HELMHOLTZ
ASSOCIATION
Alliance on Systems Biology
HelmholtzZentrum münchen
German Research Center for Environmental Health

TECHNISCHE
UNIVERSITÄT
MÜNCHEN

From single cells to genealogies: Stochastic models of stem cell differentiation

Michael Strasser

Dezember 2014

TECHNISCHE UNIVERSITÄT MÜNCHEN

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl M12 (Mathematische Modellierung biologischer Systeme)

From single cells to genealogies: Stochastic models of stem cell differentiation

Michael Korbinian Strasser

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

_____

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. F. J. Theis
2. _____
3. _____

Die Dissertation wurde am _____ bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am _____ angenommen.

"The true logic of this world is the calculus of probability."

JAMES CLERK MAXWELL

Acknowledgements

Upfront, I would like to thank quite a few people who in some way or another helped me throughout this scientific adventure which is summarized on the next 150 pages.

Thanks to Fabian and Carsten, who gave me the chance to do this project at ICB/QSCD, who guided me, kept me motivated over the years and never got tired discussing with me. Additionally, thanks for signing my paychecks!

Thanks to the members of the Schroeder lab for introducing me to the exciting world of stem cell research and for sharing their knowledge with me in countless lab-meetings.

Thanks to my former roomies Justin and Katrin for the great time I had over the years with you guys, for the entertaining discussions about potential landscapes and other weird stuff, and for helping me decipher the strange writings on my whiteboard!

Thanks to Felix and Jörg for our regular meetings at the coffee machine, arguing about science and even more important things. These breaks kept me from going froot loops.

Thanks to the our table-tennis (former table-soccer) gang, for the entertaining competitions and for luring me away from work.

Thanks to each and every one of you guys at ICB for putting a lot of fun into my everyday work.

Finally, thanks to my parents, not only for the unconditional support over the years, but also for sparking my interest in science (and biology in particular) in the first place. Last but not least at all, thanks to Flori, the architect of my past, present and future.

Abstract

During differentiation, a stem cell and its progeny cascade through a number of lineage decisions from a multipotent state over progenitor states to mature functional cells. Within the current paradigm of cell fate choice, many decisions are assumed to be binary and realized by a genetic toggle switch, a simple cellular memory device consisting of two genes that inhibit each others expression. In each differentiating cell, one gene will eventually win this biomolecular battle, inhibiting the other gene and subsequently activating its lineage-determining downstream targets. In hematopoiesis, the generation of blood cells, a series of gene switches has been found along the differentiation path of hematopoietic stem cells, potentially directing the ratio of mature blood cells. The most prominent example in this context is the mutual inhibition of Gata1 and PU.1, two transcription factors responsible for the development of erythroid and myeloid blood cells from common myeloid progenitors.

While being an intriguing and simple mechanism of cell fate choice, no definite experimental proof of the toggle switch paradigm exists: It is still unclear whether a toggle switch actively determines the cell fate choice through its dynamics or merely locks down a previously chosen fate. With the advent of new single cell technology, which allows to monitor cell fate decisions in single cells continuously over time, these questions are now being addressed. However, when observing populations of dividing cells on a single cell level, one is confronted with two challenges: cellular heterogeneity due to stochastic fluctuations and inherent genealogical structure of the data due to cell division. In this thesis, these two challenges will be addressed using stochastic, single cell models of stem cell differentiation.

We start with a theoretical investigation of the toggle switch motif as a cell-intrinsic mechanism of cell fate choice in the presence of stochastic fluctuations. Specifically, we show how the dynamics of the system are altered compared to previous studies when accounting for small mRNA numbers in the gene expression process. We find that the switching process can be regarded as a point process with a fixed rate and provide analytical expressions for the switching rates of the system. Using Approximate Bayesian Computation we show that a stochastic toggle switch model is capable of explaining the differentiation dynamics in the granulocyte/monocyte cell fate decision.

Next, we present a method based on generalized linear models to infer potential cell-extrinsic features (e.g. local cell density) causing differentiation events observed in cellular genealogies. We analyze the required sample sizes and the influence of cell tracking error on the results. Furthermore, we utilize the genealogical information to validate our model, i.e. to test whether the model is able to explain the correlation structure observed in sister cells.

Finally, we combine the ideas of cell-intrinsic and cell-extrinsic processes impacting on cell fate choice into a single model to explain correlated cell fate marker onsets in genealogies. Motivated by our findings in the stochastic toggle switch, we assume that differentiation is a point process potentially modulated by external factors while the onset of the cell fate marker in response to differentiation is delayed due to an intrinsic stochastic gene expression process. We develop an inference method tailored to this model which allows us to predict the timepoint of differentiation from the observed correlations of marker onsets in the genealogies. After testing the method on various synthetic datasets, it is applied to the myeloid/erythroid fate choice in order to investigate the role of the

PU.1/Gata1 toggle switch. Utilizing available timecourse information on PU.1 expression levels, we find that PU.1 dynamics at the predicted timepoints of differentiation deviate significantly from the standard PU.1/Gata1 toggle switch model.

Summarizing, in this thesis we develop methods and models to analyze differentiation in cellular genealogies and give insight into cell fate choice in hematopoiesis.

Zusammenfassung

Während der Differenzierung durchlaufen eine Zelle und ihre Nachkommen eine Hierarchie aus Zellzuständen. An der Spitze dieser Hierarchie steht das Stammzellstadium, gefolgt von diversen Vorläufterzelltypen auf den darauf folgenden Ebenen bis hin zu final ausdifferenzierten, funktionellen Zelltypen am unteren Ende der Hierarchie. Mit jedem Schritt muss die Zelle sich für einen der typischerweise zwei möglichen nachfolgenden Zelltypen entscheiden. Derzeit nimmt man an, dass diese binären Differenzierungsentscheidungen auf molekularer Ebene durch sogenannte genetische Schalter realisiert sind. Unter einem genetischen Schalter versteht man ein Paar von Transkriptionsfaktoren, die gegenseitig ihre Expression inhibieren. Weiterhin nimmt man an, dass in der sich entscheidenden Zelle einer der beiden Transkriptionsfaktoren dieses molekulare Kräftemessen für sich entscheidet, damit die für seinen Zelltyp charakteristischen Gene aktiviert und so die Zelle in einen der beiden möglichen Zustände treibt.

Das meist untersuchte Beispiel für solch einen genetischen Schalter stammt aus der Hämatopoese: Hier glaubt man, dass ein genetischer Schalter aus den beiden Transkriptionsfaktoren PU.1 und Gata1 die Entscheidung zwischen der myeloiden und der erythroiden Linie der Blutzellen bestimmt. Diese Vermutung konnte bislang allerdings nicht experimentell bestätigt werden: Es ist unbekannt ob die Dynamik dieser beiden Faktoren aktiv die Entscheidung beeinflusst, oder ob der genetische Schalter aus PU.1 und Gata1 lediglich als Konsequenz einer Entscheidung, die an anderer Stelle getroffen wurde, umgelegt wird.

Durch die Entwicklung neuer Technologien, im Besonderen der “time-lapse” Mikroskopie, ist es möglich Zellentscheidungen auf Einzelzellebene kontinuierlich über die Zeit zu beobachten und somit dieses Fragen zu addressieren. Aus diesen Einzelzelldaten ergeben sich jedoch neben zahlreichen technischen auch neue theoretische Herausforderungen: Zum einen wird auf Einzelzellebene die Heterogenität der individuellen Zellen sichtbar, die sich z.B. aus den inhärent stochastischen Genexpressionsprozessen innerhalb der Zellen ergibt. Zum anderen erhält man aufgrund der Zellteilungen Daten, denen eine Baumstruktur zugrunde liegt, sogenannte zelluläre Genealogien. Im Verlauf dieser Arbeit werden wir diese beiden Aspekte untersuchen.

Wir beginnen mit einer theoretischen Analyse der stochastischen Dynamik eines genetischen Schalters. Im Unterschied zu vorherigen Arbeiten untersuchen wir den Einfluss von kleinen mRNA Zahlen und finden, dass die Dynamik des Systems sich dadurch grundlegend verändert. Weiterhin untersuchen wir Zustandsübergänge im System und kommen zu dem Ergebnis, dass das Schalten des Systems zwischen zwei Zuständen durch einen Punktprozess angenähert werden kann, dessen Rate wir also Funktion der Systemparameter herleiten. Unter Anwendung von Approximate Bayesian Computation wird gezeigt, dass die beobachtete Dynamik der Differenzierungsentscheidung zwischen Granulocyten und Monocyten durch ein stochastisches Modell eines genetischen Schalters erklärt werden kann.

Im Anschluss entwickeln wir eine Methode basierend auf generalisierten linearen Modellen, um externe Einflussgrößen, wie z.B. lokale Zelldichte, auf Zelldifferenzierung zu identifizieren. Die erforderliche Stichprobengröße sowie der zulässige Fehler im Tracking der Zellen wird analysiert. Weiterhin zeigen wir, wie die Baumstruktur der Daten benutzt werden kann um das gelernte Modell zu validieren.

Im letzten Teil dieser Arbeit werden die zuvor entwickelten zell-intrinsischen und zell-extrinsischen Differenzierungsmodelle zu einem einzigen Modell kombiniert. Hier nehmen wir an, dass die Zelldifferenzierung durch einen Punktprozess dargestellt werden kann, jedoch die Beobachtung dieser Differenzierung verzögert erfolgt, wodurch sich Korrelationsstrukturen in den zellulären Genealogien ergeben. Wir entwickeln eine Inferenzmethode, welche die Parameter des Modells anhand der beobachteten Genealogien schätzt, und benutzen dieses Modell um Differenzierungszeitpunkte in Genealogien vorherzusagen. Angewandt auf zelluläre Genealogien der Hämatopoese finden wir signifikante Unterschiede in der Dynamik von PU.1 an den vom Modell vorhergesagten Zeitpunkt der Differenzierung im Vergleich zu den Erwartungen unter Annahme eines genetischen Schalters. Somit können wir eine aktive Rolle von PU.1 in der Differenzierungsentscheidung zwischen myeloiden und erytroiden Zelltypen ausschließen.

Publications

- **M. Strasser**, F. J. Theis, and C. Marr. Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophysical journal*, 2012.

- C. Marr*, **M. Strasser*** , M. Schwarzfischer, T. Schroeder, and F. J. Theis. Multiscale modeling of GMP differentiation based on single-cell genealogies. *The FEBS Journal*, 2012.

- P. S. Hoppe, M. Schwarzfischer, D. Loeffler, K. D. Kokkaliaris, O. Hilsenbeck, N. Moritz, M. Endele, A. Filipczyk, M. A. Rieger, C. Marr, **M. Strasser**, B. Schauberger, I. Burtscher, O. Ermakova, A. Bürger, H. Lickert, C. Nerlov, F. J. Theis and T. Schroeder. Random PU.1 / Gata1 protein ratios do not induce early myeloid lineage choice. Under review at *Nature*.

- M. Schwarzfischer*, O. Hilsenbeck* , B. Schauberger, S. Hug, A. Filipczyk, P. S. Hoppe, **M. Strasser**, F. Buggenthin, J. S. Feigelman, J. Krumsiek, D. Loeffler, K. D. Kokkaliaris, A. J. J. van den Berg, M. Endele, S. Hastreiter, C. Marr, F. J. Theis, and T. Schroeder. Single-cell quantification of cellular and molecular behavior in long-term time-lapse microscopy. Under review at *Nature Biotechnology*.

- T. Blasi, F. Buettner, **M. Strasser**, S. Linnarsson, C. Marr, and F. J. Theis. CGcorrect: A method to correct for confounding cell-cell variation due to cell growth in single-cell transcriptomics. Under review at *Bioinformatics*.

* These authors contributed equally.

Contents

Chapter 1

Introduction

Mathematical models are an integral part of science. It represents an abstraction and approximation of a real world phenomenon and allows formal treatment thereof. The model helps to explain and understand observations but can also be used to predict yet unknown situations.

Mathematical models are particularly prominent in physics. Newton's equations of classical mechanics not only explain equally well the deterministic motion of baseballs, planets and galaxies, but also allow to make predictions: For example, discrepancies between predicted and observed trajectories of Uranus led to the discovery of Neptune. While it was long thought that Newton's equations are universal, at the beginning of the 20th century two limitations were discovered. Newton's equations must be augmented by the laws of relativity when velocities approach the speed of light, and fail entirely when looking at atomic levels, where the character of physics changes fundamentally due to the quantum nature of matter: While in classical mechanics the world is deterministic and perfectly predictable, on small scales, the world is inherently probabilistic. Predictions of experimental outcomes can only be phrased in terms of probability distributions. On this scale, classical mechanics is then replaced by the theory of quantum mechanics, which does account for the inherent randomness on atomic and subatomic scales. However, it emerges from quantum mechanics as the probabilistic behavior of individual atoms is averaged out when considering large numbers of atoms. While quantum mechanical models, such as the standard model of particle physics are highly abstract, their predictions are nonetheless powerful (e.g. the prediction of the Higgs boson) and the insight these models gave is now used to exploit the quantum nature of matter in every day life, e.g. in lasers or transistors.

Mathematical models are not constrained to describing inanimate matter only, but are equally powerful when used to investigate living systems. In a seminal study, Luria and Delbrück (1943) investigated the resistance of bacteria to virus infections and from their mathematical model and experimental data could conclude that resistance in a bacterial population arises due to random mutations instead of acquired immunity when exposed to the virus. The prevalence of mathematical models in biology has increased tremendously with the advent of Systems Biology during the last decade with the goal to gain quantitative insight biological systems. Here, an interesting parallel to physics is found: Classical mathematical models in biology, e.g. the Lotka-Voltera model of predator and prey populations (Lotka, 1925), are deterministic as they describe large numbers of entities, e.g.

cells or molecules. However it has been realized recently, that when looking at biological systems on smaller scales, e.g. on single cells rather the populations, stochastic fluctuations become significant and one has to treat these systems probabilistically. In analogy to classical mechanics, the "classical" deterministic system is recovered when looking at large numbers of entities. However, the source of randomness is not due to the quantum nature of matter itself, but is rather a manifestation of the large number of degrees of freedom (in the same way the movement of a heavy particle in a gas seems random, but is the result of deterministic collisions) as well as the complex interactions within a cell.

While decades of research in physics have taught us how to cope and even exploit the probabilistic nature of matter, e.g. resulting in the dawn of quantum computing, in biology we have only started to understand the implications of stochasticity and resulting cellular heterogeneity on organisms, e.g. in the context of development and cell differentiation.

1.1 Stem cell biology

Stem cells form the backbone of development, tissue homeostasis and regeneration in higher organisms ranging from primitive flatworms (Wagner et al., 2011) and Hydra (Boehm et al., 2012) to mice (Becker et al., 1963) and humans (Thomson et al., 1998). A whole organism develops from a single stem cell, the zygote, giving rise to over 200 different cell types in the human body (Shevde, 2012). Homeostasis in the blood system is maintained by hematopoietic stem cells which are able to sustain the rapid turnover of 10^{12} mature blood cells per day in an adult human (Kaushansky et al., 2010). Stem cells in Axolotls, a Mexican salamander species, are capable of regenerating entire limbs or parts of the spinal cord after injury (Roy and Lévesque, 2006; Sandoval-Guzmán et al., 2014). These amazing capabilities emerge from the two defining properties of stem cells: 1) Self-renewal, which is the ability to maintain their stem cell identity across infinitely many cell division, and 2) pluripotency, the ability to give rise to multiple differentiated cell types.

