CD₁₅+) as transient state, i.e. no phenotype can transit into this state and P₁₀ is depleted from the final hypoxic equilibrium. This prediction assorts well with the experimental observation of a very low proportion of P₁₀ upon long term hypoxia and its low self-renewal potential at the initial FACS passage. These results suggest that P₁₀ could be considered as a CSC/progenitor-like state at the apex of hierarchical organization in hypoxia [19].

Moreover, our analysis yields that 13 of 16 sub-populations are predicted to require over 250 days to reach equilibrium. This implies that the experimental phenotypic compositions after 60 days have not reached equilibrium yet. Indeed, statistical differences between the 16 experimental cultures are observed with respect to the cell state compositions after 60 days [19]. One possible explanation for the slower evolution of the composition to an equilibrium might be a strongly decreased proliferation of GBM cells in severe hypoxia with a cell cycle time of about 7 days [19].

Evidence for marker-dependent transitions in hypoxia

We investigate whether the cell state transitions in normoxia and hypoxia are dependent on the current marker expression. For this purpose, we compare the predicted equilibrium proportions under marker-independent transitions with the corresponding fractions obtained from the originally derived transition matrix, see Tables 12 and 13.

It turns out that both estimates are in good accordance in normoxia for all cell phenotypes. In contrast, under hypoxic conditions there is a large deviation of the predicted equilibrium compositions for the expression of CD₄₄. This result suggests that the state transitions under hypoxia are dependent on the expression of CD₄₄ which requires further experimental investigations regarding potential underlying mechanisms.

Transition probabilities		Predicted equilibrium independent transitions		Originally predicted equilibrium
	CD133+	CD133-		
CD133+	97.48	2.52	77.18	79
CD133-	8.52	91.48	22.82	21
	CD44+	CD44-		
CD44+	95.96	4.04	46.17	41
CD44-	3.46	96.54	53.83	59
	A2B5+	A2B5-		
A2B5+	79.69	20.31	42.91	41.6
A2B5-	15.27	84.73	57.09	58.4
	CD15+	CD15-		
CD15+	93.11	6.89	36.43	39.5
CD15-	3.95	96.05	63.57	60.5

Table 12: Single marker estimates and equilibrium composition under the assumption of marker-independent transitions of GBM phenotypes in normoxia. The first two columns contain the estimated transition probabilities in % per 3 days between single marker GBM cell phenotypes in normoxia. The third column provides the equilibrium proportions from the transition matrix assuming independent marker changes, see Appendix Table 18. The last column contains the equilibrium proportions derived from the originally estimated transition matrix in Appendix Table 16.

Transition probabilities		Predicted equilibrium independent transitions	equi-assum-marker-	Originally predicted equilibrium	pre-
	CD133+	CD133-			
CD133+	99.7	0.3	97.63	96.19	
CD133-	12.4	87.6	2.37	3.81	
	CD44+	CD44-			
CD44+	98.87	1.13	73.21	55.79	
CD44-	3.09	96.91	26.79	44.21	
	A2B5+	A2B5-			
A2B5+	94.05	5.95	40.68	33.92	
A2B5-	4.08	95.92	59.32	66.08	
	CD15+	CD15-			
CD15+	92.45	7.55	16.27	18.29	
CD15-	1.47	98.53	83.73	81.71	

Table 13: Single marker estimates and equilibrium composition under the assumption of marker-independent transitions of GBM phenotypes in hypoxia. The first two columns contain the estimated transition probabilities in % per 3 days between single marker GBM cell phenotypes in hypoxia. The third column provides the equilibrium proportions from the transition matrix assuming independent marker changes, see Appendix Table 19. The last column contains the equilibrium proportions derived from the originally estimated transition matrix in Appendix Table 17.

6.4 DISCUSSION

Heterogeneity within GBM tumors is a major reason for therapy failure and cancer progression and increased heterogeneity positively correlates with decreased survival [108]. Although there is large evidence that cancer cells with stem cell properties exist within GBM, the evidence for a hierarchical organization due to distinct CSC types is highly controversial. Additionally, it is still not clear if the phenotypic heterogeneity of GBM represents an intrinsic property or whether it results from exposure to different micro-environmental cues [19].

In this chapter, we utilize data from FACS experiments under different environmental conditions to quantify the phenotypic state transitions of GBM cells in vitro. In particular, we investigate the transition hierarchy of GBM stem cell-associated cell phenotypes and enquire the dependency of these transitions on the current marker expression. Our estimates of the transition frequencies allow to discriminate rare and frequent transitions, see Tables 16 and 17. These results have the potential to detect important transitions which could be therapeutical targets to disturb tumor equilibrium compositions [19].

Moreover, our analysis suggests that phenotypic heterogeneity within GBM is achieved via stochastic state transitions between distinct phenotypes without any hierarchy or sub-hierarchy under normoxic conditions. In contrast, the estimated transition matrix under hypoxia exhibits some degree of hierarchy showing that the micro-environment of the tumor cells has a crucial impact on the transition behavior. This conclusion is further supported by evidence for marker-dependent transitions in hypoxia which we cannot confirm in normoxia. In summary, these results indicate an important role of the micro-environment on the quantity and structure of GBM cell state transitions. Further biological background and results of additional experiments investigating the heterogeneity within GBM tumors are provided in [19].

In this chapter, we also demonstrated further potential conclusions that can be drawn from the application of CELLTRANS which go beyond the introduced main applications summarized in Chapter 5. We show that the structure of the underlying estimated transition matrix allows to infer the degree of transition hierarchy between the involved cell types. Moreover, the analysis of the normoxia and hypoxia data in this chapter illustrates how CELLTRANS can quantify the cell state transitions of the same cell line under different environmental conditions. The comparison of such data might allow to test the efficacy of certain drugs in vitro or to infer patient-specific characteristics. Furthermore, the demonstrated dependency analysis of the cell state transitions could allow to identify markers which have a substantial influence on the overall phenotypic transition behavior of the cells. For example, our analysis suggests a special role of CD44 of the investigated GBM cell line under hypoxia which suggests further experiments to unravel this role. Regarding the dependency analysis, a more rigorous underlying theoretical framework would be one potential extension for the future, e.g.

based on a statistical test. All of these discussed applications represent promising features that could be implemented in CELLTRANS in the future such that the corresponding analyses are automatically performed.

CONCLUSIONS

Contents

7.1	Summary	125
7.2	Novelty and relevance of the results	126
7.3	Outlook	127

7.1 SUMMARY

In this thesis, we infer the role of genetic and phenotypic cell changes on tumor development with the help of MARKOV chain models which are calibrated with tissue-scale data. Although the biological knowledge about tumor development is continually growing, a complete understanding is still a major challenge and the numbers of diagnosed cancers will largely increase in the next years [92]. Mathematical models provide a helpful tool to shed light on the complex interactions of tumor and wild-type cells and their consequences on tumor emergence on the tissue scale. In this thesis, we formulate and analyze cell-based models which are motivated by specific underlying biological questions within the tumor context.

Highly complex processes on the intra-cellular, cellular and tissue scale determine the fate of tumor development and a unifying model of all these processes remains elusive. We concentrate on cell-based models since observable phenomena on the tissue scale like the tumor regression behavior of tumors, the emergence of different tumor subtypes and the evolution of tumor heterogeneity allow a direct quantification of underlying cellular mechanisms. Our modeling approach is thereby characterized by the incorporation of sufficiently many details in order to obtain biological predictions while keeping the models simple to enable a rigorous analytical investigation. Importantly, we make quantitative predictions that can either be experimentally tested or result in novel biological hypotheses. In particular, the models in this thesis allow to estimate the tumor regression probability of pilocytic astrocytoma (Chapter 3), suggest and estimate the size of tumor-originating niches in different tissues (Chapter 4) and

predict the frequency of phenotypic cell changes (Chapter 5) which can be used to infer underlying transition hierarchies of glioblastoma cells (Chapter 6).

In the first part, we utilize MORAN models to describe cellular processes like competition, genetic alterations, proliferation and death. The design of the models allows to describe the quantity of interest, i.e. the absorption probability of the underlying process, solely in dependency of a single parameter, namely the corresponding risk coefficients γ and γ_{1D} . The derivation of these absorption probabilities is already quite involved and would have been even more complicated or even not possible in an analytical way if the models are formulated in a more complex manner.

In the second part of this thesis, we describe cell state transitions as discrete time MARKOV chain which serves as a minimal model for the evolution of cell state proportions. The introduced case studies demonstrate the applicability of the model although no experimentally measurable parameters of cell properties are utilized. In contrast, the simplicity is an advantage by enabling the application also without such experimental measurements and the predictions might even be better compared to complicated models which incorporate such parameters as demonstrated in Chapter 5.