One distinguishes between embryonic stem cells and adult, somatic stem cells. Embryonic stem cells are derived from the inner cell mass of a developing embryo and are capable of differentiating into all three germ layers, endoderm, mesoderm and ectoderm (Evans and Kaufman, 1981; Martin, 1981). Their pluripotency and the possibility to keep them in culture indefinitely have made them an indispensable model system for stem cell and developmental biology. However, the use of embryonic stem cells is controversial because they are derived by sacrificing a living embryo. During normal development, the once pluripotent cells of the inner cell mass start to loose this property quickly thereafter. With gastrulation, each cell has assumed one of three germlayer identities and their progeny will be restricted to this germlayer. With further development most cells in the organism will at some point differentiate terminally into a mature cell type. Only a few cells retain their stem cell property in the adult organism. These cells are called adult or somatic stem cells and are responsible for tissue repair and homeostasis in the adult organism. Opposed to embryonic stem cells, their potential is limited, i.e. they can give rise only to certain cell types, a property called multipotency. A specific type of adult stem cell has been identified for many tissue types, e.g. the brain (Temple, 1989), the

blood system (Becker et al., 1963), skin (Alonso and Fuchs, 2003), and the small intestine (Barker et al., 2007).

Whereas embryonic and somatic stem cells have been known since the 1960s, recently, a third class of stem cells emerged: In 2006, Takahashi and Yamanaka discovered that mouse fibroblasts can be reprogrammed by ectopic expression of four genes (Oct3/4, Sox2, c-Myc, and Klf4) to an embryonic stem cell like state, which they termed induced pluripotent stem (iPS) cells. Similar results were obtained in humans (Takahashi et al., 2007) and recently the efficiency of reprogramming could be increased tremendously (Rais et al., 2013), making them an attractive alternative to embryonic stem cells. While iPS cells are less controversial than embryonic stem cells, their increased potential to form tumors due to the forced expression of oncogenes (e.g. Klf4) and recent evidence of transcriptional abnormalities (Ma et al., 2014) have precluded iPS cells from fully replacing embryonic stem cells as a stem cell source.

All three stem cell types hold great promise for regenerative medicine due to their self-renewal and lineage potential properties (Daley, 2012a; Shevde, 2012), especially for diseases where no conventional drug treatment is available (Daley, 2012b). Bone marrow transplantation, first performed by Thomas et al. in 1957 is oldest and currently most successful form of stem cell therapy and can be used to treat e.g. leukemia or anemia. Most recently, new stem cell therapies were tested in experimental studies: Human embryonic stem cells were used to treat patients with retinal blindness (Schwartz et al., 2012) and embryonic stem cells were suggested as a cell source to replace degenerate neurons in the striatum of patients with Huntington's disease (Nicoleau et al., 2011), for which no effective drug treatment exists. Stem cells are also readily used in tissue engineering, e.g. to form tracheae (Macchiarini et al., 2008), or to create kidney structures (Taguchi et al., 2014), which eventually could provide a source for organ transplantation without the complications of transplant rejection or graft-versus-host-disease. In another application, patient derived stem cells are used as a realistic disease model that can be used for a more efficient drug screening than the classical target-centric drug discovery (for a review, see Grskovic et al., 2011).

Even though tremendous advancement and breakthroughs in stem cell biology have been achieved in the last decades, our understanding of the mechanisms underlying self-renewal, pluripotency, cell fate choice, differentiation and regeneration remains surprisingly small and many questions remain unanswered: Why is the forced expression of only four out of 20000 genes (the "Yamanaka genes") enough to dedifferentiate fibroblasts into pluripotent cells? How does a stem cell decide what cell type it will differentiate into? Why do success rates for HSC transplantations remain low (50% survival after 5 years, see Jenq and van den Brink, 2010; Passweg et al., 2012) that these are only administered to the sickest patients, who lack other alternatives, even though hematopoiesis is the most well studied stem cell system and HSC transplantations have been performed for more than 50 years?

Here, a deeper understanding of the underlying mechanisms of cell fate choice is needed (Roeder and Radtke, 2009). This will not only help to reduce the risk of clinical applications of stem cells (e.g. the formation of stem cell derived tumors, Amariglio et al., 2009), but also increase our understanding about the origins of disease (Huang, 2013).

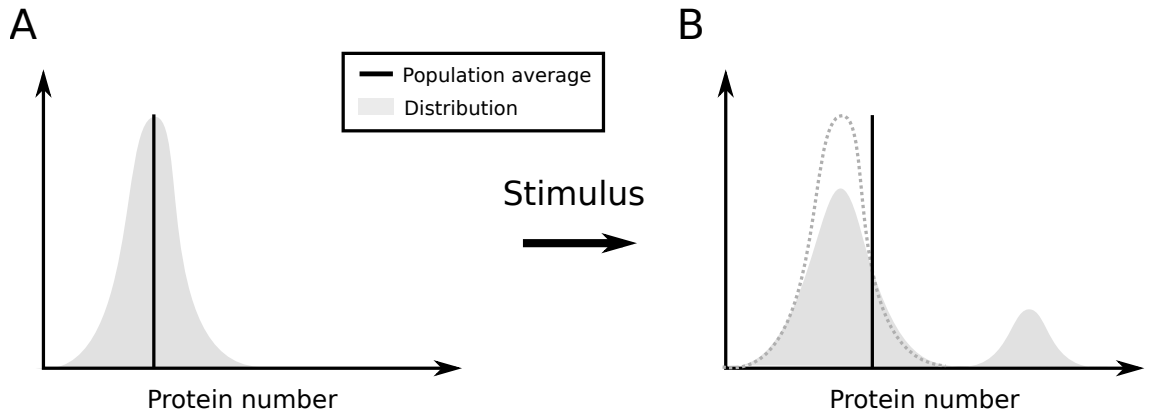


Figure 1.1: **Population averages can obscure heterogeneous cellular response.** A) A population of cells homogeneously expresses a protein of interest (mean and distribution are shown). B) Upon stimulus, a small fraction of cells responses by strongly upregulating the protein, while the other cells remain unchanged in their protein expression. The population average (black line) changes only slightly. The pre-stimulus distribution is indicated as dashed line.

1.2 Single cell experiments and cellular heterogeneity

A first step towards gaining more insight into the various phenomena of stem cell biology is to focus on the analysis of single cells instead of population measurements.

Classical tools of molecular biology need to be fueled by huge amounts of cellular material: To detect a specific proteins in a cell by western blotting, cells are lysed, and the lysate is separated via gel-electrophoresis. To detect proteins of interest, the proteins have to be transferred from the gel onto a membrane where they can then be detected by antibodies. Due to its limited sensitivity[1], typically 10^5 cells are required to perform western blotting (Ciaccio et al., 2010). DNA-microarrays are used to quantify the mRNAs contained in the cell by reverse transcribing mRNA into cDNA and amplifying the cDNA via PCR. Finally, amplified material is put on oligonucleotide-chips where it hybridizes with complementary oligonucleotides and is detected via fluorescence probes. To keep the amplification step at a minimum and to reduce the amplification bias, one must start with genetic material from several thousands of cells. As a consequence, only averages of cell populations can be measured with these methods. Also, due to the large amounts of material required, the analysis of rare cell types, such as adult stem cells, is challenging if not impossible with those methods (Chattopadhyay et al., 2014).

Within the old but flawed paradigm of "genetic determinism" (Strohman, 1997), where genotype maps linearly to phenotype (e.g. protein expression), the measurement of an average quantity seemed reasonable (in a clonal population): The average of the population would be a good representative of the individual cells and not much variation is expected. However, with new technology came new insight, showing that often the population average is not a good description of individual cells.

[1]Much of the material is already lost on the way to the final blot due to the various preprocessing steps.

Let us consider a simple example where a population averaged measurement yields misleading conclusion about the underlying phenomenon: A cell population is homogeneously expressing a certain marker protein (Fig. 1.1A) and is subjected to an external stimulus. As a response to the stimulus, a small fraction of cells upregulates the protein whereas the remaining cells does not respond and their protein levels remain unchanged (Fig. 1.1B). When analyzed e.g. with western blotting, which reports the mean protein expression in the population (black solid line in Fig. 1.1A,B), we observe a slight shift in the mean protein expression level after the stimulus. From that, one might falsely conclude that the entire cell population responded by a small upregulation of the marker, as the population mean obscures the underlying heterogeneous response.

This simple example illustrates the need for experimental techniques that allow to look beyond the mean behavior of a population and analyze single cells individually. However, note that also bulk experiments, when carefully designed and interpreted, harbor certain benefits: They are cheaper and easier to perform, more established and hence widely accepted. Intermediate approaches can also be beneficial: Instead of taking single cells, Bajikar et al. (2014) performed transcriptional profiling on aggregates of up to ten cells, reducing technical variability while single cell information was reconstructed by deconvolution.

1.2.1 Single cell technology

Over the last decade, several single-cell techniques for molecular biology have been developed. The workhorse of modern single cell biology is arguably flow cytometry (fluorescence-activated cell sorting, FACS), which allows to quantify up to 18 different fluorescence characteristics of single cells at the same time. Either by staining proteins of interest with antibodies or utilizing fluorescent reporters and fusion proteins, which are excited by several lasers inside the machine, FACS quantifies their fluorescence intensity and hence their abundance in millions of individual cells. Thus, one can assess the full protein distributions across a cell population instead of only looking at their mean value (see Fig. 1.2). Furthermore, FACS is routinely used as a purification step to sort cells according to their surface marker expression profile in order to get more homogeneous cell populations. Flow cytometry is not limited to measuring only fluorescence characteristics of single cells, but can also be used to acquire digital images of cells passing through the instrument (imaging cytometer, Basiji et al., 2007). This can for example be used to quantify the morphology of the cells (Carpenter et al., 2006) and to predict cell cycle phase (Blasi et al., submitted).

Single-cell quantitative polymerase chain reaction (qPCR) was already performed in the 1990s (Lambolez et al., 1992), but only the integration with microfluidics (Liu et al., 2003) provided the degree of automation, which is required to study the expression of many genes in several hundreds of single cells, e.g. analyzing expression changes in early embryo development (Guo et al., 2010), or characterizing regulatory networks in blood cells (Moignard et al., 2013). mRNA of selected genes are reverse transcribed into cDNA, which is then amplified in PCR cycles. The number of cycles required to yield a detectable amount of cDNA then informs about the abundance of the original mRNA template (larger cycle times mean less mRNA). Whereas single-cell qPCR yields relative transcript abundance (in terms of cycles), a modification termed digital RT-PCR produces absolute

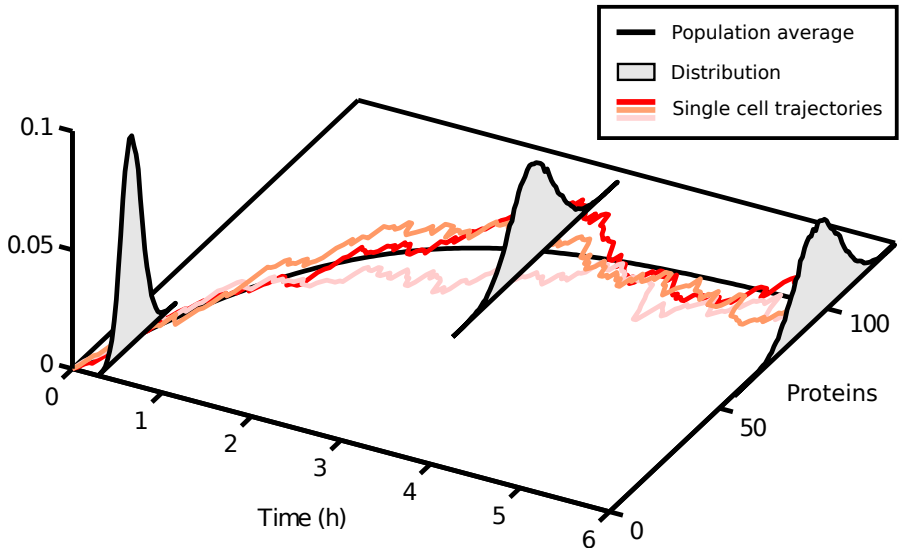


Figure 1.2: **Beyond the population average.** Instead of measuring only the population average protein expression over time (black line), single cell technology allows to observe the evolution of the entire protein distribution, e.g. via FACS snapshots, (gray areas) and the continuous observation of single-cell protein timecourses via time-lapse microscopy (colored lines).

numbers of transcripts in single cells (Warren et al., 2006). Single-cell qPCR can currently quantify the expression of 96 genes in parallel, but to quantify the whole transcriptome of a single cell, different methods must be used.

DNA-Sequencing technology has improved greatly in terms of throughput and precision, but also reduced the amount of DNA-material needed for the analysis, enabling application to single cells. Using reverse-transcriptase to create cDNA from mRNA, one can cast the problem of quantifying the transcriptome of a single cell into a DNA-sequencing problem, which can efficiently be solved with next-generation sequencing technology. Various protocols for single cell RNA-sequencing are available (Tang et al., 2009; Islam et al., 2011; Ramsköld et al., 2012; Sasagawa et al., 2013; Hashimshony et al., 2012), differing in their ways of amplification, multiplexing and sequencing (for a review of current protocols, see Shapiro et al., 2013). Again, these methods provide no absolute numbers of transcripts in the cell per se, but only read counts which have to be mapped to absolute numbers. Furthermore, technical noise is large (Brennecke et al., 2013) and only highly expressed mRNA can reliably quantified. However, recently Islam et al. (2014) tagged each individual mRNA molecule in a single cell with a different barcode before sequencing, which allows to measure directly absolute numbers of mRNAs across the whole transcriptome and reduces technical noise substantially allowing also reliable quantification of lowly expressed genes (on the order of ten mRNAs) in single cells.

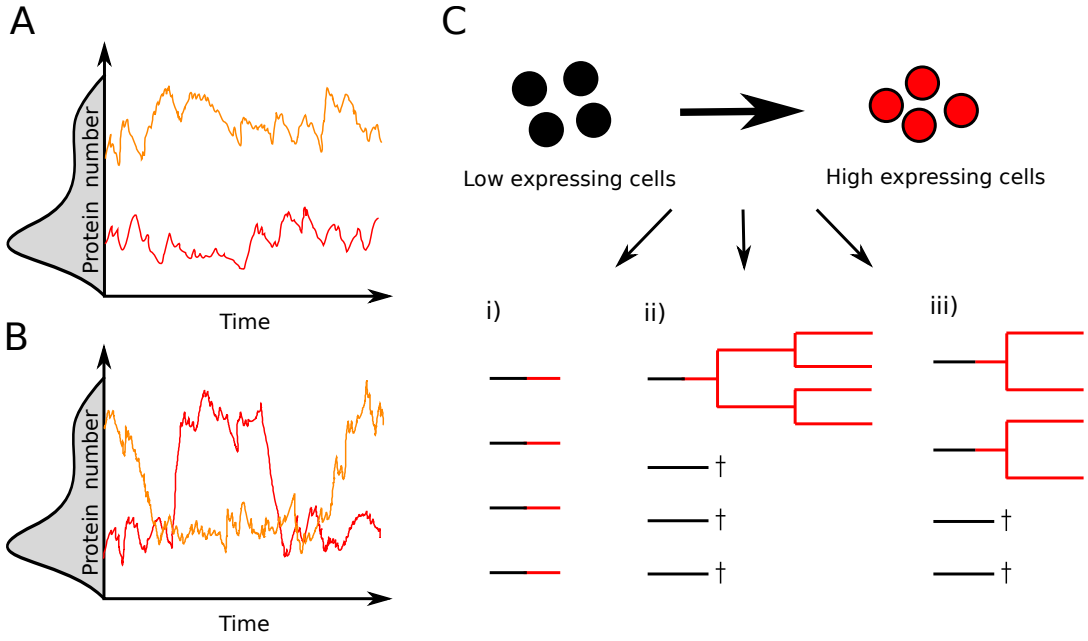


Figure 1.3: **Heterogeneity revealed by time-lapse microscopy**. A bimodal protein distribution can be generated by two separated, stable subpopulations (A) or the rapid transitions between the two modes of expression. Single-cell timecourses are shown in red and orange. C) An experiment starting with four cells with low expression and stopping with four cells with high expression can be interpreted in several ways, which can be distinguished by time-lapse microscopy.

Time-lapse microscopy

All the above methods provide only a static snapshot of the system under investigation, i.e. a measurement of the gene expression profile at a single timepoint[2]. Although being much more informative than population averages, these methods cannot elucidate how the individual cells within a population evolve over time. However this dynamic information is crucial to ultimately understand the nature of cell population heterogeneity: A bimodal distribution in a protein could result from two distinct and separated subpopulations of cells or from rapid transitions between the two peaks (Fig. 1.3A,B). Additionally, one has to consider the fact that cells can divide or die in the course of the experiment. For example, an experiment that starts with 4 cells from one peak of the protein distribution and stops with 4 cells in the the other peak of the distribution at some later timepoint can be interpreted in several ways (Fig. 1.3C): i) Either all 4 cells switched from one peak to the other, or ii) one cell switched, divided twice while the other three cells died, or iii) two cells switched, divided once while the other two died, etc... This ambiguity makes interpretation of snapshot data difficult (Schroeder, 2008).

Time-lapse microscopy is a powerful tool to dissect the ambiguities arising form snapshot data (Coutu and Schroeder, 2013; Schroeder, 2011). Here, cells are first purified by

[2]To some extent, FACS can be iteratively applied to the same set of cells, yielding snapshots of the same cells at different times. However, cell identity is lost from one iteration to the next.

FACS, put under an optical microscope and continuously imaged for the desired period of time ranging from a few hours to several days. Brightfield images provide information about the position and morphology of the cells while fluorescence images can be used to quantify surface marker expression (via in-culture-antibodies) or intracellular protein expression (via reporter constructs or fusion proteins). Additionally, the spatial arrangement of cells can be read out directly from the images, which is of interest when studying e.g. cell-cell communication and cell-niche interaction (Wang and Wagers, 2011; Rompolas et al., 2013).

In order to follow individual cells over time, single cell tracking has to be applied to the microscopy data, i.e. each cell in the current image must be identified in the consecutive image for all images of the experiment. Depending on the cell system analyzed, either automatic tracking algorithms can be applied (for a comparison of current cell tracking algorithms, see Maska et al., 2014) or cells have to be tracked manually (Rieger et al., 2009; Schwarzfischer et al., submitted). Upon cell division, both daughter cells have to be tracked, leading to entire genealogies of cells. Hence, time-lapse microscopy not only allows to observe the time evolution of e.g. proteins in single cells, but it also reveals how similar the offspring of the same common ancestor cell behaves. When combined with quantification of fluorescence-tagged proteins (Schwarzfischer et al., 2011; Schwarzfischer et al., submitted), single-cell trajectories of protein expression can be obtained (see Fig. 1.2).

Although powerful, time-lapse microscopy faces certain challenges. Imaging cells and keeping them alive over long periods of time requires optimal experimental setup, often at the expense of lower image quality (Schroeder, 2011). Tracking cells in experiments where cell density grows fast (e.g. mouse embryonic stem cells) or cells move quickly (e.g. hematopoietic stem cells) renders automatic tracking impossible and manual tracking becomes a major bottleneck in data analysis. The biggest challenge of time lapse microscopy is, similar to FACS, its limitation to just a few factors that can simultaneously be quantified. Even though the number of different fluorescence dyes is increasing, for each and every protein of interest a new genetic modification has to be introduced into the cells, be it a reporter construct or a fusion protein. While it is feasible to construct multicolored bacterial strains or yeast, and to some extend mammalian cell lines, creating a two color mouse strain is a long and costly endeavor. Hence, in the future time-lapse microscopy has to be combined with other single cell technologies to unravel the full complexity of cellular heterogeneities.

1.2.2 Examples of cellular heterogeneity

All of the mentioned single cell methods provide measurements of certain quantities (mRNA or protein expression) not as a population mean, but show how it is distributed across the entire population. Each of these methods showed that a long standing (Novick and Weiner, 1957; Spudich and Koshland, 1976) but also long neglected idea is indeed true: A supposedly homogeneous cellular population, e.g. a single cell type, is in fact often heterogeneous. Heterogeneity of a quantity simply means that its average is not a good description of the entire distribution.

For example, using flow cytometry Chambers et al. (2007) established that the pluripotency factor Nanog is heterogeneously expressed in mouse embryonic stem cells, that Nanog-low cells are more prone to differentiation than Nanog-high cells, and that the whole Nanog distribution can be reconstituted from subsets of cells. Using single-cell qPCR, Guo et al. (2010) showed heterogeneity at early stages of embryo development and Dalerba et al. (2011) analyzed heterogeneous expression profiles of colon cancer cells. Using RNA-sequencing, it was demonstrated that after triggering immune cells with lipopolysaccharide, they responded heterogeneously not only in their gene expression profile but even in which splicing variants were used (Shalek et al., 2013).

Single-cell time-lapse microscopy was used to e.g. delineate the instruction of cell fate via cytokines (Rieger et al., 2009), the origin of mammalian blood (Eilken et al., 2009), or the influence of spatial organization on stem cell fate in the hair follicle (Rompolas et al., 2013). By combining single cell tracking with fluorescence quantification (Schwarzfischer et al., submitted) and cell segmentation (Buggenthin et al., 2013), one can shed light on the transcription factor dynamics during cell differentiation (Hoppe et al., in revision).

1.2.3 Origins of heterogeneity

Single cell technology has revealed unprecedented heterogeneity in many different cell systems ranging from bacteria to mammals. Where does this heterogeneity come from?

Heterogeneity could stem from genetic differences, e.g. somatic mutations (Fig. 1.4A). In fact, in the past very often any heterogeneity has been attributed to genetic differences as postulated by genetic determinism (for a critical review and argument against this idea, see Strohman, 1994). While indeed relevant e.g. in cancer biology, where cancer cells can acquire different traits due to somatic mutations (Diaz Jr et al., 2012), somatic mutations are extremely rare in normal cell populations and cells are considered to be clonal. The fact that all different cell types (with vastly different phenotype) within an organism share identical genetic material already point out that other non-genetic mechanism must be at work (Strohman, 1997). Even in cancer cells, it was shown that some acquired resistance is not due to somatic mutations but caused by non-genetic heterogeneity (Pisco et al., 2013).

Non-genetic heterogeneity could arise from non-observed, confounding factors (Snijder and Pelkmans, 2011). An apparently heterogeneous distribution in protein expression could originate from a difference in cell cycle progression (Buettner et al., 2014) or cell growth (Blasi et al., in revision). Assuming that cells double their amount of the protein during cell cycle, and if the population of interest is not synchronized with respect to cell cycle, a protein distribution will look heterogeneous, even though originating mostly from the difference in cell cycle progression (Fig. 1.4B). In several studies it was shown that the apparent phenotypic heterogeneity can be reduced considerably when including other predetermined factors: Snijder et al. (2009) show that the heterogeneity of virus infection in *E. coli*, i.e. whether cells get infected or not, is largely determined by local cell density and the cell's position in the colony (termed population context). The upregulation of the arabinose utilization system in response to an arabinose stimulus is heterogeneous due to preexisting differences in the number of arabinose transporters (Megerle et al., 2008; Fritz et al., 2014). Similarly, Spencer et al. (2009) showed that the response of a cell

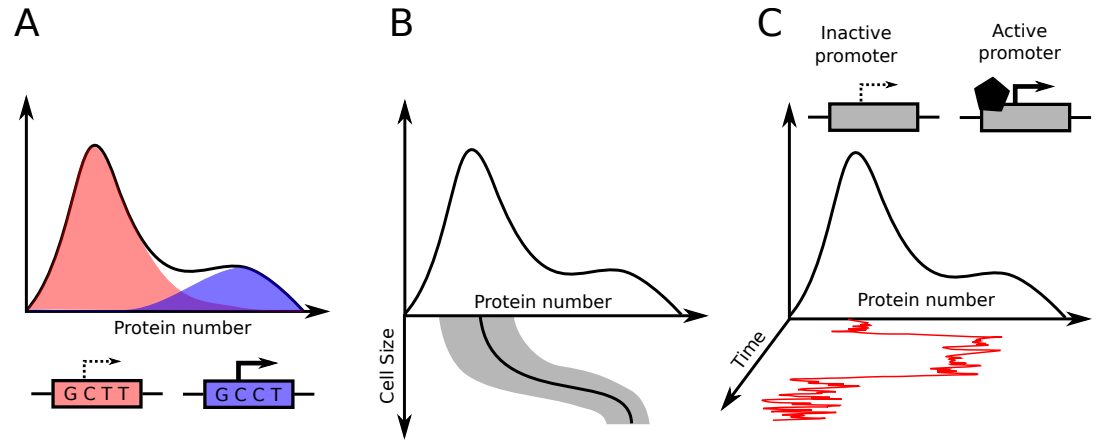


Figure 1.4: **Origins of cellular heterogeneity.** A) Genetic differences cause heterogeneity in protein expression. A somatic mutation in the promoter causes increased expression in a fraction of cells. B) Non-genetic heterogeneity can arise from unaccounted confounding factors that are different within the population. For example, cell growth causes large differences in protein levels. C) Non-genetic heterogeneity results from intrinsically stochastic gene expression, which can be observed from single-cell protein timecourses.

to TRAIL induced apoptosis is heterogeneous but can to a large extent be explained by preexisting differences in protein expression. Whereas phenotypic heterogeneity due to different population context (i.e. external signals) is sensible, phenotypic heterogeneity due to heterogeneous protein expression (as found e.g. by Spencer et al., 2009) raises the question where that internal variability originates from. Furthermore, confounding factors often explain much but not all of the variability observed.

Ultimately, cellular heterogeneity can be traced back to the process of gene expression itself (Fig. 1.4C). Gene expression is an intrinsically stochastic process (for a review, see Kaern et al., 2005). Transcription of a gene is initiated by factors that bind at or upstream of the gene's promoter, e.g. core transcription factors and activators. These factors then facilitate the binding of the RNA-polymerase, which in turn synthesizes mRNA from the DNA template. These binding events result from random collisions of these molecules, and thus considered to be inherently stochastic. For the same reasons, the complex processes of translation, mRNA and protein decay are stochastic. Due to the potentially low number of molecules involved (e.g. two DNA molecules containing the gene) these stochastic fluctuations, which often referred to as "gene expression noise", have to be accounted for (McAdams and Arkin, 1999). The importance of stochastic fluctuations in genetic circuits was already appreciated in 1985 by Shea and Ackers in a model of the lysogenic-lytic switch in the λ-phage and was suggested to contribute to unexplained phenotypic variation by McAdams and Arkin (1997). However, only the seminal study of Elowitz et al. (2002) provided experimental prove of intrinsic fluctuations of proteins due to gene expression in *E. coli*: Using two fluorescence reporters of different color but controlled by identical regulatory sequences, the authors could show that within the same cell, the expression of these two genes was indeed intrinsically noisy. However, some variation in the overall protein level could not be attributed to intrinsic noise, since it affected both genes

in the same way. It was suggested to originate from upstream factors, such as the amount of RNA-polymerase per cell and was termed extrinsic noise. Using the same two color reporter assay, the similar results were obtained also for yeast (Raser and O'Shea, 2004). Whereas these studies were limited to single genes, genome wide measurements of yeast (Newman et al., 2006) and *E. coli* (Taniguchi et al., 2010) protein distributions showed that gene expression noise is not an isolated incident but a genome-wide phenomenon.

With these experimental confirmations of gene expression noise, new theories of gene expression were required that account for the inherent stochasticity as classical deterministic models could no longer describe the observed data (for an introduction to stochastic systems, see chapter 2). For example, Thattai and van Oudenaarden (2001) showed that the variance in protein expression of a single unregulated gene mainly depended on the "translational burst size", that is, the number of proteins translated per mRNA. Additionally, it was discovered that also "transcriptional bursts" contributed strongly to the variance in protein expression (Raj et al., 2006). These studies showed how fluctuations in low copy number species (DNA, mRNA) can propagate to the protein level and cause large variability, even in the regime of large protein numbers.

To describe the observed distributions, analytic expressions for the protein distribution were derived for many small gene expression circuits, for example a two or three stage model of gene expression (Shahrezaei and Swain, 2008; Bokes et al., 2011; Elgart et al., 2011; for a review of stochastic gene expression models, see Paulsson, 2005), a model including cell division (Friedman et al., 2006), self-regulating genes (Ramos et al., 2011; Hornos et al., 2005; Grima et al., 2012) and cascades (Walczak et al., 2009). However, analytical expression for the distributions of more complicated gene expression networks are not available. Here, one has to resort to Monte Carlo methods to approximate the underlying distribution of the system (see chapter 2, section 2.2.3).

While the stochasticity of the process complicates theoretical analysis, it must be stressed that the fluctuations also harbor additional information which can be exploited (Munsky and Neuert, 2012). For example, one can identify additional parameters of the gene expression model (Munsky et al., 2009), infer the transcriptional dynamics (Suter et al., 2011; Harper et al., 2011), distinguish different promoter models (Neuert et al., 2013), determine the influence of cytokines on the frequency and magnitude of transcriptional bursts (Molina et al., 2013) and even use these fluctuations for network inference (Dunlop et al., 2008).

Apart from being an interesting subject of study for the theoretician, however, one has to raise the question how the cell handles these noisy signals. Does a cell merely cope with the fluctuations, or does noise have a function that a cell benefits from (for a review, see Eldar and Elowitz, 2010). Indeed, evidence was found that yeast cells utilize stochasticity to coordinate the expression of many genes in response to calcium signaling (Cai et al., 2008). In bacteria, random switching of cell fate caused by fluctuating protein levels can provide a fitness advantage: due to random switching, a small portion of the population remains in a persister state, which is not susceptible to antibiotic and can repopulate after antibiotic treatment, a phenomenon called "bet hedging" (Lewis, 2007; Veening et al., 2008). Furthermore, noise enables probabilistic differentiation, a simple mechanism to determine cell fate: Identical cells choose their fate randomly due to the fluctuations in important key regulators, as for example observed at the first differentiation decision in the

inner cell mass (Morris et al., 2010; Dietrich and Hiiragi, 2007). Without noise, external signals would be needed to instruct the otherwise identical cells into different fates.

Random cell fate choices at first seem to contradict the observed precise development of an organism. Note however, that these fluctuations could simply be used to break the symmetry and start differentiation in otherwise identical cells, whereas cell-cell communication creates feedback mechanisms to ensure precise development. For example, in the Drosophila epidermis, one cell in a proneural cluster stochastically differentiates into a neuroblast and in turn inhibits all other cells in the cluster from becoming neuroblasts via Delta-Notch signaling, a mechanism called lateral inhibition (Skeath and Carroll, 1992; Losick and Desplan, 2008).

In conclusion, there is strong evidence that noisy gene expression is not just an unavoidable physical phenomenon that has to be controlled, but it can be exploited by organism to gain evolutionary benefits.