7.2 NOVELTY AND RELEVANCE OF THE RESULTS

In the first part of this thesis, we extended established MORAN models and derived novel analytical results which are not only relevant for the biological questions investigated in this thesis but also for other applications. In order to describe the emergence of different tumor subtypes we altered the established MORAN model with mutations to a stochastic process with two absorbing states representing these subtypes. The analysis of the resulting process in Chapters 3 and 4 resulted in the absorption probabilities considering different spatial cell arrangements. In particular, we find that, assuming a biologically motivated parameter regime, the absorption probabilities depend solely on the risk coefficients γ and γ_{1D} , respectively. Moreover, we show that these asymptotic results provide reasonable approximations for relatively small finite values of the population size which has several applications also beyond the scope of this thesis. For example, these novel insights allow to distinguish different regimes in the MORAN model not only asymptotically but also for small values of the population size N . Our theoretical results in Chapter 4 obviate the need of ad-hoc rules for the application of asymptotic results for finite values of N and might for instance improve estimates about the waiting times until two mutations accumulated in some population, see for example [28].

In the second part, we demonstrated in which way our newly developed R package CELLTRANS automatizes the whole process of mathematical modeling and estimation of cell state transitions from experimental data. Experimentalists can directly utilize the publicly available package to analyze cell state transition

# cell types	analytical results	no space	space
2	absorption probabilities & times	-	novel MORAN model on a crypt-like structure (Chapter 2)
3	absorption probabilities	extended free model (Chapter 3)	space-extended MORAN model (Chapter 4)

Table 14: Novel analytical results for the MORAN models in this thesis.

data. Hence, we contribute to bridge the gap between the experimental analysis of cell state data and mathematical modeling. Moreover, we utilized the QOM algorithm [74] to achieve probability matrix regularization which is frequently used in credit risk models. We demonstrated that this technique proves useful whenever transition matrices are approximated and need to be regularized.

7.3 OUTLOOK

We demonstrate in this thesis in which way highly complex biological processes on the cellular scale can be abstracted to mathematical models which are motivated by specific biological problems. The applications of extended MORAN models in the first part of this thesis illustrates the versatility of this model class and exemplifies how new biological hypotheses can be generated with the help of these models. The estimated tumor regression probabilities of PA in Chapter 3 motivate further clinical studies that include volumetric data and form a statistical basis for clinical decision rules. Similarly, our proposal of the existence of a tumor-originating niche in Chapter 4 has far reaching clinical consequences. The existence of these niches and the estimated sizes should be further investigated both from a modeling and an experimental perspective. From a mathematical point of view, a rigorous derivation of the calculated determinant in the proof of Theorem 3.4 remains a task for future work. Moreover, the absorption probabilities of the one-dimensional MORAN model which have been approximated in Theorem 4.3 might be derived in an exact manner in the future.

Furthermore, we hope that our developed R package CELLTRANS proves useful for experimentalists and contributes to a deeper understanding of the mechanisms which underlie phenotypic plasticity and their consequences. Promising extensions of CELLTRANS include the automatization of the analysis of transition hierarchies and the possibility to apply this package if cells do not divide equally often.

I hope that this thesis contributes to bridging the gap between mathematical modeling and the analysis of experimental data in the cancer context by demonstrating possible ways to utilize observable quantities on the tissue scale to infer mechanisms on the cellular scale.

A

APPENDIX

A.1 RESIDUAL CEREBELLAR POSTOPERATIVE PILOCYTIC ASTROCYTOMA VOLUME AND OUTCOME

residual tumor size (cm ³)	outcome	reference
0.1	tumor regression	[44]
0.1	tumor regression	[125]
2	tumor regression	[44]
2	tumor regression	[44]
2	tumor regression	[44]
2	tumor regression	[44]
2.5	tumor regression	[125]
6	tumor regression or stable	[44]
8.8	tumor regression or stable	[44]
10	tumor growth	[44]
13.6	tumor growth	[125]
16.8	tumor growth	[44]
43.2	tumor growth	[44]

Table 15: Residual cerebellar postoperative pilocytic astrocytoma volume and outcome. This table contains the results of our literature research regarding volumetric data of residual cerebellar pilocytic astrocytoma and the corresponding patient outcome.

A.2 CELLTRANS VIGNETTE

Installing CELLTRANS

CELLTRANS is freely available via GitHub (<http://github.com/tbuder/CellTrans>). In order to install it directly type within R

```
if (!require("devtools")) install.packages("devtools")
library("devtools")
install_github("tbuder/CellTrans", build_vignettes=TRUE)
```

Note that you need the *devtools* package in order to install packages from GitHub and that it might be necessary to adjust your connection settings if you use a proxy server.

Then, load the package with

```
library("CellTrans")
```

To demonstrate how CELLTRANS is used directly within R, we utilize FACS data from experiments with two different cancer cell subpopulations in colon cancer [147]. The data describe the evolution of the proportion of cancer stem cells (CSC) and non-stem cancer cells (NSCC) during cell line experiments in time. Two experiments have been performed with pure initial cell states in the beginning and the evolution of the proportions has been measured each 2 days for 24 days.

Data preparation

In order to analyze data from FACS and flow cytometry experiments in CellTrans, they have to be appropriately formatted as text files. These text files contain line-wise the cell state proportions after the corresponding time of the two cell line experiments, see <https://github.com/tbuder/CellTrans>. For example, the content of 'W6.txt' is

```
0.7323 0.2677
0.4740 0.5260
```

meaning that the experiment with pure CSC cells in the beginning evolved to 73.23% CSC and 26.77% NSCC cells after 6 days and the experiment with pure NSCC cells evolved to 47.4% CSC and 52.6% NSCC, respectively.

A.2.1 *Reading the data and configuring CELLTRANS*

All relevant information for the analysis and visualization of the data with CELLTRANS can be read by utilizing the function `READEXPERIMENTALDATA()`. It is recommended to store these data in a variable, e.g.

```
my_CellTransdata<- readExperimentalData()
```

This command opens a sequence of windows asking for the details of the experimental setup, see Figure 21.

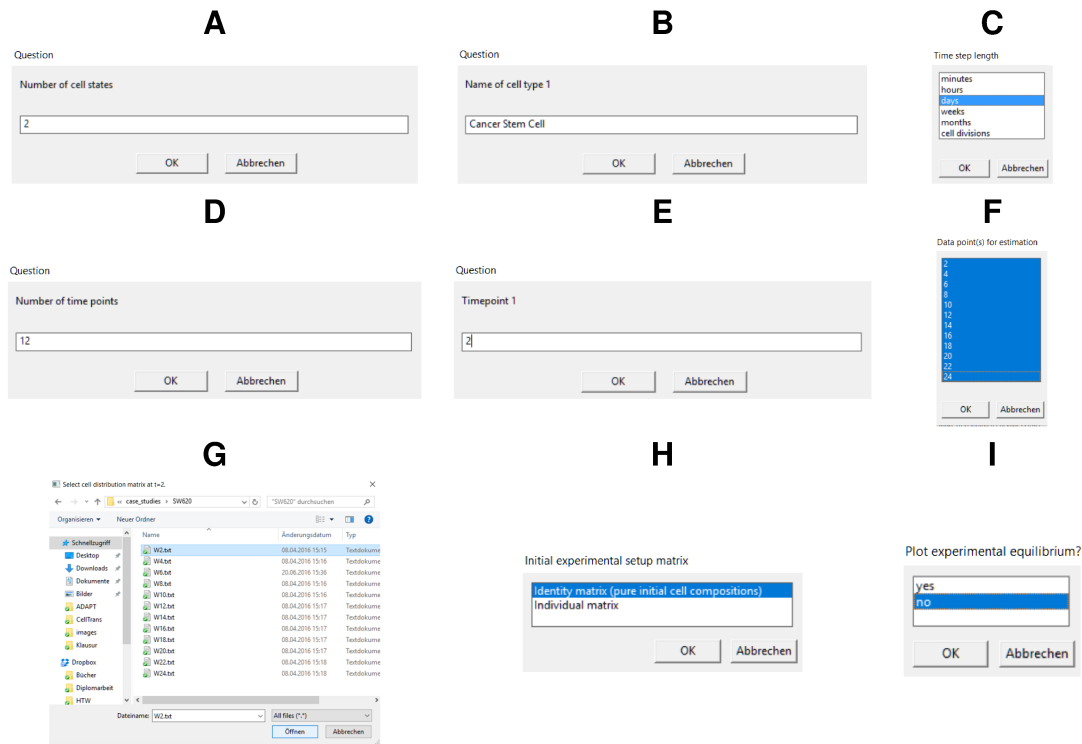


Figure 21: How to read data into CELLTRANS. *A.* The number of cell states m , *B.* the names of the states, *C.* the time unit of measurement, *D.* the number of time points and *E.* the time points of measurements are read into the R package. *F.* The data points considered for the estimations can be chosen and *G.* the experimental data matrices corresponding to the time points have to be selected. *H.* The initial cell state distributions of the experiments have to be chosen. If the experiments do not use pure initial cell state proportions, an individual matrix containing row-wise the initial cell state compositions can be provided. *I.* Within plots, the experimental equilibrium can be added at a chosen time point.

Now, all relevant data are read and can be used for the analysis and visualization of the data.

Deriving cell state transition probabilities

The function `CELLTRANSITIONS()` derives and prints an estimate of the transition probabilities between the cell states and an estimation of the equilibrium composition. Applying this function to the read data, i.e.

```
celltransitions(my_CellTransdata)
my_CellTransdata<-readExperimentalData()
```

allows to select the data points considered for the estimation, see Figure 21F.

`CELLTRANS` prints the estimated transition probabilities between both cell states and the predicted equilibrium distribution.

```
[1] "Results of CellTrans"
[1] "#####"
[1] "used timepoints:"
[1] "2" "4" "6" "8" "10" "12" "14" "16" "18" "20" "22" "24"
Markov Chain
A 2 - dimensional discrete Markov Chain defined by the following states:
Cancer Stem Cells, Non-Stem Cancer Cells
The transition matrix (by rows) is defined as follows:
                Cancer Stem Cells Non-Stem Cancer Cells
Cancer Stem Cells      0.9455135          0.05448652
Non-Stem Cancer Cells  0.1029900          0.89701002

[1] "predicted equilibrium distribution"
                Cancer Stem Cells Non-Stem Cancer Cells
[1,]           0.6540022          0.3459978
[1] "#####"
```

Estimating time to equilibrium

In order to estimate the time needed from a specific initial cell state composition to the equilibrium composition, the function

```
timeToEquilibrium(input,initialDistribution, tol)
```

can be utilized. The first parameter represents the read data from the function `readExperimentalData()` and the second parameter describes the initial cell state proportion. This initial cell state composition has to be described as a vector in R. The third parameter `TOL` describes the maximum deviation of the proportion of each cell state from the equilibrium distribution.

To illustrate this application, consider 50% of CSC and 50% of NSCC as initial composition in above introduced example. This composition can be described in R as vector `c(0.5, 0.5)`. Setting a tolerance deviation of 0.01 yields an expected time of 16 days to equilibrium:

```
> timeToEquilibrium(my_CellTransdata,c(0.5,0.5),0.01)
[1] "Expected time to equilibrium: 16 days"
[1] "Predicted cell state proportions at this time:"
      [,1]      [,2]
[1,] 0.6440747 0.3559253
```

A.2.2 Creating Plots

Another feature of CellTrans is the creation of plots of the data and the corresponding predictions. The function

```
celltrans_plotPDF(input)
```

allows to choose the data points considered for the estimation and asks whether the experimentally measured equilibrium should also be plotted, see Figure 21I. If yes, a text file containing the experimentally measured equilibrium proportions has to be selected and is then plotted at a time point of your choice.

Subsequently, the experiments for which a plot shall be created can be selected. For each selected experiment one plot is created with the corresponding evolution of the cell state proportions and the predicted evolution estimated by CELLTRANS. Furthermore, you can choose which cell states shall be included in the plot, see Figure 22. The function `celltrans_plot()` offers the same options,

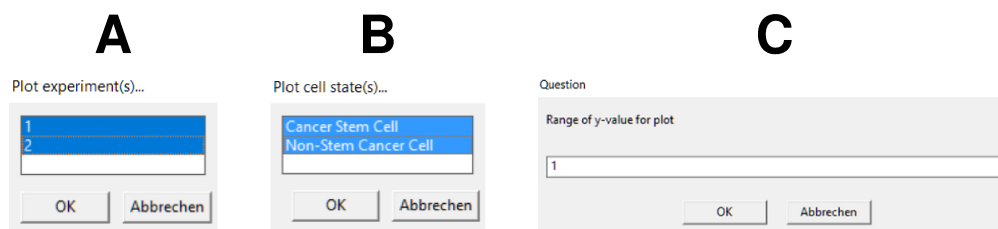


Figure 22: Options for plots in CELLTRANS.

but directly produces a plot in R without creating a pdf.

A.3 ADDITIONAL MODEL PREDICTIONS FOR GBM CELL STATE TRANSITIONS

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P1	86.81	1.08	6.52	0.83	1.22	0.11	0.25	0.47	0.00	0.00	2.07	0.00	0.01	0.00	0.00	0.63
P2	0.66	78.75	7.24	3.52	0.00	0.55	1.34	3.52	0.00	0.10	0.96	0.00	0.00	0.05	0.48	2.80
P3	3.84	1.19	88.11	1.99	0.40	0.35	0.35	0.73	0.33	0.01	2.24	0.22	0.04	0.00	0.00	0.19
P4	0.00	9.90	5.27	71.42	0.65	2.91	1.02	1.92	0.00	2.03	0.03	2.63	0.53	1.36	0.00	0.34
P5	1.93	0.00	4.40	0.00	78.63	2.46	5.99	0.16	0.00	0.00	1.66	0.67	0.00	0.00	0.93	3.17
P6	0.69	1.76	1.70	1.95	1.12	83.68	1.68	5.36	0.18	0.00	0.22	0.11	0.27	1.23	0.05	0.00
P7	0.36	0.47	1.29	0.03	4.44	9.22	80.93	0.42	1.00	0.65	0.00	0.00	0.00	0.56	0.64	0.00
P8	0.00	0.00	3.21	0.49	1.05	8.23	4.13	79.34	0.00	0.00	0.35	0.00	0.00	0.00	0.42	2.77
P9	6.11	1.58	0.00	2.47	0.92	0.00	0.24	1.70	77.27	0.40	6.50	0.00	2.68	0.14	0.00	0.00
P10	0.24	5.56	3.15	14.38	2.96	3.23	0.00	2.35	0.00	56.37	1.61	4.67	0.00	0.00	1.08	4.41
P11	1.39	1.79	5.79	0.00	0.00	0.17	0.43	0.00	1.80	0.00	85.91	0.53	0.00	0.00	2.19	0.00
P12	4.20	12.45	0.57	0.56	0.22	5.54	0.00	0.67	0.24	2.81	0.53	71.95	0.19	0.00	0.00	0.06
P13	0.66	0.81	0.00	0.17	1.43	1.53	0.00	0.00	1.66	0.00	2.01	0.00	81.00	2.68	7.58	0.48
P14	0.71	2.63	0.00	0.08	0.20	2.52	0.00	6.56	0.00	0.91	0.76	0.63	0.00	78.26	1.35	5.40
P15	0.00	0.00	1.36	0.15	4.09	0.52	0.74	1.43	0.65	0.01	0.83	0.40	4.87	1.17	81.18	2.61
P16	0.82	0.40	1.71	4.20	0.09	13.53	0.00	2.99	0.37	0.29	0.00	0.60	0.00	0.79	5.58	68.63

Table 16: Estimated transition probabilities per normoxia cell cycle length of 3 days in % between the 16 GBM cell phenotypes in normoxia.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P1	90.01	0.57	5.77	1.14	2.12	0.26	0.01	0.00	0.00	0.00	0.09	0.00	0.00	0.04	0.00	0.00
P2	0.03	77.22	8.17	1.66	0.92	7.71	1.63	1.49	0.27	0.00	0.89	0.00	0.00	0.00	0.00	0.00
P3	2.68	0.17	95.88	0.78	0.09	0.00	0.33	0.00	0.00	0.00	0.04	0.03	0.00	0.00	0.00	0.00
P4	0.00	2.37	1.92	83.53	0.49	1.47	3.91	4.92	0.02	0.00	0.00	0.18	0.00	0.00	0.97	0.22
P5	0.40	0.00	0.74	0.35	91.56	0.00	4.55	0.34	0.00	0.00	0.00	0.00	0.00	0.00	1.44	0.62
P6	0.00	1.20	0.08	0.00	5.95	87.97	0.00	4.52	0.00	0.00	0.00	0.00	0.15	0.13	0.00	0.00
P7	0.00	0.00	1.66	0.00	2.60	1.40	92.67	1.54	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00
P8	0.00	0.00	0.00	0.73	0.00	2.82	9.97	86.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P9	10.86	0.00	0.00	0.00	4.38	0.00	0.00	1.09	79.49	0.00	0.92	0.00	3.25	0.00	0.00	0.00
P10	0.00	0.00	30.94	0.00	0.00	0.00	0.00	9.76	0.74	58.01	0.54	0.00	0.00	0.00	0.00	0.00
P11	5.78	0.15	11.98	0.00	0.00	0.00	1.53	0.00	1.12	0.00	78.32	0.03	0.37	0.00	0.00	0.72
P12	0.00	0.00	5.85	0.00	0.55	11.40	0.00	10.80	5.10	0.00	2.04	62.78	0.00	1.48	0.00	0.00
P13	0.00	0.00	0.00	0.00	7.72	6.25	0.00	0.00	0.48	0.00	0.00	0.00	85.55	0.00	0.00	0.00
P14	0.00	0.00	3.21	0.31	0.00	14.05	0.00	0.00	0.00	0.00	0.00	0.00	6.53	75.10	0.74	0.06
P15	0.00	0.00	0.00	0.00	0.30	1.23	4.03	1.58	0.74	0.00	1.63	0.00	4.93	0.31	85.27	0.00
P16	0.00	1.59	0.00	2.85	0.00	8.76	1.64	0.73	2.22	0.00	0.00	0.00	0.00	2.25	0.00	79.96

Table 17: Estimated transition probabilities per 3 days in % between the 16 GBM cell phenotypes in hypoxia.