1.3 Stem cells in the epigenetic landscape

To gain a deeper understanding about the mechanisms of pluripotency, differentiation, self renewal and cell state transitions in general in the presence of stochastic fluctuations, one has to define what constitutes the "state of a cell". Clearly, the state of the cell is not solely determined by its nucleotide sequence, as all cells within a single organism carry the same genetic material. It is rather determined by the composition of transcripts, proteins (most importantly transcription factors), metabolites and also the state of the DNA itself such as DNA-methylation, chromatin/histone modifications and transcription factor occupation. This is called the epigenetic state of the cell, because it determines the cell's identity (the phenotype) beyond its genetic material (the genotype) and can be inherited from one cell to its offspring[3]. In simpler words, in the epigenetic state of an enterocyte, an intestinal cell type, certain genes e.g. important for the digestion and uptake of food are expressed, whereas proteins required for the formation of neural synapses are repressed. The same genes are of course present in neurons, but enterocyte-specific genes are suppressed in this epigenetic state, whereas neuron-specific genes are upregulated.

However, a cell cannot reach all possible epigenetic states[4] simply because of the regulatory interactions encoded in the genome (Davidson and Erwin, 2006): Genes influence the expression of other genes through an intricate network of molecular interactions, known as the gene regulatory network (GRN, Kauffman, 1969). For example, a transcription factor can bind to the promoter of another gene and repress its expression. Due to this interaction, an epigenetic state where both the transcription factor and its target are expressed is not reachable for the cells. The GRN not only imposes constraints on the epigenetic states but also determines how the epigenetic state of a cell changes over time. Here it is important to clarify that the network topology and hence the regulatory links are fixed on the timescale of cell lifespan and only the state of the cell changes. The topology of the

[3] Note that there is an unfortunate redefinition of the term "epigenetics" in molecular biology, where it only describes the recently discovered phenomena of DNA-methylation and covalent modifications of histones.

[4] Assuming the epigenetic state is just composed of on/off states of 25000 genes in a human cell, this results in 2^{25000} possible states, far more than there are atoms in the observable universe.

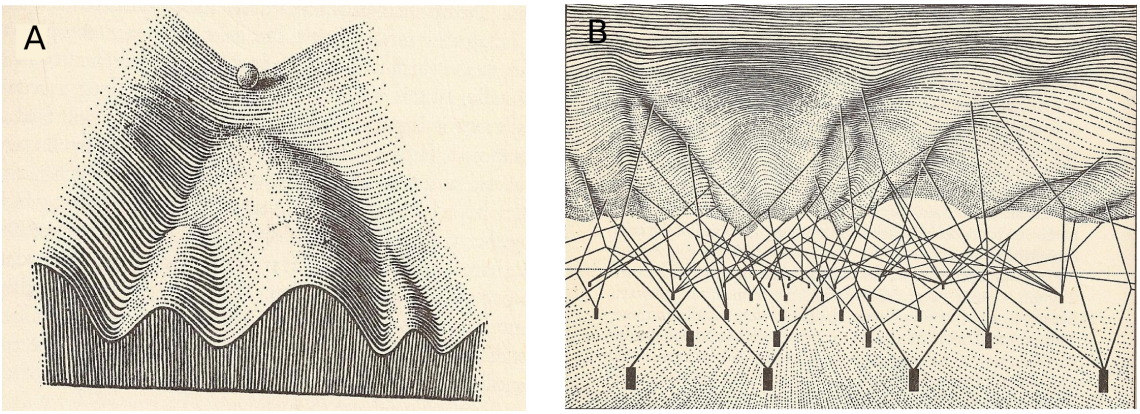


Figure 1.5: **The epigenetic landscape** as originally depicted by C. Waddington (Waddington, 1957).

network only changes on much longer timescale due to mutations (for a detailed argument, see Huang, 2012).

One can then analyze the dynamics of such GRN's using mathematical tools (see chapter 2, section 2.1.2) to find attractor states of the network, i.e. states towards which the systems tends to move and returns to after small perturbations (Beuter, 2003). These states are associated with stable cell types (Huang et al., 2005). One GRN can give rise to multiple attractor states and cell state transitions, such as differentiation and apoptosis occur if the state of the system moves from one attractor to another.

To gain a more intuitive understanding of this abstract concept, it is illustrative to take a detour in history into the 1950's, when Conrad Waddington put forward his idea of the epigenetic landscape (Waddington, 1957). He describes cell differentiation during embryogenesis as a marble rolling down a hilly landscape. Starting at high elevation, the cell is pluripotent and as it starts to descend, it encounters several branching points, where it has to decide its future path. After several of those branching points, the cell will come to rest a one of several possible valleys, corresponding to different mature cell fates. This is the famous picture of the epigenetic landscape (Fig. 1.5A) which has recently been revived and is used as a metaphor in modern stem cell biology (Graf and Enver, 2009; Fisher and Merkenschlager, 2010; Furusawa and Kaneko, 2011; Sisan et al., 2012; Wang et al., 2011; Qiu et al., 2012; Wang et al., 2010). Waddington also gave a hint on what determines the shape of this landscape in a less famous picture (Fig. 1.5B), showing the bottom side of the landscape which is supported by a network of wires anchored by genes. This is a surprisingly accurate description of our current understanding that the GRN constrains the possible epigenetic states (hilltops in the landscapes are not reachable) and the network determines how the cell's state changes over time by shaping this landscape (Huang, 2012). Attractor states correspond to local minima in the landscape, into which the system will eventually settle. However, the motion is not a continuous downward flow in the landscape until the cell reaches its final attractor as depicted by Waddington, but the cell can temporarily get trapped in local minima (attractors) in the landscape, which correspond to intermediate cell types.

What is still missing to close the gap between the GRN and the epigenetic landscape, is how to translate the wiring of the GRN into a landscape, such that state changes allowed by the network correspond to downhill movements in the landscape. In physics terminology: how to find a potential such that the forces acting on the system due to the GRN correspond to the potential gradient? Note that this is not only useful for visualization, but also defines the relative stability of the states (that is, which states correspond to mature cells and which to stem cells) and allows to judge the height of barriers between states. Unfortunately this construction is not possible in general[5], but several approximations were proposed (Zhou et al., 2012; Bhattacharya et al., 2011; Wang et al., 2011), resulting in "quasi-potentials", i.e. potentials whose gradient most accurately reflects the dynamics of the GRN.

Here it is noteworthy that, even though these quasi-potential landscapes offer a intuitive visual link to Waddington's epigenetic landscape, interpretation can be difficult. First, being only an approximation to a true potential, path independence in a quasi-potential is not fulfilled, i.e. the action of a certain path does not only depend on the value of the potential at the start and endpoint, but on the actual path. Second, the state of the system (the marble in Waddington's picture) does not have inertia. Third and most important: The state does not strictly move down the gradient for two reasons: Additionally to the force along gradient, there is a remainder force in the system, which can be for example perpendicular to the gradient (depending on the chosen approximation as reviewed by Zhou et al., 2012). Apart from these deterministic components, there is also a stochastic component to the time-evolution of the state due to gene expression noise. Fluctuation in protein levels can move the system against the gradient and eventually lead to barrier crossings, i.e. cell state transitions such as differentiation.

1.4 Toggle switches in binary cell fate choice

Since studying cell state transitions within the entire regulatory network is still difficult[6], it is useful to study smaller regulatory motifs of just a few genes.

A simple motif of cell fate choice was suggested by discoveries in the blood system (for a review, see Graf and Enver, 2009): Forced expression of the myeloid-associated transcription factor PU.1 in a erythroid-megakaryocytic cell lineage led to activation of myeloid lineage markers and also downregulated erythroid genes, effectively converting them into myeloid cells (Nerlov and Graf, 1998). On the other hand, myeloid cells could also be converted into erythroid cells by forced expression of the erythroid transcription factor Gata1 (Kulessa et al., 1995; Visvader and Elefanty, 1992; Heyworth et al., 2002). In combination with the finding that both factors mutually inhibit each other's expression (Zhang et al., 1999; Stopka et al., 2005) and autoactivate (Okuno et al., 2005; Yu et al., 2002), this established the idea of transcription factor cross-antagonism in binary cell fate choice (Graf and Enver, 2009; Zhou and Huang, 2011) (Fig. 1.6A): The cell fate choice is implemented molecularly by two mutually inhibiting transcription factors and

[5]The driving force from the GRN can contain curl, which cannot be represented by a gradient, as it is non-integrable.

[6]The epigenetic landscape is hard or impossible to obtain mathematically for large networks and the network structure might not be known completely.

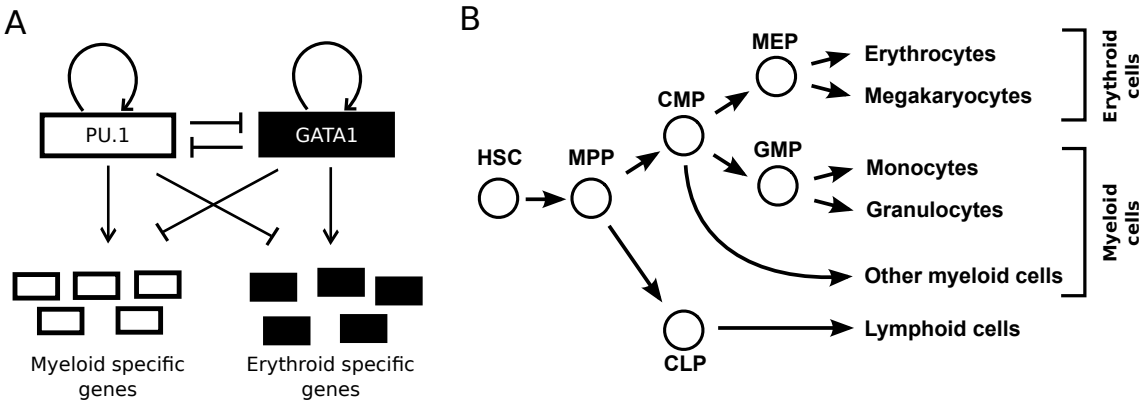


Figure 1.6: **Toggle switches implement binary lineage choice via cross antagonism.** A) PU.1 and Gata1 autoactivate their expression and mutually inhibit their expression. Additionally, PU.1 (Gata1) activates myeloid (erythroid) genes and inhibits erythroid (myeloid) genes. B) Hematopoietic differentiation hierarchy. Hematopoiesis is currently viewed as a hierarchy of differentiation processes (Orkin and Zon, 2008). From a hematopoietic stem cell (HSC), mature blood cells are replenished via a series of progenitor cells with limited potential. Abbreviations: MPP, multipotent progenitor; CMP, common myeloid progenitors; MEP, megakaryocyte-erythrocyte progenitor; GMP, granulocyte-monocyte progenitor; CLP, common lymphoid progenitor.

their balance determines the outcome of the lineage choice. Initially, in a phase called "priming", both transcription factors are balanced, such that none overwhelms the other and the cell is not yet committed to either lineage. If at some point the balance is tilted in favor of PU.1, it will repress Gata1, activate the myeloid-specific genes and lead to myeloid commitment. If the balance is shifted towards Gata1, erythroid genes are activated, PU.1 is repressed and the cell will commit to the erythroid lineage. Due to their switch like behavior, these regulatory motifs are called genetic "toggle switches".

Many other cross-antagonistic transcription factors and their involvement in binary lineage decisions have been identified lately (Graf and Enver, 2009; Zhou and Huang, 2011), either in individual experiments (e.g. granulocytes against macrophages (Laslo et al., 2006), or trophectoderm against inner cell mass (Niwa et al., 2005)) or computational studies (Heinäniemi and Nykter, 2013). By combining multiple toggle switches, one can recapitulate the hierarchical branching of different cell types (Foster et al., 2009). For example, Krumsiek et al. (2011) showed how a network of 11 transcription factors, including several toggle switches, can give rise to the experimentally observed hierarchy of blood cell progenitors.

The PU.1/Gata1 motif still serves as the leading paradigm of cross antagonism and cell fate choice, also because of the multitude of theoretical work: To gain more insight into the experimental findings, several theoretical studies analyzed the properties of the PU.1/Gata1 toggle switch in detail (Roeder and Glauche, 2006; Huang et al., 2007; Chickarmane et al., 2009; Bokes et al., 2009; Duff et al., 2011; Foster et al., 2009), assuming deterministic dynamics. Of main interest in those studies is if indeed two mutually inhibiting transcription factors can generate multi-stable dynamics (in terms of the epige-

netic landscape, multiple attractors) which give rise to the observed progenitor-, myeloid- and erythroid-states. For example, Roeder and Glauche (2006) investigate different mechanisms of mutual inhibition, apply bifurcation analysis to determine under which conditions the system is multi-stable, and show in silico how the forced expression of one or the other factor leads to lineage conversion, recovering experimental results. While this study does not explain how the toggle switch actively carries out a lineage decision, but only how it is stabilized, Huang et al. (2007) proposed that a change in autoactivation strength leads to the loss of the progenitor state, thereby forcing cells to differentiate into either lineage. Furthermore, the authors analyze the details of transition dynamics following the loss of the progenitor state and conclude that the transition dynamics recapitulate microarray measurements.

However, these studies are based on deterministic models, neglecting low copy numbers of e.g. PU.1 mRNA (Warren et al., 2006) and hence the possibility of stochastic cell fate transitions. Furthermore population-averaged measurements were used, hindering clear interpretation on the single cell level. Whereas is is undoubted that the transcription factors PU.1 and Gata1 are involved in the myeloid/erythroid lineage decision, no evidence is yet available whether these two factors indeed actively "make" the decision or if they just implement and lock down an upstream signal. Recent experiments using single cell time-lapse microscopy have questioned the active decision making function of PU.1/Gata1 and provide the necessary data to study the stochastic dynamics of this switch (Hoppe et al., in revision).

1.5 Research questions

The main goal of this thesis is to investigate cell fate decisions in time-lapse data at the single cell level using mathematical models and to assess the role of stochasticity on cell fate choice, with a focus on genetic toggle switches as a molecular implementation of cell fate choice.

First, we ask how stochasticity impacts on cell fate choice and whether it is compatible with observed data. As cell fate choice is implemented molecularly in the gene regulatory network, e.g. via a toggle switches, which itself is subject to fluctuations from gene expression, the process of cell fate choice itself has a stochastic component. However, to which extent this stochasticity plays a role in cell fate choice is still being discussed: Cell differentiation might be entirely stochastic and cell intrinsic (Gomes et al., 2011; Till et al., 1964; Roeder et al., 2005; Abkowitz et al., 1996), regulated by external stimuli (Rieger et al., 2009; Moore and Lemischka, 2006), or even be predetermined to a large degree (Müller-Sieburg et al., 2002). In this thesis, we develop methods that allow to analyze the mechanisms behind cell fate choice, show that simple stochastic models can recapitulate seemingly complex phenomena such as lineage priming or lineage bias, and that these models can explain observed data.

Second, we ask how one can incorporate and utilize the genealogical structure inherently emerging from a growing and dividing cell population. On the one hand, these cellular genealogies present an obstacle to standard statistical analysis, as related cells do not qualify as independent samples. On the other hand, genealogies also provide valuable information. This is for example utilized in "paired daughter cell assays" (Suda et al.,

1984), where the two daughter cells arising from a cell division are separated via micromanipulation and individual experiments are performed on these cells, e.g. colony assays to quantify how variable these daughter cells are in terms of their lineage potential (Takano et al., 2004). However, it is yet unclear how to exploit this information systematically on a larger scale, i.e. not only on pairs of daughter cells, but whole genealogies. Hence, new methods and models have to be developed which account for, and utilize this genealogical information.