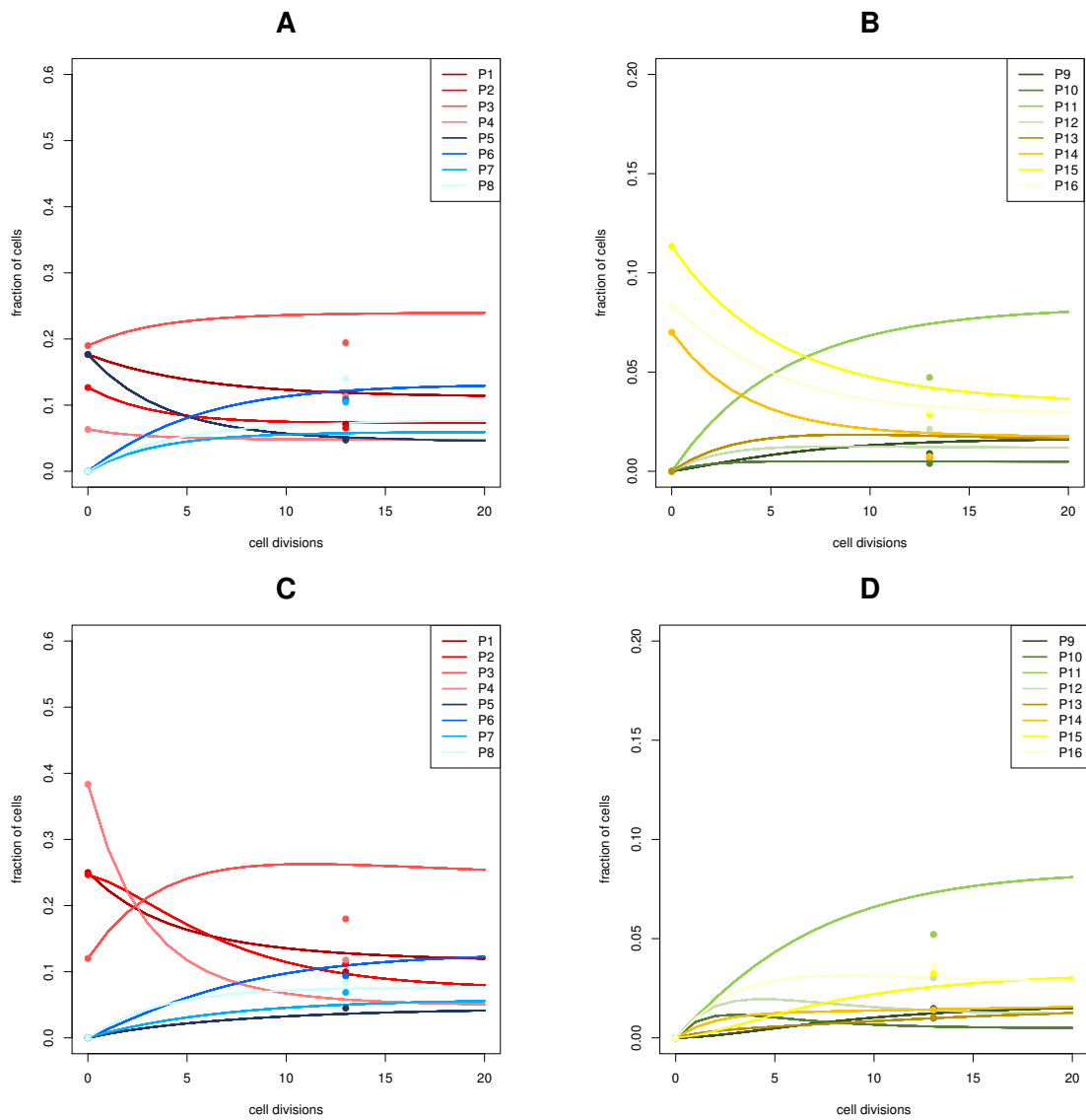


Figure 23: Estimated and measured GBM phenotypic compositions from experiments with two different initial compositions of GBM phenotypes in normoxia. We utilized CELLTRANS to estimate the time needed to equilibrium for two given different initial cell population compositions. The solid lines correspond to the predicted evolution of the cell phenotype compositions and the dots indicate the measured composition after the predicted time to the equilibrium composition. A-B. Composition P1 - P16: 53, 38, 57, 19, 53, 0, 0, 0, 0, 0, 0, 0, 0, 0, 21, 34, 25 cells, predicted time to equilibrium: 13 cell divisions. C-D. Composition P1 - P16: 75, 74, 36, 115, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 cells, predicted time to equilibrium: 13 cell divisions.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P1	72.03	2.96	18.36	0.75	2.58	0.11	0.66	0.03	1.86	0.08	0.47	0.02	0.07	0.00	0.02	0.00
P2	5.17	69.82	1.32	17.80	0.19	2.51	0.05	0.64	0.13	1.80	0.03	0.46	0.00	0.06	0.00	0.02
P3	13.80	0.57	76.59	3.15	0.50	0.02	2.75	0.11	0.36	0.01	1.98	0.08	0.01	0.00	0.07	0.00
P4	0.99	13.38	5.49	74.24	0.04	0.48	0.20	2.66	0.03	0.35	0.14	1.92	0.00	0.01	0.01	0.07
P5	3.01	0.12	0.77	0.03	71.60	2.94	18.25	0.75	0.08	0.00	0.02	0.00	1.85	0.08	0.47	0.02
P6	0.22	2.92	0.06	0.74	5.14	69.41	1.31	17.69	0.01	0.08	0.00	0.02	0.13	1.79	0.03	0.46
P7	0.58	0.02	3.20	0.13	13.72	0.56	76.13	3.13	0.01	0.00	0.08	0.00	0.35	0.01	1.97	0.08
P8	0.04	0.56	0.23	3.11	0.98	13.30	5.46	73.80	0.00	0.01	0.01	0.08	0.03	0.34	0.14	1.91
P9	6.30	0.26	1.60	0.07	0.23	0.01	0.06	0.00	67.59	2.78	17.23	0.71	2.43	0.10	0.62	0.03
P10	0.45	6.10	0.12	1.56	0.02	0.22	0.00	0.06	4.85	65.52	1.24	16.70	0.17	2.35	0.04	0.60
P11	1.21	0.05	6.69	0.28	0.04	0.00	0.24	0.01	12.95	0.53	71.87	2.95	0.46	0.02	2.58	0.11
P12	0.09	1.17	0.48	6.49	0.00	0.04	0.02	0.23	0.93	12.56	5.16	69.67	0.03	0.45	0.18	2.50
P13	0.26	0.01	0.07	0.00	6.26	0.26	1.60	0.07	2.83	0.12	0.72	0.03	67.19	2.76	17.13	0.70
P14	0.02	0.26	0.00	0.07	0.45	6.07	0.11	1.55	0.20	2.74	0.05	0.70	4.82	65.13	1.23	16.60
P15	0.05	0.00	0.28	0.01	1.20	0.05	6.65	0.27	0.54	0.02	3.01	0.12	12.87	0.53	71.44	2.94
P16	0.00	0.05	0.02	0.27	0.09	1.16	0.48	6.45	0.04	0.53	0.22	2.91	0.92	12.48	5.12	69.26

Table 18: Estimated transition probabilities in % between GBM cell phenotypes in normoxia per 3 days assuming independent cell state transitions.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16
P1	89.54	1.33	5.67	0.08	2.85	0.04	0.18	0.18	0.27	0.00	0.02	0.00	0.01	0.00	0.00	0.00
P2	6.86	84.01	0.43	5.32	0.22	0.22	0.01	0.00	0.02	0.25	0.00	0.02	0.00	0.01	0.00	0.00
P3	3.89	0.06	91.32	1.36	0.12	0.00	2.91	0.04	0.01	0.00	0.28	0.00	0.00	0.00	0.01	0.00
P4	0.30	3.65	0.30	85.68	0.01	0.12	0.22	2.73	0.00	0.01	0.02	0.26	0.00	0.00	0.00	0.01
P5	1.04	0.02	0.07	0.00	91.34	1.36	5.78	0.09	0.00	0.00	0.00	0.00	0.28	0.00	0.02	0.00
P6	0.08	0.98	0.01	0.06	7.00	85.70	0.44	5.43	0.00	0.00	0.00	0.00	0.02	0.26	0.00	0.02
P7	0.05	0.00	1.06	0.02	3.97	0.06	93.16	1.39	0.00	0.00	0.00	0.00	0.01	0.00	0.28	0.00
P8	0.00	0.04	0.08	1.00	0.30	3.72	7.14	87.41	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.26
P9	11.14	0.17	0.71	0.01	0.35	0.01	0.02	0.00	78.67	1.17	4.98	0.07	2.51	0.04	0.16	0.00
P10	0.85	10.45	0.05	0.66	0.03	0.33	0.00	0.02	6.03	73.81	0.38	4.67	0.19	2.35	0.01	0.15
P11	0.48	0.01	11.36	0.17	0.02	0.00	0.36	0.01	3.42	0.05	80.23	1.19	0.11	0.00	2.56	0.04
P12	0.04	0.45	0.87	10.66	0.00	0.01	0.03	0.34	0.26	3.20	6.15	75.28	0.01	0.10	0.20	2.40
P13	0.13	0.00	0.01	0.00	11.36	0.17	0.72	0.01	0.92	0.01	0.06	0.00	80.26	1.20	5.08	0.08
P14	0.01	0.12	0.00	0.01	0.87	10.66	0.06	0.67	0.07	0.86	0.00	0.05	6.15	75.30	0.39	4.77
P15	0.01	0.00	0.13	0.00	0.49	0.01	11.59	0.17	0.04	0.00	0.94	0.01	3.48	0.05	81.85	1.22
P16	0.00	0.01	0.01	0.12	0.04	0.46	0.89	10.87	0.00	0.04	0.07	0.88	0.27	3.27	6.27	76.80

Table 19: Estimated transition probabilities in % between GBM cell phenotypes in hypoxia per 3 days assuming independent cell state transitions.

BIBLIOGRAPHY

- [1] M. Abramowitz. *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables*, Dover Publications, Incorporated, 1974.
- [2] M. Amoyel and E. A. Bach. "Cell competition: how to eliminate your neighbours." In: *Development* 141.5 (2014), pp. 988–1000.
- [3] T. Antal, S. Redner, and V. Sood. "Evolutionary dynamics on degree-heterogeneous graphs." In: *Physical Review Letters* 96.18 (2006), p. 188104.
- [4] T. Antal and I. Scheuring. "Fixation of strategies for an evolutionary game in finite populations." In: *Bulletin of Mathematical Biology* 68.8 (2006), pp. 1923–1944.
- [5] S. Asmussen. *Applied probability and queues*. Vol. 51. Springer Science & Business Media, 2008.
- [6] I. Baccelli and A. Trumpp. "The evolving concept of cancer and metastasis stem cells." In: *Journal of Cell Biology* 198.3 (2012), pp. 281–293.
- [7] A.-M. Baker, B. Cereser, S. Melton, et al. "Quantification of crypt and stem cell evolution in the normal and neoplastic human colon." In: *Cell Reports* 8.4 (2014), pp. 940–947.
- [8] S. Bao, Q. Wu, R. E. McLendon, et al. "Glioma stem cells promote radioresistance by preferential activation of the DNA damage response." In: *Nature* 444.7120 (2006), p. 756.
- [9] O. Basak, T. G. Krieger, M. J. Muraro, et al. "Troy+ brain stem cells cycle through quiescence and regulate their number by sensing niche occupancy." In: *Proceedings of the National Academy of Sciences* (2018), p. 201715911.
- [10] N. Beerenwinkel, T. Antal, D. Dingli, et al. "Genetic progression and the waiting time to cancer." In: *PLoS Computational Biology* 3.11 (2007), e225.
- [11] D. Beier, P. Hau, M. Proescholdt, et al. "CD133+ and CD133- glioblastoma-derived cancer stem cells show differential growth characteristics and molecular profiles." In: *Cancer Research* 67.9 (2007), pp. 4010–4015.
- [12] B. Biteau, C. E. Hochmuth, and H. Jasper. "Maintaining tissue homeostasis: dynamic control of somatic stem cell activity." In: *Cell Stem Cell* 9.5 (2011), pp. 402–411.
- [13] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*. Vol. 31. Springer Science & Business Media, 2013.

- [14] T. Buder, A. Deutsch, B. Klink, and A. Voss-Böhme. "Model-based evaluation of spontaneous tumor regression in pilocytic astrocytoma." In: *PLoS Computational Biology* 11.12 (2015), e1004662.
- [15] T. Buder, A. Deutsch, M. Seifert, and A. Voss-Böhme. "CellTrans: An R Package to Quantify Stochastic Cell State Transitions." In: *Bioinformatics and Biology Insights* 11 (June 2017), p. 1177932217712241. DOI: [10.1177/1177932217712241](https://doi.org/10.1177/1177932217712241). URL: <http://insights.sagepub.com/celltrans-an-r-package-to-quantify-stochastic-cell-state-transitions-article-a6410>.
- [16] G. Civenni, A. Walter, N. Kobert, et al. "Human CD271-positive melanoma stem cells associated with metastasis establish tumor heterogeneity and long-term growth." In: *Cancer Research* 71.8 (2011), pp. 3098–3109.
- [17] R. M. Cymerman, Y. Shao, K. Wang, et al. "De novo vs nevus-associated melanomas: differences in associations with prognostic indicators and survival." In: *JNCI: Journal of the National Cancer Institute* 108.10 (2016).
- [18] J. E. Dick. "Stem cell concepts renew cancer research." In: *Blood* 112.13 (2008), pp. 4793–4807.
- [19] A. Dirkse, A. Golebiewska, T. Buder, et al. "Stem cell-associated heterogeneity in Glioblastoma is a result of intrinsic tumor plasticity shaped by the microenvironment." submitted.
- [20] C. Dirven, J. Mooij, and W. Molenaar. "Cerebellar pilocytic astrocytoma: a treatment protocol based upon analysis of 73 cases and a review of the literature." In: *Child's Nervous System* 13.1 (1997), pp. 17–23.
- [21] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow. "Rates of spontaneous mutation." In: *Genetics* 148.4 (1998), pp. 1667–1686.
- [22] D. K. Driman and R. H. Riddell. "Flat adenomas and flat carcinomas: do you see what I see?" In: *Gastrointestinal Endoscopy* 40.1 (1994), pp. 106–109.
- [23] B. J. Due-Tønnessen, E. Helseth, D. Scheie, K. Skullerud, G. Aamodt, and T. Lundar. "Long-term outcome after resection of benign cerebellar astrocytomas in children and young adults (0–19 years): report of 110 consecutive cases." In: *Pediatric Neurosurgery* 37.2 (2002), pp. 71–80.
- [24] R. Durrett. *Probability models for DNA sequence evolution*. Springer Science & Business Media, 2008.
- [25] R. Durrett, J. Foo, and K. Leder. "Spatial Moran models, II: cancer initiation in spatially structured tissue." In: *Journal of Mathematical Biology* 72.5 (2016), pp. 1369–1400.