Finally, we investigate the current paradigm of the PU.1/Gata1 toggle switch governing the hematopoietic cell fate decision between myeloid and erythroid lineages applying the methods developed in this thesis to hematopoietic stem cell genealogies generated by Hoppe et al., in revision. Specifically, we assess whether the PU.1/Gata1 toggle switch actively determines the cell fate choice via its dynamics, or if PU.1 and Gata1 merely serve as a memory device that is linked to an unobserved upstream decisions (e.g. via mechanisms proposed by Fritz et al., 2007; Hillenbrand et al., 2013).

1.6 Overview of this thesis

In chapter 2 we briefly introduce the formalism of dynamical stochastic systems, show how the deterministic reaction rate equations are derived from the stochastic system, and review Approximate Bayesian Computation as a tool to perform inference for stochastic systems.

In chapter 3, we study the dynamics of a two stage toggle switch as a potential internal mechanism of binary cell fate decisions. We investigate the system's quasi-potential landscape and find that using a two stage gene expression model induces four attractor states as opposed to using a one stage expression model, where transcription and translation are lumped together. We analyze the dynamics of the system in the quasi-potential and associate two attractors with differentiated cell types and two attractors with undifferentiated cell types, which are however already biased in their lineage choice. Furthermore, we provide analytical expressions of the attractors residence times. Finally, we fit a toggle switch model to single cell data from GMP differentiation using Approximate Bayesian Computation. The presented methods, figures and results have been published in Strasser et al. (2012) and Marr et al. (2012). We thus contribute to the current understanding of stochastic toggle switch models and their dynamics and furthermore show how stochastic, mechanistic models can be linked to experimental observation of cell fate choice.

In chapter 4, we study different coarse grained models of cell fate decision, where the cell fate decision does not depend on internal dynamics but is governed by global, external influences, such as local cell density. We present an inference framework based on regularized linear models, which identifies the relevant external influences impacting on differentiation from cellular genealogies with annotated differentiation events. We use the framework to predict the required sample size for the analysis and explore the impact of tracking errors within the genealogies on our results. Here, our contribution is the adaptation of available statistical methods to tree-structured data.

In chapter 5 we consider models where cell fate decisions depend both on external influences and cell internal dynamics and present an algorithm that estimates parameters of this model from cellular genealogies. We show that the model can accurately detect differentiation events when the underlying dynamics are governed by a toggle switch. Applying the method to genealogies of differentiating blood stem cells, we present evidence against the long standing paradigm of the PU.1/Gata1 toggle switch in hematopoietic lineage decisions. We contribute a novel model and inference method for genealogies that allows to infer unobserved state changes (e.g. cell differentiation) from correlated fate of genealogically related cells.

In chapter 6, we summarize the thesis and present an outlook on future research directions.

Chapter 2

Methods

In this chapter, we introduce the basic mathematical methods used in chapters 3–5. The first part revolves around chemical reaction networks and their associated deterministic and stochastic dynamics. In the second part, we briefly review likelihood-based inference in general and outline Approximate Bayesian Computation as a likelihood free inference method. In the last part of this chapter, we introduce graphical models as well as inference in graphical models via the sum-product algorithm.

2.1 Chemical reaction kinetics

Throughout this thesis, dynamical systems are modeled as chemical reaction networks and in the following, we briefly introduce the basic notations and concepts.

Typically, biochemical processes are modeled as chemical reactions (Wilkinson, 2011; Alon, 2006). Here, the model consists of N species $X_1, \ldots, X_N$ and a set of M reactions $\mu_1, \ldots, \mu_M$. The reaction μ can be written in stoichiometric form as

$$e_{\mu 1} X_1 + e_{\mu 2} X_2 \ldots \xrightarrow{k_\mu} p_{\mu 1} X_1 + p_{\mu 2} X_2 \ldots \ , \tag{2.1}$$

where $e_{\mu i} \in \mathbb{N}_0$ ($p_{\mu i} \in \mathbb{N}_0$) is the number of molecules X_i consumed (produced) by reaction μ. Arranging these coefficients as matrices

$$\begin{aligned} (E)_{ij} &= e_{ij}, \ E \in \mathbb{N}_0^{M \times N} \\ (P)_{ij} &= p_{ij}, \ P \in \mathbb{N}_0^{M \times N} \ , \end{aligned}$$

we can obtain the stoichiometric matrix V of the system as $V = -E + P$. Each row vector $\nu_\mu = [V_{\mu,1}, \ldots, V_{\mu,N}]$ corresponds to the change in species numbers upon one occurrence of reaction μ. The order of reaction μ is defined as

$$o_\mu = \sum_{i=1}^{N} e_{\mu i} \ .$$

The reaction rate constants k_μ quantify the reaction's speed and are related to its activation energy via the Arrhenius equation

$$k = A \cdot e^{-\frac{E_a}{k_B T}} \ ,$$

where E_a is the activation energy, k_B is Boltzmann's constant, T is the temperature, and A is a reaction specific prefactor. The units of the reaction rate constant depend on the reaction's order: For a reaction of order o_μ the rate constant k_μ has units of $\left[s^{-1}M^{1-o_\mu}\right]$, where $M = \frac{\text{mol}}{\text{liter}}$ is molarity. For example, for second order reactions ($o_\mu = 2$), the reaction rate has units of $\left[M^{-1}s^{-1}\right]$.

2.1.1 Reaction rate equations

Up to now, for a given system we only have specified how reactions change the number of molecules, but not the rules that determine the dynamics of the reactions. The traditional method for modeling cellular dynamics is based on ordinary differential equations, which are called reaction rate equations and describe macroscopic dynamics:

$$\frac{\partial x}{\partial t} = f(x,t) \tag{2.2}$$

with $x \in \mathbb{R}_0^N$ and $f : \mathbb{R}_0^N \times \mathbb{R} \to \mathbb{R}_0^N$. This equation describes the deterministic time evolution of continuous species concentrations $x(t)$. Typical choices for the function f include mass action (Guldberg and Waage, 1879) or Michaelis-Menten kinetics (Michaelis and Menten, 1913). Applying the law of mass action, one obtains:

$$f_i(x,t) = \sum_\mu \left[\nu_{\mu i} \cdot k_\mu \cdot \prod_{j=1}^{N} x_j^{e_{\mu j}} \right] , \tag{2.3}$$

with $i = 1 \ldots N$. The rate of change in concentration of species X_i is the sum of contributions from the individual reactions μ, where each reaction changes the amount of X_i by $\nu_{\mu i}$. The product over j derives from the mass action assumption, where one assumes that the reaction rate is proportional to the possible number of educt molecule collisions, and hence proportional to powers of the educt concentrations.

In section 2.2.4, we derive the macroscopic reaction rate equations from a more accurate, microscopic and stochastic description of the dynamics and the required assumptions are discussed.

2.1.2 Steady state solutions and stability analysis

Given the set of reaction rate equations of the underlying dynamical system, one is typically interested in the asymptotic behavior of the system, i.e. $t \to \infty$. In the reaction rate equations of the form shown in Eq. (2.2) one sets $\partial_t x(t) = 0$ and solves for x, yielding steady state solutions $\{x^* | f(x^*) = 0\}$.

Linear stability analysis can be applied to characterize the steady states further (Murray, 2002; Beuter, 2003). Here one studies how the system behaves upon infinitesimal perturbations from the steady state and classifies the steady states as either "stable", "unstable", or "saddle point". In the following, this procedure is shortly reviewed.

We start by considering the time evolution of a small perturbation $x - x^*$ from a steady state x^* and linearize f at x^*:

$$\begin{aligned} \partial_t (x - x^*) &= f(x - x^*, t) \\ &\approx J \cdot (x - x^*) , \end{aligned}$$

where

$$J = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1} & \cdots & \frac{\partial f_N}{\partial x_N} \end{pmatrix} ,$$

is the Jacobian of the system. Performing eigendecomposition of J one obtains

$$\begin{aligned} \partial_t(x - x^*) &\approx J \cdot (x - x^*) \\ &= ADA^{-1} \cdot (x - x^*) , \end{aligned}$$

with D as a diagonal matrix of eigenvalues and A the matrix of eigenvectors. Multiplying both sides by A^{-1} from the left yields

$$\begin{aligned} \partial_t \underbrace{A^{-1} \cdot (x - x^*)}_{y} &= D \underbrace{A^{-1} \cdot (x - x^*)}_{y} \\ \partial_t y &= D \cdot y . \end{aligned}$$

Since D is diagonal, we obtain decoupled differential equations for each transformed variable $y_j \in \mathbb{R}_0$ $(j = 1, \ldots, N)$ with solution

$$y_j(t) = e^{\lambda_j t} \cdot y_j(0),$$

where λ_i is the j-th eigenvalue of the Jacobian. In general $\lambda_j = a_j + \imath b_j$ is complex and therefore

$$y_j(t) = e^{a_j t} \underbrace{[\cos(b_j t) + \imath \sin(b_j t)]}_{e^{\imath b_j t}} \cdot y_j(0) .$$

The imaginary part of the eigenvalue induces oscillations (second term), whereas the real part determines the stability of x^*: For $t \to \infty$ and if all $a_j < 0$, we see that all $y_i \to 0$ and therefore also $x - x^* \to 0$. The perturbation vanishes and the system returns to its former steady state. The steady state x^* is stable and often referred to as an attractor since systems near that state evolve towards it. On the other hand, if any $a_j > 0$, the perturbation will not vanish for $t \to \infty$ and the system will leave the former steady state x^*. The steady state x^* is called unstable (or saddle point if its eigenvalues have both positive and negative real parts). If all $a_j = 0$ and any $b_j \neq 0$, the system will oscillate around the center x^* (Murray, 2002).

To conclude the analysis, one classifies the dynamical system (Eq. 2.2), depending on the number of distinct stable steady states x^* as "monostable", "bistable", etc... Note that the number of stable states in general depends on the parameters of the system (e.g. the reaction rates), which is studied in the field of bifurcation theory. Furthermore, for each stable state x^*, one can determine the set of states $\mathcal{A}(x^*)$ which evolve towards x^* as $t \to \infty$:

$$\mathcal{A}(x^*) = \{x_0 \in \mathbb{R}_0^N | x(0) = x_0 \wedge \lim_{t \to \infty} x(t) = x^*\}$$

This set $\mathcal{A}(x^*)$ is called the basin of attraction of stable state (or attractor) x^*.

2.2 Stochastic systems

As discussed in section 1.2.3, the dynamics of gene expression processes have to be considered stochastic rather than deterministic owing to small molecule numbers of e.g. mRNAs. In the following, we recapitulate the formalism to describe these stochastic dynamical systems.

2.2.1 Stochastic chemical kinetics

We assume a well stirred system in a volume Ω, such that positions and velocities of individual molecules can be considered random. This allows to represent the entire state of the system at time t by the vector

$$x(t) = [x_1(t), \ldots, x_N(t)] \ ,$$

where $x_i(t)$ denotes the number of molecules of species X_i at time t. The vector $x(t) \in S$ is called the state vector of the system where $S \subseteq \mathbb{N}_0^N$ is called the state space of the system. Note that the state $x(t)$ of the system is discrete as opposed to the reaction rate equations considered in section 2.1.1.

Next, we define the propensity function a_μ of a reaction μ as

$$\begin{aligned} a_\mu(x)dt :=& \text{Probability that reaction } \mu \text{ occurs in the infinitesimal} \\ & \text{interval } [t, t+dt] \text{ given the system is in state } x \text{ at time } t \end{aligned} \tag{2.4}$$

The form of the function a_μ is derived from molecular physics taking into account collision probabilities and reaction probabilities (Gillespie, 1992):

$$a_\mu(x) = c_\mu \cdot \prod_{i=1}^{N} \binom{x_i}{e_{\mu i}} \ , \tag{2.5}$$

where c_μ is a constant. For unimolecular reactions ($o_\mu = 1$), c_μ is the probability per unit time of a spontaneous quantum-mechanical transition from educt to product. The second term in Eq. (2.5) accounts for the number of molecules that can undergo this conversion. In case of bimolecular reactions ($o_\mu = 2$), c_μ is the probability per unit time that a collision of educt molecules results in a successful reaction. The potential number of educt collisions is specified by the second term of Eq. (2.5).

Note that in general $c_\mu \neq k_\mu$, where the latter is the familiar chemical reaction rate constant. In fact they are related via (Gillespie, 2007)

$$c_\mu = \begin{cases} k_\mu & \text{for unimolecular reactions} \\ k_\mu \Omega^{-1} & \text{for bimolecular reactions with different educts} \\ 2k_\mu \Omega^{-1} & \text{for bimolecular reactions with same educts} \ . \end{cases}$$

Higher order (e.g. trimolecular) reactions are neglected because these are approximations to sequences of bimolecular reactions. In section 2.2.4 we will see the connection between the propensities and the mass-action rates from Eq (2.3).

Ultimately, one is interested in the time evolution of the state vector $x(t)$ through a series of reactions. However, as the timing and order of reaction firings is inherently stochastic (the propensities are formulated as probabilities and $x(t)$ is a Markov Jump process), one can only derive an equation that describes the time evolution of the probability distribution of the state vector $x(t)$ (Gillespie, 1992). This equation is known as the Chemical Master Equation (CME):

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

It describes how the probability $\mathcal{P}(x,t)$ of being in a certain state x at time t changes in an infinitesimal time interval due to reactions into and out of the state x. The first term accounts for reactions leading into x from adjacent states (states that are one reaction away from x), which therefore increase the probability of state x. Similarly, the second term accounts for reactions leading away from the current state x, thereby decreasing the probability to be in x.

Let us briefly characterize Eq. (2.6). First, we find that it is not a single ordinary differential equation, but a system of coupled differential equations, one for each possible state x of the system. Even if our system comprises only a single species, the state space S is potentially infinite, e.g. if the number of molecules is unbounded ($S = \mathbb{N}_0$). If the number of molecules is bounded, the size of the statespace is growing exponentially in the number of species. Second, we observe that Eq. (2.6) is linear in $\mathcal{P}(x,t)$. Note that the propensities can nevertheless be nonlinear in x. If we choose some arbitrary enumeration X of the state space, one can rewrite Eq. (2.6) as

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

where the matrix Q is defined via its elements as

$$(Q)_{xy} = \begin{cases} -\sum_{\mu=1}^{M} a_\mu(x) & x = y \\ a_\mu(x) & y = x + \nu_\mu \\ 0 & \text{otherwise} . \end{cases}$$

x and y are elements of the state space, thus the dimension of Q is equal to the size of the possibly infinite state space.

We conclude that the CME is a system of coupled linear differential equations and the size of the equation system corresponds to the size of the state space.

2.2.2 Analytic solutions to the CME

Analytic solutions to the CME are only available for certain simple systems (Jahnke and Huisinga, 2007; Ramos et al., 2011; Pendar et al., 2013) restricted to steady state distributions (Raj et al., 2006; Friedman et al., 2006; Paulsson and Ehrenberg, 2000; Bokes et al., 2011; Hornos et al., 2005) or are valid only in certain regimes (Shahrezaei and Swain, 2008).

Instead, numerical methods have been applied. Finite state projection (Munsky and Khammash, 2006) tries to find suitable small truncations of the statespace S, allowing for a direct numerical solution of Eq. (2.7). Spectral methods (Walczak et al., 2009; Mugler et al., 2009) exploit the linearity of Eq. (2.7) by solving it in terms of its eigenfunctions without state space truncation. Moment equations methods consider not the time evolution of the whole distribution but of its moments (Lee et al., 2009; Engblom, 2006; Hasenauer et al., 2013).