- [26] R. Durrett and S. Moseley. "Spatial Moran models I. Stochastic tunneling in the neutral case." In: *The Annals of Applied Probability* 25.1 (2015), p. 104.
- [27] R. Durrett and D. Schmidt. "Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution." In: *Genetics* 180.3 (2008), pp. 1501–1509.
- [28] R. Durrett and D. Schmidt. "Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution." In: *Genetics* 180.3 (2008), pp. 1501–1509.
- [29] R. Durrett, D. Schmidt, and J. Schweinsberg. "A waiting time problem arising from the study of multi-stage carcinogenesis." In: *The Annals of Applied Probability* (2009), pp. 676–718.
- [30] R. Durrett, D. Schmidt, and J. Schweinsberg. "A waiting time problem arising from the study of multi-stage carcinogenesis." In: *The Annals of Applied Probability* 19.2 (Apr. 2009), pp. 676–718. DOI: [10.1214/08-AAP559](https://doi.org/10.1214/08-AAP559). URL: <http://dx.doi.org/10.1214/08-AAP559>.
- [31] I. Y. Eyüpoglu, M. Buchfelder, and N. E. Savaskan. "Surgical resection of malignant gliomas—role in optimizing patient outcome." In: *Nature Reviews Neurology* 9.3 (2013), pp. 141–151.
- [32] C. Fernandez, D. Figarella-Branger, N. Girard, et al. "Pilocytic astrocytomas in children: prognostic factors—a retrospective study of 80 cases." In: *Neurosurgery* 53.3 (2003), pp. 544–555.
- [33] J. Foo, K. Leder, and F. Michor. "Stochastic dynamics of cancer initiation." In: *Physical Biology* 8.1 (2011), p. 015002.
- [34] J. Foo, K. Leder, and F. Michor. "Stochastic dynamics of cancer initiation." In: *Physical Biology* 8.1 (2011), p. 015002.
- [35] J. Foo, K. Leder, and M. D. Ryser. "Multifocality and recurrence risk: A quantitative model of field cancerization." In: *Journal of Theoretical Biology* 355.0 (2014), pp. 170–184. ISSN: 0022-5193. DOI: <http://dx.doi.org/10.1016/j.jtbi.2014.02.042>. URL: <http://www.sciencedirect.com/science/article/pii/S0022519314002203>.
- [36] J. Foo, K. Leder, and M. D. Ryser. "Multifocality and recurrence risk: a quantitative model of field cancerization." In: *Journal of Theoretical Biology* 355 (2014), pp. 170–184.
- [37] J. Francois. "Genetics of retinoblastoma." In: *Modern Problems in Ophthalmology* 18 (1977), pp. 165–172.
- [38] M. R. Freeman. "Specification and morphogenesis of astrocytes." In: *Science* 330.6005 (2010), pp. 774–778.
- [39] C. W. Gardiner. *Stochastic methods*. Vol. 4. Springer-Verlag, 1985.

- [40] Y. Geng, S. Chandrasekaran, S. Agastin, J. Li, and M. R. King. "Dynamic switch between two adhesion phenotypes in colorectal cancer cells." In: *Cellular and Molecular Bioengineering* 7.1 (2014), pp. 35–44.
- [41] R. J. Gilbertson and J. N. Rich. "Making a tumour's bed: glioblastoma stem cells and the vascular niche." In: *Nature Reviews Cancer* 7.10 (2007), pp. 733–736.
- [42] C. Ginestier, M. H. Hur, E. Charafe-Jauffret, et al. "ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome." In: *Cell Stem Cell* 1.5 (2007), pp. 555–567.
- [43] J. Gronych, A. Korshunov, J. Bageritz, et al. "An activated mutant BRAF kinase domain is sufficient to induce pilocytic astrocytoma in mice." In: *The Journal of Clinical Investigation* 121.4 (2011), p. 1344.
- [44] R. S. Gunny, R. D. Hayward, K. P. Phipps, B. N. Harding, and D. E. Saunders. "Spontaneous regression of residual low-grade cerebellar pilocytic astrocytomas in children." In: *Pediatric Radiology* 35.11 (2005), pp. 1086–1091.
- [45] P. B. Gupta, C. M. Fillmore, G. Jiang, et al. "Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells." In: *Cell* 146.4 (2011), pp. 633–644.
- [46] M. Al-Hajj, M. S. Wicha, A. Benito-Hernandez, S. J. Morrison, and M. F. Clarke. "Prospective identification of tumorigenic breast cancer cells." In: *Proceedings of the National Academy of Sciences* 100.7 (2003), pp. 3983–3988.
- [47] D. Hanahan and R. A. Weinberg. "The hallmarks of cancer." In: *Cell* 100.1 (2000), pp. 57–70.
- [48] D. Hanahan and R. A. Weinberg. "Hallmarks of cancer: the next generation." In: *Cell* 144.5 (2011), pp. 646–674.
- [49] L. A. Herzenberg, S. C. De Rosa, and L. A. Herzenberg. "Monoclonal antibodies and the FACS: complementary tools for immunobiology and medicine." In: *Immunology Today* 21.8 (2000), pp. 383–390.
- [50] N. J. Higham and L. Lin. "On pth roots of stochastic matrices." In: *Linear Algebra and its Applications* 435.3 (2011), pp. 448–463.
- [51] L. Hindersin and A. Traulsen. "Counterintuitive properties of the fixation time in network-structured populations." In: *Journal of The Royal Society Interface* 11.99 (2014), p. 20140606.
- [52] H. J. Hoffman, D. S. Soloniuk, R. P. Humphreys, et al. "Management and outcome of low-grade astrocytomas of the midline in children: a retrospective review." In: *Neurosurgery* 33.6 (1993), pp. 964–971.

- [53] C. Horbinski, M. N. Nikiforova, J. M. Hagenkord, R. L. Hamilton, and I. F. Pollack. "Interplay among BRAF, p16, p53, and MIB1 in pediatric low-grade gliomas." In: *Neuro-oncology* 14.6 (2012), pp. 777–789.
- [54] E. H. Huang, M. J. Hynes, T. Zhang, et al. "Aldehyde dehydrogenase 1 is a marker for normal and malignant human colonic stem cells (SC) and tracks SC overpopulation during colon tumorigenesis." In: *Cancer Research* 69.8 (2009), pp. 3382–3389.
- [55] A. Humphries and N. A. Wright. "Colonic crypt organization and tumorigenesis." In: *Nature Reviews Cancer* 8.6 (2008), p. 415.
- [56] W. R. Inc. *Mathematica, Version 11.3*. Champaign, IL, 2018.
- [57] M.-d.-M. Inda, R. Bonavia, J. Seoane, et al. "Glioblastoma multiforme: a look inside its heterogeneous nature." In: *Cancers* 6.1 (2014), pp. 226–239.
- [58] Y. Iwasa, F. Michor, N. L. Komarova, and M. A. Nowak. "Population genetics of tumor suppressor genes." In: *Journal of Theoretical Biology* 233.1 (2005), pp. 15–23.
- [59] Y. Iwasa, F. Michor, N. L. Komarova, and M. A. Nowak. "Population genetics of tumor suppressor genes." In: *Journal of Theoretical Biology* 233.1 (2005), pp. 15–23. ISSN: 0022-5193. DOI: [10.1016/j.jtbi.2004.09.001](https://doi.org/10.1016/j.jtbi.2004.09.001). URL: <http://dx.doi.org/10.1016/j.jtbi.2004.09.001>.
- [60] D. R. Johnson, P. D. Brown, E. Galanis, and J. E. Hammack. "Pilocytic astrocytoma survival in adults: analysis of the Surveillance, Epidemiology, and End Results Program of the National Cancer Institute." In: *Journal of Neuro-oncology* 108.1 (2012), pp. 187–193.
- [61] L. A. Johnston. "Competitive interactions between cells: death, growth, and geography." In: *Science* 324.5935 (2009), pp. 1679–1682.
- [62] D. T. Jones, J. Gronych, P. Lichter, O. Witt, and S. M. Pfister. "MAPK pathway activation in pilocytic astrocytoma." In: *Cellular and Molecular Life Sciences* 69.11 (2012), pp. 1799–1811.
- [63] D. T. Jones, B. Hutter, N. Jäger, et al. "Recurrent somatic alterations of FGFR1 and NTRK2 in pilocytic astrocytoma." In: *Nature Genetics* 45 (2013), pp. 927–932.
- [64] H. Kang and D. Shibata. "Direct measurements of human colon crypt stem cell niche genetic fidelity: the role of chance in non-Darwinian mutation selection." In: *Frontiers in Oncology* 3 (2013).
- [65] I. Kareva. "What can ecology teach us about cancer?" In: *Translational Oncology* 4.5 (2011), pp. 266–270.
- [66] A. L. Kennedy, J. P. Morton, I. Manoharan, et al. "Activation of the PIK3CA/AKT pathway suppresses senescence induced by an activated

- RAS oncogene to promote tumorigenesis." In: *Molecular Cell* 42.1 (2011), pp. 36–49.
- [67] K. Kerlikowske. "Epidemiology of ductal carcinoma in situ." In: *Journal of the National Cancer Institute Monographs* 2010.41 (2010), pp. 139–141.
- [68] J. H. Kim, J. W. Kim, Y. N. Kim, et al. "Progression of Monoclonal Gammopathy with Undetermined Significance to Multiple Myeloma Diagnosed by Kidney Biopsy: A Case Report." In: *Case Reports in Nephrology and Dialysis* 5.3 (2015), pp. 180–186.
- [69] A. G. Knudson. "Mutation and cancer: statistical study of retinoblastoma." In: *Proceedings of the National Academy of Sciences* 68.4 (1971), pp. 820–823.
- [70] N. L. Komarova. "Spatial stochastic models for cancer initiation and progression." In: *Bulletin of Mathematical Biology* 68.7 (2006), pp. 1573–1599.
- [71] N. L. Komarova, A. Sengupta, and M. A. Nowak. "Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability." In: *Journal of Theoretical Biology* 223.4 (2003), pp. 433–450.
- [72] N. L. Komarova, A. Sengupta, and M. A. Nowak. "Mutation–selection networks of cancer initiation: tumor suppressor genes and chromosomal instability." In: *Journal of Theoretical Biology* 223.4 (2003), pp. 433–450.
- [73] D. Krackhardt. "Graph theoretical dimensions of informal organizations." In: *Computational Organization Theory* 89.112 (1994), pp. 123–140.
- [74] A. Kreinin and M. Sidelnikova. "Regularization algorithms for transition matrices." In: *Algo Research Quarterly* 4.1/2 (2001), pp. 23–40.
- [75] Y. Kuang, J. D. Nagy, and S. E. Eikenberry. *Introduction to mathematical oncology*. Vol. 59. CRC Press, 2016.
- [76] M. A. LaBarge. "The difficulty of targeting cancer stem cell niches." In: *Clinical Cancer Research* 16.12 (2010), pp. 3121–3129.
- [77] O. Landgren, R. A. Kyle, R. M. Pfeiffer, et al. "Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study." In: *Blood* 113.22 (2009), pp. 5412–5417.
- [78] S. Larjavaara, H. Haapasalo, R. Sankila, P. Helén, and A. Auvinen. "Is the incidence of meningiomas underestimated? A regional survey." In: *British Journal of Cancer* 99.1 (2008), p. 182.
- [79] J. D. Lathia, J. Gallagher, J. M. Heddleston, et al. "Integrin alpha 6 regulates glioblastoma stem cells." In: *Cell Stem Cell* 6.5 (2010), pp. 421–432.
- [80] J. D. Lathia, S. C. Mack, E. E. Mulkearns-Hubert, C. L. Valentim, and J. N. Rich. "Cancer stem cells in glioblastoma." In: *Genes & Development* 29.12 (2015), pp. 1203–1217.

- [81] W. M. Lau, E. Teng, H. S. Chong, et al. "CD44v8-10 is a cancer-specific marker for gastric cancer stem cells." In: *Cancer Research* 74.9 (2014), pp. 2630–2641.
- [82] E. Lieberman, C. Hauert, and M. A. Nowak. "Evolutionary dynamics on graphs." In: *Nature* 433.7023 (2005), p. 312.
- [83] S. R. A. Llaguno and L. F. Parada. "Cell of origin of glioma: biological and clinical implications." In: *British Journal of Cancer* 115.12 (2016), p. 1445.
- [84] H. Lodish, A. Berk, S. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell. "Tumor cells and the onset of cancer." In: *Molecular Cell Biology* (2000).
- [85] J.-K. Loh, A.-S. Lieu, C.-Y. Chai, et al. "Arrested growth and spontaneous tumor regression of partially resected low-grade cerebellar astrocytomas in children." In: *Child's Nervous System* 29.11 (2013), pp. 2051–2055.
- [86] H. Lopez-Bertoni, Y. Li, and J. Laterra. "Cancer Stem Cells: Dynamic Entities in an Ever-Evolving Paradigm." In: *Biology and Medicine* 2015 (2015).
- [87] H. Lopez-Bertoni, Y. Li, and J. Laterra. "Cancer stem cells: dynamic entities in an ever-evolving paradigm." In: *Biology and Medicine (Aligarh)* 7.Suppl 2 (2015).
- [88] D. N. Louis, H. Ohgaki, O. D. Wiestler, et al. "The 2007 WHO classification of tumours of the central nervous system." In: *Acta Neuropathologica* 114.2 (2007), pp. 97–109.
- [89] S. A. Mani, W. Guo, M.-J. Liao, et al. "The epithelial-mesenchymal transition generates cells with properties of stem cells." In: *Cell* 133.4 (2008), pp. 704–715.
- [90] E. A. Martens, R. Kostadinov, C. C. Maley, and O. Hallatschek. "Spatial structure increases the waiting time for cancer." In: *New Journal of Physics* 13.11 (2011), p. 115014.
- [91] A. Mazloom, J. C. Hodges, B. S. Teh, M. Chintagumpala, and A. C. Paulino. "Outcome of patients with pilocytic astrocytoma and leptomeningeal dissemination." In: *International Journal of Radiation Oncology* Biology* Physics* 84.2 (2012), pp. 350–354.
- [92] S. McGuire. "World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO Press, 2015." In: *Advances in Nutrition: An International Review Journal* 7.2 (2016), pp. 418–419.

- [93] L. M. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley. "Cancer as an evolutionary and ecological process." In: *Nature Reviews Cancer* 6.12 (2006), pp. 924–935.
- [94] F. Michor, S. A. Frank, R. M. May, Y. Iwasa, and M. A. Nowak. "Somatic selection for and against cancer." In: *Journal of Theoretical Biology* 225.3 (2003), pp. 377–382. ISSN: 0022-5193. DOI: [http://dx.doi.org/10.1016/S0022-5193\(03\)00267-4](http://dx.doi.org/10.1016/S0022-5193(03)00267-4). URL: <http://www.sciencedirect.com/science/article/pii/S0022519303002674>.
- [95] F. Michor, Y. Iwasa, and M. A. Nowak. "Dynamics of cancer progression." In: *Nature Reviews Cancer* 4.3 (2004), pp. 197–205.
- [96] F. Michor, Y. Iwasa, H. Rajagopalan, C. Lengauer, and M. A. Nowak. "Linear Model of Colon Cancer Initiation." In: *Cell Cycle* 3 (2004), pp. 356–360. ISSN: 1469-8064.
- [97] A. V. Molofsky, R. Krenick, E. Ullian, et al. "Astrocytes and disease: a neurodevelopmental perspective." In: *Genes & Development* 26.9 (2012), pp. 891–907.
- [98] P. A. P. Moran. "Random processes in genetics." In: *Mathematical Proceedings of the Cambridge Philosophical Society* 54 (01 Jan. 1958), pp. 60–71. ISSN: 1469-8064. DOI: [10.1017/S0305004100033193](http://journals.cambridge.org/article_S0305004100033193). URL: http://journals.cambridge.org/article_S0305004100033193.
- [99] J. R. Norris. *Markov chains*. 2. Cambridge university press, 1998.
- [100] M. A. Nowak. *Evolutionary dynamics*. Exploring the equations of life. Cambridge, MA: The Belknap Press of Harvard University Press, 2006.
- [101] M. A. Nowak, N. L. Komarova, A. Sengupta, et al. "The role of chromosomal instability in tumor initiation." In: *Proceedings of the National Academy of Sciences* 99.25 (2002), pp. 16226–16231.
- [102] M. A. Nowak, F. Michor, N. L. Komarova, and Y. Iwasa. "Evolutionary dynamics of tumor suppressor gene inactivation." In: *Proceedings of the National Academy of Sciences* 101.29 (2004), pp. 10635–10638.
- [103] P. C. Nowell. "Mechanisms of tumor progression." In: *Cancer Research* 46.5 (1986), pp. 2203–2207.
- [104] C. A. O'Brien, A. Pollett, S. Gallinger, and J. E. Dick. "A human colon cancer cell capable of initiating tumour growth in immunodeficient mice." In: *Nature* 445.7123 (2007), p. 106.
- [105] H. Ogiwara, R. M. Bowman, and T. Tomita. "Long-term follow-up of pediatric benign cerebellar astrocytomas." In: *Neurosurgery* 70.1 (2011), pp. 40–48.