2.2.3 Stochastic simulation

The above methods cannot be applied if either the relevant state space of the system is to large or if the system itself is too complicated (e.g. because of multiple feedback loops), eluding any sophisticated mathematical treatment. However, one can apply a simple algorithm to draw samples from the stochastic process $x(t)$ governed by the CME, which is known as Gillespie's algorithm or stochastic simulation algorithm (Kendall, 1950; Gillespie, 1976). Drawing sufficiently many samples, one can then approximate $\mathcal{P}(x,t)$, the solution of the CME.

The core algorithm is based on the "reaction probability density function" (Gillespie, 1976):

$$\begin{aligned} p(\tau,\mu|x,t)d\tau :=& \text{Probability at time t that the next reaction to happen is of type} \\ & \mu \text{ and it occurs in the infinitesimal interval } [t+\tau, t+\tau+d\tau] \\ & \text{given the system is in state } x \text{ at time } t. \end{aligned} \tag{2.8}$$

Note the difference to the definition of the propensity in Eq. (2.4): The propensity is the instantaneous probability of a certain reaction, whereas $p(\tau,\mu|x,t)$ informs also about the waiting time.

One can easily derive the form of Eq. (2.8) (Gillespie, 1976, 1992): It is the product of the probability $p_1(\tau|t,x)$ that no reaction happens in $[t,t+\tau]$ and the probability $p_2(\tau|t,x)d\tau$ of reaction μ happening in $[t+\tau, t+\tau+d\tau]$:

$$\begin{aligned} p(\tau,\mu|x,t)d\tau &= p_1(\tau|t,x)\cdot p_2(\tau|t,x)d\tau \\ &= e^{-\sum_i a_i(x)\tau}\cdot a_\mu(x)\ . \end{aligned} \tag{2.9}$$

The first factor is derived from the density of a Poisson distribution with rate $\sum_i a_i(x)\tau$ evaluated at 0, and the second factor follows from the definition of the propensity (Eq. 2.4).

Defining $a(x)=\sum_i a_i(x)$ and rewriting the above as

$$p(\tau,\mu|x,t)d\tau = \left[e^{-a(x)\tau}a(x)\right]\cdot\left[\frac{1}{a(x)}a_\mu(x)\right]\ ,$$

it is apparent that $p(\tau,\mu|x,t)d\tau$ is the joint density of two independent random variables: An exponential random variable with mean $a^{-1}(x)$ (first bracket) and a categorical random variable with probability vector $\left[\frac{a_1(x)}{a(x)},\ldots,\frac{a_M(x)}{a(x)}\right]$ (second bracket).

Hence, one can easily draw samples from Eq. (2.9) and thereby determine the timing and type of the next reaction given the current state x. With that, we can construct the

Algorithm 1: The stochastic simulation algorithm (SSA).

Input: Initial condition x_0, maximal simulation time t_{max}, propensity functions a_i and stoichiometric matrix V

Output: Timecourse of species abundances $x(t)$

$j = 0$;
$t_0 = 0$;
while $t < t_{max}$ **do**
  $a = \sum_\mu a_\mu(x_j)$;
  $\tau \sim Exp(a)$; // Exponential random variable
  $\mu \sim Cat\left(\left[\frac{a_1(x)}{a(x)}, \ldots, \frac{a_M(x)}{a(x)}\right]\right)$; // Categorical random variable
  $t_{j+1} = t_j + \tau$;
  $x_{j+1} = x_j + \nu_\mu$;
  $j = j + 1$
end

stochastic simulation algorithm (Algorithm 1), which starts from an initial state x_0 and initial time t_0 and iteratively updates time and state by executing reactions in accordance to Eq. (2.9). This extremely simple algorithm allows one to obtain samples (i.e. timecourses) and hence to approximate the solution of the CME even if it is infeasible to solve the CME directly.

However, two complications arise: Being a Monte Carlo method, the rate of convergence of this sampling approximation to the solution of the CME is only of order $1/\sqrt{N}$, where N is the number of samples[1]. Hence, a huge number of samples (typically $> 10^3$) is required to achieve acceptable accuracy. Furthermore, the computation of a single sample becomes time-consuming if the number of reactions taking place in the desired time interval $[t_0, t_{\max}]$ is large, as each reaction is simulated individually. This happens, for example, if overall molecule numbers in the system or reaction constants are large. Therefore, the reaction propensities $a_\mu(x)$ and the factor $a(x)$ become large. This in turn leads to small time steps τ by which the algorithm advances, thereby requiring much more iterations to completely simulate the desired time interval $[t_0, t_{\max}]$.

Various exact variations the stochastic simulation algorithm have been developed in the last decades, which are devoted to improve the algorithm's scaling in terms of species number N and number of reactions (Ramaswamy et al., 2009; McCollum et al., 2006; Gibson and Bruck, 2000; Slepoy et al., 2008; Cao et al., 2004). However, all these methods are still subject to the above mentioned problems, as they create exact samples of the underlying stochastic process.

To overcome the issue of computational complexity, numerous approximate algorithms have been developed. They all evolve around the idea that under certain circumstances, it is valid to simulate not every reaction, but lump together many individual reactions into a single simulation step. The idea of τ-leaping (Gillespie, 2001) was proposed first,

[1]Suppose you want to estimate $\mu = \mathbb{E}[f(Y)]$ where Y is a random variable and f is some function. Drawing samples $y_1, \ldots, y_n$, we find $\hat{\mu} = n^{-1}\sum_i f(y_i)$. The variance of the estimator is $\sigma^2(\hat{\mu}) = n^{-2} \cdot n \cdot \sigma^2(Y)$ and the standard deviation $\sigma(\hat{\mu}) = n^{-1/2} \cdot \sigma(Y)$

where one assumes that within a certain time interval the propensities $a_i(x)$ are effectively constant. Therefore the numbers of reactions of each type happening in this interval can be approximated as Poisson random variables and all these reaction are executed in a single simulation step simultaneously. Based on this idea, many modifications of τ-leaping have been derived (Mjolsness et al., 2009; Cai and Xu, 2007; Auger et al., 2006; Peng and Wang, 2007). The state space simulated by τ-leaping methods is still discrete. If the state space is approximated by continuous variables (e.g. because the molecule numbers are large) one naturally arrives at a stochastic differential equation approximation of the CME (Gillespie, 2000). With any of these approximations, one trades accuracy for speed: For example, by leaping over larger time intervals and executing many reactions in parallel, τ-leaping can considerably reduce computational time compared to standard stochastic simulation. However, the larger these intervals, the more the assumption of constant propensity will be violated, resulting in a larger error of τ-leaping compared to the exact stochastic simulation algorithm. In general, these methods perform well if species numbers in the system are large, but can introduce large errors if some species numbers are close to zero. Unfortunately it is often these low species numbers that induce the characteristic behavior of the system (for an example, see Schultz et al., 2008).

Another class of approximations is based on the idea of time scale separation (Cao et al., 2005; Haseltine and Rawlings, 2002), where ones partitions the dynamics of the system into a fast and a slow set and only simulates the slow set explicitly via stochastic simulation, but approximates the evolution of the fast set (for example) deterministically. Huge speedups can be achieved if there is a clear time-scale separation in the system. However, the degree of time scale separation often changes dynamically with the state of the system itself (some regions in state space satisfy time scale separation, whereas others do not) and with the parameters of the system. This renders an automated treatment of the problem difficult and requires some user-specified prior knowledge of the system. Often, a separation of time scales is simply not possible, even tough some reactions are much faster then others (termed the "weakly" adiabatic regime by Walczak et al., 2005a).

2.2.4 Derivation of the reaction rate equation from the CME

In the following, we show how and under what assumptions the classical reaction rate equations emerge as a limit of the stochastic dynamics described by the CME. Note that there are different ways of deriving the reaction rate equations from the CME and for brevity only the approach by Gillespie (2000) will be discussed here (for overviews see e.g. Grima et al., 2011; Grima, 2010a; Gardiner, 2004).

One starts from the stochastic simulation algorithm, which generates exact samples and apply the τ-leaping approximation, where the number of reaction per simulation step follows a Poisson distribution. If this number is large on average it is suitable to approximate the Poisson by a Gaussian distribution. Casting the resulting difference equation into a differential equation, we obtain the following stochastic differential equation known as the Chemical Langevin Equation (Gillespie, 2000)

$$\frac{\partial x}{\partial t} = \underbrace{\sum_{\mu} \nu_\mu a_\mu(x)}_{\text{Drift}} + \underbrace{\sum_{\mu} \nu_\mu \sqrt{a_\mu(x)} \Gamma_\mu(t)}_{\text{Diffusion}} \tag{2.10}$$

for the time evolution of the continuous state vector $x(t) \in \mathbb{R}_0^N$. ν_μ denotes the row of the stoichiometric matrix corresponding to reaction μ. The right-hand side of Eq. (2.10) consists of a deterministic drift term and a stochastic diffusion term driven by independent white noise terms $\Gamma_\mu(t)$. The classical reaction rate equation is recovered in the thermodynamic limit (i.e. system volume and species numbers approach infinity) where the propensities $a_\mu(x)$ become so large that the diffusion term can be neglected due to the square root scaling:

$$\frac{\partial x}{\partial t} = \sum_\mu \nu_\mu a_\mu(x) + \mathcal{O}(a^{1/2}) \; .$$

Using the definition of the propensities Eq. (2.5) and approximating the numerator of the binomial coefficient[2] we find

$$\begin{aligned} a_\mu(x) &= c_\mu \cdot \prod_{i=1}^{N} \binom{x_i}{e_{\mu i}} \\ &\approx k_\mu \cdot \prod_{i=1}^{N} x_i^{e_{\mu i}} \; . \end{aligned}$$

Comparing to Eq. (2.3), we find that, in the thermodynamic limit, we recover the reaction rate equations and see that the propensities correspond to the mass-action rates.

However, one must keep in mind that this derivation and therefore the reaction rate equations itself only holds if the system of interest fulfills all the assumptions of the Langevin Equation and the system size approaches infinity. As most processes of interest happen within a cell's volume or a sub-compartment thereof, the validity of the reaction rate equations in the realm of cell biology is questionable. Several studies indicated the breakdown of classical reaction rate equations at volumes comparable to cells (Grima, 2009b,a; Ramaswamy et al., 2012). To account for these small volume effects, the effective mesoscopic rate equations have been proposed, which augment the classical reaction rate equation by a volume correction (Grima, 2010b).

2.2.5 Stability of states in deterministic and stochastic systems

In section 2.1.2 we calculated the stable states and their stability in the deterministic system (the reaction rate equations), whereas in section 2.2.1 we applied the CME, which describes the stochastic dynamics in terms of probability distributions over state space. What can be concluded from stability analysis of the deterministic system about the underlying stochastic system from which it was derived?

First, one has to note that there are generally no stable states (attractors) in a stochastic system: Due to the inherent probabilistic nature of the process, no single state x will be stable. There is always non-zero probability of leaving the state[3]. Hence, the system will not converge to a single state in the long term limit, even though its distribution

[2] For example, $\binom{n}{2} = \frac{n(n-1)}{2} \approx \frac{n^2}{2}$ if n is large.

[3] Exceptions are absorbing states where no reaction leads out of the state, i.e. $\forall_\mu \; a_\mu(\tilde{x}) = 0$ for the absorbing state $\tilde{x}$.

converges under mild assumptions (Van Kampen, 1992) and one has to come up with a different definition of stable states in a stochastic system. An intuitive choice for an analog of a deterministic stable state is a mode of the probability distribution of the stochastic process. Note that there can be multiple modes, e.g. in the protein number distribution of a self-activating gene (Hornos et al., 2005). The idea is that a single trajectory will fluctuate near a mode of the distribution for a long time and a transition to another mode will happen on a much longer timescale. The region around the mode, the analog to the deterministic basin of attraction, is at least approximately stable on timescales much shorter then the escape time. Note that the escape time can be much longer than the average lifetime of a cell.

Second, the question arises how deterministic stable states and the modes of the stochastic system relate. Unfortunately, no general correspondence exists: Existence of a deterministic stable state does not imply that the solution of the master equation has a mode at this location[4]. On the other hand, the existence of a mode does not imply the existence of a stable state in the deterministic system.

An example discussed in this thesis (see chapter 3 and Strasser et al., 2012) is a model of a two-stage toggle switch without cooperative binding. Stability analysis predicts one stable state (see supplements of Strasser et al., 2012) whereas the distribution of the stochastic model shows four distinct peaks (Fig. 3 of Strasser et al., 2012). Additionally, the deterministic stable state is located in a region of state space where the distribution has negligible probability mass. Therefore, conclusions drawn from the deterministic systems can be arbitrarily wrong in context of the (more accurate) stochastic counterpart.

Finally, it has to be noted that these discrepancies are not due to an inherent flaw of the reaction rate equations. Such discrepancies simply arise because the assumptions needed to derived the reaction rate equations from the Chemical Master Equation are violated for the system of interest. Therefore, one should either study the stochastic system directly, or if complexity does not allow this, one should rigorously check if the assumptions of the reaction rate equations hold.

2.3 Parameter inference

Having set up a (stochastic or deterministic) model of the system of interest, one has to fit the model to observed data to infer the unknown parameters of the system, e.g. the reaction rates. Here, one has to derive the likelihood function $\mathcal{L}(X|\theta)$ of the system, which gives the probability of observing the data X given a set of parameters θ.

2.3.1 Likelihood-based inference

To fit the system to the data in a frequentist approach (Sivia and Skilling, 2006), the maximum likelihood estimate $\hat{\theta}$ is obtained by optimizing the likelihood with respect to the parameters:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ \mathcal{L}(X|\theta) \tag{2.11}$$

[4]Note that there are certainly systems where this correspondence holds, for example a simple birth death process.

Consider the following example: We observe n independent and identically distributed data points $X = (x_1, \ldots, x_n)$ from a Gaussian distribution, whose mean μ and variance σ^2 are unknown and have to be inferred. The likelihood of a single datum x_i given parameters μ, σ^2 is a Gaussian density:

$$\mathcal{L}(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Due to the independence and identical distribution of the x_i, the likelihood of the whole dataset X is the product of individual likelihoods:

$$\begin{aligned}\mathcal{L}(X|\mu, \sigma^2) &= \prod_{i=1}^{n} \mathcal{L}(x_i|\mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\end{aligned}$$

To obtain e.g. the maximum likelihood estimate $\hat{\mu}$ via Eq. (2.11) we have to find the maximum of $\mathcal{L}(X|\mu, \sigma^2)$ with respect to μ. Here, it is more convenient to use the log-likelihood, which has the same maxima as the likelihood, as the logarithm is a monotone function, but is easier to handle mathematically:

$$\log(\mathcal{L}(X|\mu, \sigma^2)) = n \log(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \ .$$

To find the maxima, we calculate the derivate with respect to μ and, set it to 0 and solve for μ:

$$\begin{aligned}\frac{\partial}{\partial \mu} \log(\mathcal{L}(X|\mu, \sigma^2)) &= 0 \\ \sum_{i=1}^{n} \frac{(x_i - \mu)}{\sigma^2} &= 0 \\ \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i\end{aligned}$$

Here, we found the maximum of the log-likelihood (since $\frac{\partial^2}{\partial \mu^2} \log(\mathcal{L}(X|\mu, \sigma^2)) = -n/\sigma^2 < 0$) at the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$ as expected by intuition. As similar calculation for $\hat{\sigma}^2$ yields

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \ ,$$

which is the sample variance.