- [106] H. Ohgaki and P. Kleihues. "The definition of primary and secondary glioblastoma." In: *Clinical Cancer Research* 19.4 (2013), pp. 764–772.
- [107] L. Palma, P. Celli, and A. Mariottini. "Long-term follow-up of childhood cerebellar astrocytomas after incomplete resection with particular reference to arrested growth or spontaneous tumour regression." In: *Acta Neurochirurgica* 146.6 (2004), pp. 581–588.
- [108] A. P. Patel, I. Tirosh, J. J. Trombetta, et al. "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma." In: *Science* 344.6190 (2014), pp. 1396–1401.
- [109] A. I. Persson, C. Petritsch, F. J. Swartling, et al. "Non-stem cell origin for oligodendroglioma." In: *Cancer Cell* 18.6 (2010), pp. 669–682.
- [110] I. Petrovchich and J. M. Ford. "Genetic predisposition to gastric cancer." In: *Seminars in Oncology*. Vol. 43. 5. Elsevier. 2016, pp. 554–559.
- [111] P. F. Pinsky. "A multi-stage model of adenoma development." In: *Journal of Theoretical Biology* 207.2 (2000), pp. 129–143.
- [112] N. Privault. *Understanding Markov chains: examples and applications*. Springer Science & Business Media, 2013.
- [113] H. Qin, D. Bao, X. Tong, Q. Hu, G. Sun, and X. Huang. "The role of stem cells in benign tumors." In: *Tumor Biology* 37.12 (2016), pp. 15349–15357.
- [114] E. H. Raabe, K. S. Lim, J. M. Kim, et al. "BRAF activation induces transformation and then senescence in human neural stem cells: a pilocytic astrocytoma model." In: *Clinical Cancer Research* 17.11 (2011), pp. 3590–3599.
- [115] J. Reid, E. Lee, A. Paladin, C. Carrico, and W. Davros. *Pediatric Radiology*. New York: Oxford University Press, 2013. ISBN: 978-0-199-75532-5.
- [116] E. F. Rodriguez, B. W. Scheithauer, C. Giannini, et al. "PI3K / AKT pathway alterations are associated with clinically aggressive and histologically anaplastic subsets of pilocytic astrocytoma." In: *Acta Neuropathologica* 121.3 (2011), pp. 407–420.
- [117] M. D. Ryser, W. T. Lee, N. E. Ready, K. Z. Leder, and J. Foo. "Quantifying the dynamics of field cancerization in tobacco-related head and neck cancer: a multiscale modeling approach." In: *Cancer Research* 76.24 (2016), pp. 7078–7088.
- [118] A. R. Safa, M. R. Saadatzadeh, A. A. Cohen-Gadol, K. E. Pollok, and K. Bijangi-Vishehsaraei. "Glioblastoma stem cells (GSCs) epigenetic plasticity and interconversion between differentiated non-GSCs and GSCs." In: *Genes & Diseases* 2.2 (2015), pp. 152–163.

- [119] N. Sanai, M.-Y. Polley, M. W. McDermott, A. T. Parsa, and M. S. Berger. "An extent of resection threshold for newly diagnosed glioblastomas." In: *Journal of Neurosurgery* 115.1 (2011), pp. 3–8.
- [120] J. D. Schiffman, J. G. Hodgson, S. R. VandenBerg, et al. "Oncogenic BRAF mutation with CDKN2A inactivation is characteristic of a subset of pediatric malignant astrocytomas." In: *Cancer Research* 70.2 (2010), pp. 512–519.
- [121] J. Schweinsberg et al. "Waiting for m mutations." In: *Electronic Journal of Probability* 13 (2008), pp. 1442–1478.
- [122] E. M. Shriver, J. Ragheb, and T. T. David. "Combined transcranial-orbital approach for resection of optic nerve gliomas: a clinical and anatomical study." In: *Ophthalmic Plastic & Reconstructive Surgery* 28.3 (2012), pp. 184–191.
- [123] A. J. Sievert and M. J. Fisher. "Pediatric low-grade gliomas." In: *Journal of Child Neurology* 24.11 (2009), pp. 1397–1408.
- [124] J. S. Smith, E. F. Chang, K. R. Lamborn, et al. "Role of extent of resection in the long-term outcome of low-grade hemispheric gliomas." In: *Journal of Clinical Oncology* 26.8 (2008), pp. 1338–1345.
- [125] D. W. Smoots, J. R. Geyer, D. M. Lieberman, and M. S. Berger. "Predicting disease progression in childhood cerebellar astrocytoma." In: *Child's Nervous System* 14.11 (1998), pp. 636–648.
- [126] V. Sood, T. Antal, and S. Redner. "Voter models on heterogeneous networks." In: *Physical Review E* 77.4 (2008), p. 041121.
- [127] P. S. Steeg. "Targeting metastasis." In: *Nature Reviews Cancer* 16.4 (2016), pp. 201–218.
- [128] P. Steinbok, K. Poskitt, and G. Hendson. "Spontaneous regression of cerebellar astrocytoma after subtotal resection." In: *Child's Nervous System* 22.6 (2006), pp. 572–576.
- [129] C. Stürer, B. Vilz, M. Majores, A. Becker, J. Schramm, and M. Simon. "Frequent recurrence and progression in pilocytic astrocytoma in adults." In: *Cancer* 110.12 (2007), pp. 2799–2808.
- [130] D. Temko, Y.-K. Cheng, K. Polyak, and F. Michor. "Mathematical modeling links pregnancy-associated changes and breast cancer risk." In: *Cancer Research* 77.11 (2017), pp. 2800–2809.
- [131] C. J. Thalhauser, J. S. Lowengrub, D. Stupack, and N. L. Komarova. "Selection in spatial stochastic models of cancer: migration as a key modulator of fitness." In: *Biology Direct* 5.1 (2010), p. 21.

- [132] C. Tomasetti, B. Vogelstein, and G. Parmigiani. "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation." In: *Proceedings of the National Academy of Sciences* 110.6 (2013), pp. 1999–2004.
- [133] A. Traulsen, T. Lenaerts, J. M. Pacheco, and D. Dingli. "On the dynamics of neutral mutations in a mathematical model for a homogeneous stem cell population." In: *Journal of The Royal Society Interface* 10.79 (2013), p. 20120810.
- [134] A. L. Vescovi, R. Galli, and B. A. Reynolds. "Brain tumour stem cells." In: *Nature Reviews Cancer* 6.6 (2006), pp. 425–436.
- [135] A. Vijay, A. Elaffandi, and H. Khalaf. "Hepatocellular adenoma: An update." In: *World Journal of Hepatology* 7.25 (2015), p. 2603.
- [136] A. Vishwakarma and J. M. Karp. *Biology and Engineering of Stem Cell Niches*. Academic Press, 2017.
- [137] J. E. Visvader. "Cells of origin in cancer." In: *Nature* 469.7330 (2011), pp. 314–322.
- [138] B. Vogelstein and K. W. Kinzler. "The multistep nature of cancer." In: *Trends in Genetics* 9.4 (1993), pp. 138–141.
- [139] B. Vogelstein and K. W. Kinzler. *The genetic basis of human cancer*. McGraw-Hill, 2002.
- [140] B. Vogelstein, K. W. Kinzler, et al. "Cancer genes and the pathways they control." In: *Nature Medicine* 10.8 (2004), pp. 789–799.
- [141] H. Wang, M. Naghavi, C. Allen, et al. "Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015." In: *The Lancet* 388.10053 (2016), pp. 1459–1544.
- [142] J. C. Wang and J. E. Dick. "Cancer stem cells: lessons from leukemia." In: *Trends in Cell Biology* 15.9 (2005), pp. 494–501.
- [143] W. Wang, Y. Quan, Q. Fu, et al. "Dynamics between cancer cell subpopulations reveals a model coordinating with both hierarchical and stochastic concepts." In: *PloS One* 9.1 (2014), e84654.
- [144] M. J. Williams, B. Werner, C. P. Barnes, T. A. Graham, and A. Sottoriva. "Identification of neutral tumor evolution across cancer types." In: *Nature Genetics* 48.3 (2016), pp. 238–244.
- [145] D. Wodarz and N. L. Komarova. *Dynamics of cancer: mathematical foundations of oncology*. World Scientific, 2014.

- [146] C.-I. Wu, H.-Y. Wang, S. Ling, and X. Lu. "The ecology and evolution of cancer: the ultra-microevolutionary process." In: *Annual Review of Genetics* 50 (2016), pp. 347–369.
- [147] G. Yang, Y. Quan, W. Wang, et al. "Dynamic equilibrium between cancer stem cells and non-stem cancer cells in human SW620 and MCF-7 cancer cell populations." In: *British Journal of Cancer* 106.9 (2012), pp. 1512–1519.
- [148] J. Yang, D. E. Axelrod, and N. L. Komarova. "Determining the control networks regulating stem cell lineages in colonic crypts." In: *Journal of Theoretical Biology* 429 (2017), pp. 190–203.
- [149] Z. F. Yang, D. W. Ho, M. N. Ng, et al. "Significance of CD90+ cancer stem cells in human liver cancer." In: *Cancer Cell* 13.2 (2008), pp. 153–166.
- [150] A. Zeuner, M. Todaro, G. Stassi, and R. De Maria. "Colorectal cancer stem cells: from the crypt to the clinic." In: *Cell Stem Cell* 15.6 (2014), pp. 692–705.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Diese Dissertation wurde an der Technischen Universität Dresden, Fachrichtung Mathematik am Institut für Mathematische Stochastik unter der wissenschaftlichen Betreuung von Prof. Dr. Dietmar Ferger sowie an der HTW Dresden, Fakultät Informatik/Mathematik unter der wissenschaftlichen Betreuung von Prof. Dr. Anja Voß-Böhme angefertigt.

Hiermit versichere ich, dass ich keine früheren erfolglosen Promotionsverfahren bestritten habe. Hiermit erkenne ich die Promotionsordnung der Fakultät Mathematik und Naturwissenschaften an der Technischen Universität Dresden in der Fassung vom 23. Februar 2011 an.

Dresden, 04.04.2018