Although in this example, the maximum likelihood estimators of the parameters can be calculated in closed form, this is not possible in general. Then, the maximum likelihood estimate has to be obtained by numerically minimizing the log-likelihood function via optimization algorithms (e.g. gradient descend).

In a Bayesian approach (as reviewed by Bishop, 2006), one is not only interested in a point estimate, such as the maximum likelihood estimate, but in the full posterior distribution $P(\theta|X)$ of the parameters θ, which is obtained using Bayes' rule:

$$P(\theta|X) = \frac{\mathcal{L}(X|\theta)\pi(\theta)}{\int d\theta \mathcal{L}(X|\theta)\pi(\theta)} \tag{2.12}$$

Here, $\pi(\theta)$ is a prior distribution over parameters, which can be used to include prior knowledge about the parameters (e.g. from previous experiments). Except for the simplest models, the posterior distribution is analytically intractable and Markov Chain Monte Carlo methods are used to draw samples from the posterior instead (Metropolis et al., 1953). In analogy to the maximum likelihood estimate, one can use the mode, mean or median of the posterior as a point estimate of the parameters.

Parameter uncertainty

However, apart from point estimates, one is also interested in their uncertainty, i.e. how much one can trust the estimate. This uncertainty should then be propagated into e.g. model predictions.

In the frequentist approach it is assumed that the parameters θ of the model are unknown but fixed, while the observed data X is considered random sample, and one reports the maximum likelihood estimate $\hat{\theta}$ as the point estimate of θ. However, this point estimate will vary depending on the particular sample of data X: If we observe another dataset X' (generated with the same unknown parameters), the estimate $\hat{\theta}$ will now be different. To account for this uncertainty, one constructs confidence intervals around the maximum likelihood estimate. These intervals are constructed from the sampling distribution of $\hat{\theta}$, i.e. the distribution of $\hat{\theta}$ across an infinite number of future datasets. The sampling distribution can be derived analytically in some cases (e.g. the t-distribution for the estimator of the mean of a Gaussian) or has to be obtained via bootstrap. The interpretation of a 95% confidence intervals is as follows: If we happen to obtain new datasets (but from the same underlying model and the same parameters) and we calculate the maximum likelihood for each dataset separately, in 95 % of cases the new estimate will be contained in the confidence interval.

In the Bayesian approach, quantifying the uncertainty of the parameters θ is much more intuitive. Having obtained the posterior $p(\theta|X)$, one can report the posterior mode as a point estimate (analogous to the frequentist $\hat{\theta}$), and can quantify the uncertainty in the parameters by calculating credibility intervals (regions) that measure the width of the posterior distribution, where a small width indicates low uncertainty in θ. Often the credibility intervals are constructed such that 95% of the posterior mass is contained within and that equal mass is located in each tail (central intervals). Interpretation of credibility intervals is simple: They represent our degree of belief given the data X that the parameter lies within the constructed region.

Despite their similar usage in quantifying the uncertainty of a parameter, frequentist confidence intervals and Bayesian credibility intervals are different conceptually and also numerically[5].

2.3.2 Approximate Bayesian Computation

Unfortunately, the exact likelihood function is intractable for most stochastic systems, as it requires the solution of the underlying CME. Instead, one can approximate the system of interest, such that the approximate system has a tractable likelihood function. Examples include moment equations (Lee et al., 2009; Engblom, 2006; Hasenauer et al., 2013; Zechner et al., 2012), the Linear Noise Approximation (Elf and Ehrenberg, 2003; Komorowski et al., 2009; Van Kampen, 1992) and diffusion approximations (Fuchs, 2013; Golightly and Wilkinson, 2005). However, these approximations are not valid in general (e.g. for multi-stable models) or require additional ad hoc assumptions (e.g. moment closure). Hence, exact likelihood based inference is not applicable for many stochastic models.

However, as we can easily generate samples from the CME using Gillespie's algorithm, we can perform simulation based (likelihood free) inference using Approximate Bayesian Computation (ABC, Marjoram et al., 2003; Sisson et al., 2007; Toni et al., 2009), which casts likelihood-free inference into a Bayesian framework. In ABC, the evaluation of the likelihood is replaced by forward simulation of the model and a comparison of simulated and observed data via a distance function $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_0^+$, where $\mathcal{D}$ is the domain of the data. Instead of optimizing the analytical likelihood, one instead minimizes the distance between observed and simulated data with respect to parameters.

ABC rejection sampling

A simple algorithm reminiscent of rejection sampling (Pritchard et al., 1999) can be formulated (Algorithm 2), which generates a sample $\theta^* \sim \pi(\theta)$ from the prior distribution $\pi(\theta)$, simulates data $x^* \in \mathcal{D}$ from this parameter, and accepts sample θ^* based on a distance function d between x^* and the observed data $x_0 \in \mathcal{D}$ with a given threshold ϵ. Thereby, we obtain an approximation of the posterior distribution $\pi(\theta|d(x_0, x^*) < \epsilon)$. The smaller ϵ the better the approximation to the true posterior. However, this procedure is ineffective if the prior distribution and the posterior are very different as most samples will be rejected.

ABC sequential Monte Carlo sampling

Therefore, one typically performs a sequential Monte Carlo (particle filtering) variant of ABC (Toni et al., 2009). Sequential Monte Carlo applies multiple rounds of rejection sampling with decreasing tolerance levels $\epsilon_1, \ldots, \epsilon_M$ and proper reweighing of parameters. The posterior from the last iteration is used as a prior for the next iteration, after applying small perturbations via a kernel K (Algorithm 3). The algorithm yields a sample from the approximate posterior $\pi(\theta|d(x_0, x^*) < \epsilon_M)$, but is more efficient than rejection sampling

[5]However, a notable exception is the estimation of the mean of a Gaussian. Here, both sampling distribution and posterior are identical t-distributions, hence giving the same results for confidence and credibility intervals.

Algorithm 2: ABC rejection sampling

Input: Prior $\pi(\theta)$, observed data x_o, tolerance level ϵ, distance function d, number of samples N

Output: Sample θ from the approximate Posterior $\pi(\theta|d(x_0, x^*) < \epsilon)$

```
θ = ∅ ;                                          // set of accepted parameters
i = 0;
while i < N do
    θ* ~ π(θ) ;                                  // sample from prior
    Simulate data x* from θ*;
    d* = d(x*, x_o);
    if d* ≤ ε then
        θ = θ ∪ θ* ;                             // Accept θ*
        i = i + 1;
    else
        Reject θ*;
    end
end
```

due to its lower rejection rate. For a detailed derivation and analysis of the algorithm, see Toni et al. (2009).

To achieve optimal performance several parameters of the algorithm need to be tuned: First, the number of accepted particles per iterations has to be set so large that the accepted particles give a good approximation of the distribution, but on the other hand, so small that it is computationally still tractable. Typical choices are in the order of 10^3 (Toni et al., 2009). Second, a perturbation kernel K has to be selected. Here, usually a uniform or Gaussian kernel is selected whose width decreases with decreasing tolerance level (Lillacci and Khammash, 2013). Last, the number of iterations and the associated tolerance schedule $[\epsilon_1, \ldots, \epsilon_M]$ has to be defined. Optimally, one chooses a schedule where the distances between consecutive distributions θ_m is minimal (high acceptance rates) but at the same time keeping the number of iterations to a minimum (computational efficiency). Several methods of automatically determining the tolerance schedule have been proposed recently (Beaumont et al., 2009; Silk et al., 2012; Moral et al., 2011).

Note that for a stochastic model, it is not enough to create a single simulation from a parameter set which is compared to the observed data. Due to the inherent randomness the distance between a single realization of the model and the observed data can be large even if the true parameters were used. Hence, one has to simulate many realizations for given parameters, making the method computationally challenging.

Distance function

For applications, it is most crucial to design an appropriate distance function d that determines if the simulated and observed data are similar. Often it is difficult (and ineffective due to the curse of dimensionality) to design a distance function between the complete datasets x_0 and x^*. Therefore, the distance function is typically computed on summary

Algorithm 3: ABC sequential Monte Carlo sampling

Input: Prior $\pi(\theta)$, observed data x_o, tolerance levels $\epsilon_1, \ldots, \epsilon_M$, distance function d, number of samples N, kernel K

Output: Sample θ_M from the approximate Posterior $\pi(\theta|d(x_0, x^*) < \epsilon_M)$

$m = 0$;

while $m \leq M$ **do**

  $n = 0$;

  $\theta_m = \emptyset$; `// set of accepted parameters`

  while $n \leq N$; `// iterate until N particles accepted`

  do

    if $m == 0$ **then**

      $\theta^{**} \sim \pi(\theta)$; `// sample from prior`

    else

      Sample θ^* from population θ_{m-1} with weights w_{m-1};

      $\theta^{**} \sim K(\theta^{**}|\theta^*)$; `// perturb` θ^* `with kernel K`

    end

    Simulate data x^{**} from θ^{**};

    $d^{**} = d(x^{**}, x_o)$;

    if $d^{**} \leq \epsilon_m$ **then**

      $\theta_m = \theta_m \cup \theta^{**}$; `// Accept` θ^{**}

      $w_m^{(n)} = \begin{cases} 1 & m = 0 \\ \frac{\pi(\theta^{**})}{\sum_j w_{m-1}^{(j)} K(\theta^{**}|\theta_{m-1}^{(j)})} & m > 0 \end{cases}$; `// Calculate weight for` θ^{**}

      $n = n + 1$;

    end

  end

  $m = m + 1$;

end

statistics of the datasets, e.g. $d(S(x_0), S(x^*))$, where $S(x) \in \mathbb{R}^k$ is a vector of k summary statistics of the data x. Hence, one has to carefully choose these summary statistics to ensure that they capture the relevant features of the system. Usually one chooses a distance function based on the L_p-norm ($p \geq 1$)

$$d_{L_p}(S(x_0), S(x^*)) = \|S(x_0) - S(x^*)\|_p = \left(\sum_{i=1}^{k} |S_i(x_0) - S_i(x^*)|^p \right)^{1/p},$$

with e.g. $p = 2$ ("Euclidean distance") or $p = 1$ ("Manhattan distance"). However, it has been shown that the infinity norm $L_\infty(x) = \max(|x_1|, \ldots, |x_k|)$ is beneficial under certain circumstances as it allows to derive an upper bound on the number of simulations per parameter required (Lillacci and Khammash, 2013).

2.4 Graphical models

In this section, we will briefly introduce graphical models as a concept to represent the joint probability distribution of various variables, e.g. observed quantities, hidden variables and unknown parameter. We also discuss inference in tree-structured graphical models via the sum-product algorithm, which is applied in context of cellular genealogies in chapter 5. For a comprehensive treatment of graphical models, we refer to Murphy (2012) or Bishop (2006).

Consider a set of variables $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ (sets of variables are denoted in bold). Generally, any joint probability distribution $p(\boldsymbol{X})$ of the n variables can be decomposed into

$$\begin{aligned} p(\boldsymbol{X}) &= p(X_1, \ldots, X_n) \\ &= p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_2, X_1) \cdot \ \ldots \ \cdot p(X_n|X_1, \ldots, X_{n-1}) \ . \end{aligned} \tag{2.13}$$

Note that the factorization is not unique, because one can simply relabel the variables. This representation of the joint distribution is however not very useful, as every variable depends on all the previous variables.

2.4.1 Conditional independence

The key idea to represent this joint density more compactly is to make some problem-specific conditional independence assumptions. Two variables X, Y are conditionally independent given a third variable Z, written as $(X \perp Y)|Z$ if and only if

$$p(X, Y|Z) = p(X|Z) \cdot p(Y|Z) \ ,$$

or equivalently

$$p(X|Y, Z) = p(X|Z) \ .$$

These conditional independence assumptions then simplify the right hand side of Eq. (2.13). For example, consider a Markov chain (X_1, X_2, X_3, X_4) of length four, i.e. we assumed $(X_{i-1} \perp X_{i+1})|X_i$ such that "past and future are independent given the present". Instead of the general factorization in Eq. (2.13), we obtain the much simpler expression

$$p(X_1, X_2, X_3, X_4) = p(X_4|X_3) \cdot p(X_3|X_2) \cdot p(X_2|X_1) \cdot p(X_1) \ ,$$

where every variable is just conditioned on its immediate predecessor.

2.4.2 Directed graphical models

Any given factorization of the joint probability density can be visualized as a directed acyclic graph $G = (\boldsymbol{V}, \boldsymbol{E})$ where the set of nodes $\boldsymbol{V} = \{X_1, \ldots, X_n\}$ are the variables in joint distribution and there exists a directed edge (X_i, X_j) from X_i to X_j if any term of the factorization has X_j conditioned on X_i. These models are termed directed graphical models, Bayesian networks or belief networks.

In case of the general factorization of Eq. (2.13), the resulting graph will be fully connected. By applying conditional independence assumptions, some edges will be removed

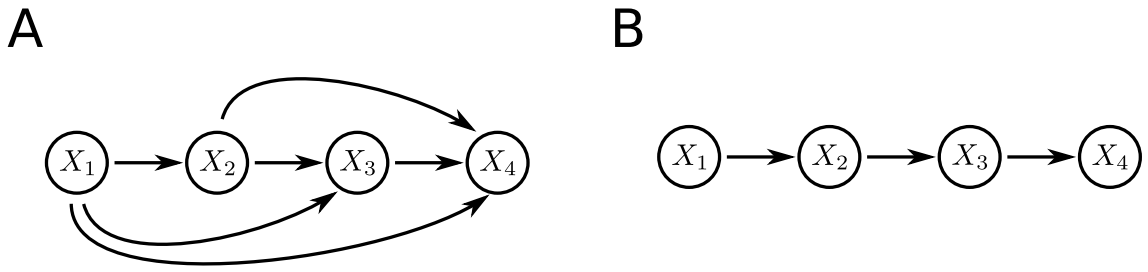


Figure 2.1: **Directed graphical models.** A) A directed graphical model representing the general factorization of the joint distribution $p(X_1, X_2, X_3, X_4) = p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_2, X_1) \cdot p(X_4|X_3, X_2, X_1)$ of Eq. (2.13). The graph is fully connected. B) Applying particular conditional independence assumptions (in this case the Markov property $(X_{i-1} \perp X_{i+1})|X_i$) simplifies the factorization into $p(X_1, X_2, X_3, X_4) = p(X_4|X_3) \cdot p(X_3|X_2) \cdot p(X_2|X_1) \cdot p(X_1)$ and yields a simpler graphical model.

from the graph (see Fig. 2.1). For a given directed graphical model $G = (\boldsymbol{V}, \boldsymbol{E})$ we can read off the factorization that it encodes via

$$p(X_1, \ldots, X_n) = \prod_i^n P(X_i|\text{pa}(X_i)) \ , \tag{2.14}$$

where $\text{pa}(X_i) = \{X_j \in \boldsymbol{V}|(X_j, X_i) \in \boldsymbol{E}\}$ denotes the parents of node X_i. To derive the conditional independence assumptions a directed graphical model encodes, the concept of d-separation is used (Geiger et al., 1990).

2.4.3 Inference

Given a graphical model $G = (\boldsymbol{V}, \boldsymbol{E})$ and observations of a subset of nodes $\boldsymbol{X}_v \subset \boldsymbol{V}$, one can now perform inference on G, that is, estimate the hidden variables $\boldsymbol{X}_h = \boldsymbol{V} \setminus \boldsymbol{X}_v$ given the observed variables $\boldsymbol{X}_v$[6]. The task is to calculate the posterior of the hidden variables via Bayes' theorem:

$$p(\boldsymbol{X}_h|\boldsymbol{X}_v) = \frac{p(\boldsymbol{X}_v|\boldsymbol{X}_h)\pi(\boldsymbol{X}_h)}{p(\boldsymbol{X}_v)} = \frac{p(\boldsymbol{X}_v|\boldsymbol{X}_h)\pi(\boldsymbol{X}_h)}{\sum_{\boldsymbol{X}'_h} p(\boldsymbol{X}_v|\boldsymbol{X}'_h)\pi(\boldsymbol{X}'_h)} \ .$$

Here, we will introduce the "sum-product" algorithm (also known as belief propagation) which performs exact inference on trees, i.e. connected graphs where there exists only a single path between any two nodes. For more general graphs, the "junction tree" algorithm is an alternative for exact inference, but will not be discussed here (see e.g. Murphy, 2012).

[6]Unknown parameters can simply be included into $\boldsymbol{X}_h$ and inferred from the observed data together with the hidden variables.

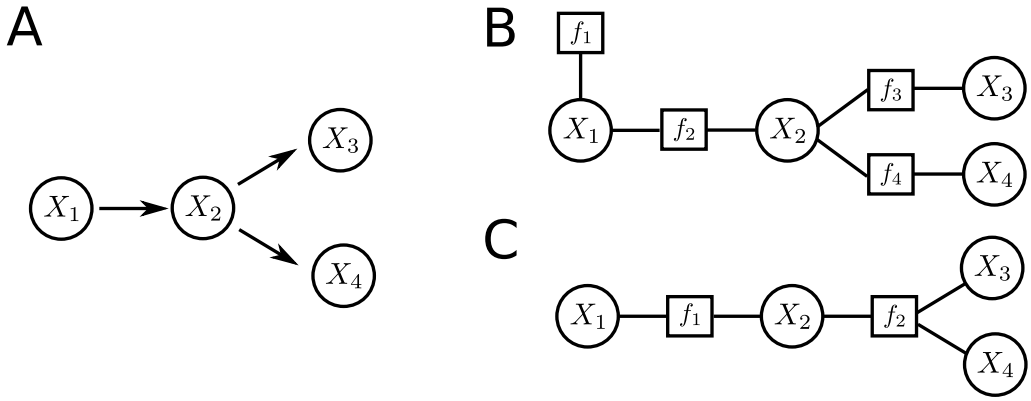


Figure 2.2: **Factor graph representation.** A) A directed graphical model representing the joint distribution $p(X_1, X_2, X_3, X_4) = p(X_1) \cdot p(X_2|X_1) \cdot p(X_3|X_2) \cdot p(X_4|X_2)$. B) Factor graph representing the same joint distribution as the model in A), with $f_1(X_1) = p(X_1)$, $f_2(X_1, X_2) = p(X_2|X_1)$, $f_3(X_2, X_3) = p(X_3|X_2)$, and $f_4(X_3, X_4) = p(X_4|X_2)$. C) Another factor graph for the model in A) is less explicit about the factorization, with $f_1(X_1) = p(X_1) \cdot p(X_2|X_1)$ and $f_2(X_2, X_3, X_4) = p(X_3|X_2) \cdot p(X_4|X_2)$.

Factor graphs

To derive the sum-product algorithm in a simple form, we first introduce the idea of a factor graph. Suppose we decompose the joint distribution of over the variables $\boldsymbol{X} = \{X_1, \dots X_n\}$ into a product of factors f_s, such that

$$p(\boldsymbol{X}) = p(X_1, \dots, X_n) = \prod_s f_s(\boldsymbol{X}_s) \,, \tag{2.15}$$

where each $\boldsymbol{X}_s \subset \boldsymbol{X}$ is a subset of variables (the sets can overlap). Comparing to Eq. (2.14), we see that the factors f_s correspond to the conditional probability distributions $P(X_i|\text{pa}(X_i))$ in directed graphical models.

Eq. (2.15) can be represented as a bipartite graph, which is a graph with two kinds of nodes and edges exist only between nodes of different kind. In a factor graph, one set of nodes represents the variables $\boldsymbol{X}$, the other set of nodes represents the factors f_s. Undirected edges link a factor $f_s(\boldsymbol{X}_s)$ to its associated variables $\boldsymbol{X}_s$. One can now convert a directed graphical model into a factor graph (see Fig. 2.2). However, multiple factor graphs correspond to the same directed graphical model, depending how explicit we are about the factorization (see Fig. 2.2B,C). Note that any tree-structured directed graphical model will result in a tree-structured factor graph (Bishop, 2006).

The sum-product algorithm

To goal of the sum-product algorithm is to calculate marginal distributions $p(X_i)$ of a single variable X_i in the model. Note that from now on, we consider only discrete variables X_i for simplicity. Naively, one can invoke the definition of the marginal distribution,

$$p(X_i) = \sum_{X_1} \dots \sum_{X_{i-1}} \sum_{X_{i+1}} \dots \sum_{X_n} p(X_1, \dots, X_n) \,, \tag{2.16}$$

i.e. summing the joint distribution over all other variables. However, the amount of operations required to evaluate the right hand side scales exponentially with the number of nodes. If the graph has n nodes, each having K possible discrete states, the memory required to store the joint distribution is K^n. The key idea is to substitute the factorization of the joint distribution from Eq. (2.15) into Eq. (2.16),

$$p(X_i) = \sum_{X_1} \dots \sum_{X_{i-1}} \sum_{X_{i+1}} \dots \sum_{X_n} \prod_s f_s(\boldsymbol{X}_s) \;, \tag{2.17}$$

and to rearrange summations and multiplications, since some factors are independent of certain summation variables. This leads to a more efficient calculation as sums are performed locally, operating only on functions f_s of a subset of variables $\boldsymbol{X}_s$. While this algebraic manipulations could be done manually in principle, one can formulate this calculation more abstractly on the factor graph in terms of message passing, i.e. local messages being sent between nodes. We refer the interested reader to Bishop (2006) for a derivation and here present only the resulting algorithm.

The main result is that the calculation of the marginal distribution of a variable X_i can be expressed in terms of two types of messages: messages being sent from variable nodes to factor nodes and messages being sent from factor nodes to variable nodes.

First, messages originating at an unobserved variable node $X_i \in \boldsymbol{X}_h$ and being sent to an adjacent factor node f_s are defined as

$$\mu_{X_i \to f_s}(X_i) = \prod_{g \in \mathrm{ne}(X_i) \setminus f_s} \mu_{g \to X_i}(X_i) \;, \tag{2.18}$$

where $\mathrm{ne}(X_i)$ refers to all factor nodes connected to X_i. The message of sent from X_i to f_s is just the product of messages $\mu_{g \to X_i}$ received from all neighboring factors g except f_s. If X_i is observed ($X_i \in \boldsymbol{X}_v$) and has value a the message is

$$\mu_{X_i \to f_s}(X_i) = \delta_{X_i, a} \prod_{g \in \mathrm{ne}(X_i) \setminus f_s} \mu_{g \to X_i}(X_i) \;, \tag{2.19}$$

where we "clamp" the observed variable to its value via the Kronecker-δ. Second, a message from a factor node f_s to a variable node X_i is defined as

$$\mu_{f_s \to X_i}(X_i) = \sum_{X'_1} \dots \sum_{X'_m} f_s(\boldsymbol{X}_s) \prod_{Y \in \boldsymbol{X}_s \setminus X_i} \mu_{Y \to f_s}(Y) \;, \tag{2.20}$$

where we have labeled the neighbors of node f_s as $\boldsymbol{X}_s = \{X'_1, \dots, X'_m, X_i\}$. The message sent from factor f_s to variable X_i is calculated by multiplying incoming messages, and marginalizing this product together with f_s over all associated variables except X_i.

Consider two special cases of Eqs. (2.18) and (2.20): if a variable node X_i is a leaf of the factor graph, the message sent from this node via its only edge is $\mu_{X_i \to f}(X_i) = 1$. If a factor node is a leaf, the message sent by this factor to its variable node is $\mu_{f \to X_i} = f(X_i)$.

With these definitions the calculation of the marginal distribution of variable X_i reduces to a multiplication of incoming messages at node X_i in the factor graph

$$p(X_i) = \prod_{f \in \mathrm{ne}(X_i)} \mu_{f \to X_i}(X_i) \;, \tag{2.21}$$

which is the main result of the sum-product algorithm. Note that in the presence of observed variables, the algorithm calculates the joint distribution $p(X_i, \boldsymbol{X}_v)$ between a single unobserved variable $X_i \in \boldsymbol{X}_h$ and the set of observed variables $\boldsymbol{X}_v$) and we obtain the evidence of the observed data as $p(\boldsymbol{X}_v) = \sum_{X_i} p(X_i, \boldsymbol{X}_v)$. In order to calculate the marginal distribution $p(X_i)$, we proceeds as follows: (i) Consider the node X_i as the root of the factor graph tree. (ii) Find all leave nodes, which are either factors or variables. (iii) Propagate messages from the leaves up towards the roots. Nodes can send outgoing messages once they received incoming messages from all their children in the tree. (iv) When the root has received all incoming messages from its children, evaluate Eq. (2.21) to obtain the marginal distribution.

Let us illustrate the procedure with an example shown in Fig. 2.3. We want to calculate the marginal distribution of X_1 and have observed $X_4 = a$. We root the factor graph in X_1 and start propagating messages from the leaves X_4, X_5.

$$\begin{aligned} \mu_{X_4 \to f_2}(X_4) &= \delta_{X_4,a} \\ \mu_{X_5 \to f_3}(X_5) &= 1 \ , \end{aligned}$$

where we clamped the value of the observed variable via the Kronecker-δ and used the definition of a message sent by leaf variables. Next, we propagate a message from f_2 to X_2 according to Eq. (2.20)

$$\begin{aligned} \mu_{f_2 \to X_2}(X_2) &= \sum_{X_4} f_2(X_2, X_4) \mu_{X_4 \to f_2}(X_4) \\ &= \sum_{X_4} f_2(X_2, X_4) \delta_{X_4,a} \\ &= f_2(X_2, a) \ , \end{aligned}$$

where we collapsed the sum over X_4 due to the clamping. Similarly for $\mu_{f_3 \to X_3}(X_3)$ we have

$$\begin{aligned} \mu_{f_3 \to X_3}(X_3) &= \sum_{X_5} f_3(X_3, X_5) \mu_{X_5 \to f_3}(X_5) \\ &= \sum_{X_5} f_3(X_3, X_5) \ . \end{aligned}$$

Sending messages towards f_1 is simple as variable nodes with only two factors associated just pass through incoming messages according to Eq. (2.18):

$$\begin{aligned} \mu_{X_2 \to f_1}(X_2) &= \mu_{f_2 \to X_2}(X_2) \\ &= f_2(X_2, a) \\ \mu_{X_3 \to f_1}(X_3) &= \mu_{f_3 \to X_3}(X_3) \\ &= \sum_{X_5} f_3(X_3, X_5) \ . \end{aligned}$$

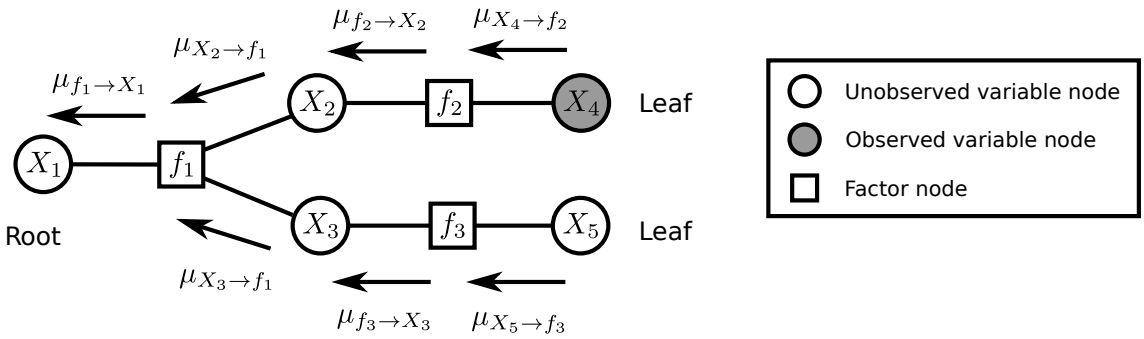


Figure 2.3: **Message passing on a factor graph.** To calculate the marginal distribution $p(X_1)$ via the sum-product algorithm, the tree-structured factor graph is rooted in X_1 and messages are passed from the leaf nodes X_4 (observed) and X_5 along the factor graph until they reach the root node X_1. Once the root has received messages from all its children (in this case only f_1) the marginal distribution is calculated via Eq. (2.21).

These messages now converge on f_1 which sends to X_1 the message (Eq. 2.20)

$$\begin{aligned}
\mu_{f_1 \to X_1}(X_1) &= \sum_{X_2} \sum_{X_3} f(X_1, X_2, X_3) \cdot \mu_{X_2 \to f_1}(X_2) \cdot \mu_{X_3 \to f_1}(X_3) \\
&= \sum_{X_2} \sum_{X_3} f(X_1, X_2, X_3) \cdot f_2(X_2, a) \cdot \mu_{X_3 \to f_1}(X_3) \\
&= \sum_{X_2} \sum_{X_3} f(X_1, X_2, X_3) \cdot f_2(X_2, a) \cdot \left[\sum_{X_5} f_3(X_3, X_5) \right] .
\end{aligned}$$

As X_1 only has one neighboring node (f_1) we find via Eq. (2.21) that the marginal evaluates to

$$\begin{aligned}
p(X_1) &= \mu_{f_1 \to X_1}(X_1) \\
&= \sum_{X_2} \sum_{X_3} f(X_1, X_2, X_3) \cdot f_2(X_2, a) \cdot \left[\sum_{X_5} f_3(X_3, X_5) \right] ,
\end{aligned}$$

where we see how the algorithm pushed the sum over X_5 into the product over factors and clamped X_4 to its observed value.

If we wanted to calculate the marginal distribution of another node X_j, we could repeat the above procedure. However, a more efficient way to calculate all marginal distributions at once is the following: Choosing an arbitrary node as the root, propagate messages from leaves to root as before. Once all messages have arrived at the root, send out messages from the root down towards the leaves. After those two passes, every node will have received messages from all its neighbors, and one can efficiently evaluate the marginal distribution of every node in terms of messages via Eq. (2.21). We only have to store the messages that were generated during the passes. Overall, we have to compute $2 \cdot m$ messages with m the number of edges in the graph, as opposed to $m \cdot n$ when running the sum-product algorithm for each of the n nodes individually.

Variants of the sum-product algorithm

Although only the sum-product algorithm for discrete variables is used in this thesis, for the sake of completeness, we mention some variants of the sum product algorithm.

A common variant of the sum-product algorithm is the max-product algorithm (for details, see Bishop, 2006), which allows to calculate the most probable configuration of the graphical model, or equivalently the mode of the joint distribution $p(\boldsymbol{X})$. Here, one simply replaced the sums within the sum-product algorithm with max()-operations. A numerically stable version replaces probabilities by log-probabilities and is called the max-sum algorithm. When one departs from discrete to continues variables, summations become integrals, and generally, no exact algorithms exist for inference, since most integrals cannot be solved analytically. An important exception are directed Gaussian graphical models[7]. While the sum-product and max-product algorithms are exact for tree-structured models, for general graphs containing cycles, different strategies have to be used. Variable elimination and the junction tree algorithm (for details, see Murphy, 2012) provide exact inference, however the computational complexity of both is exponential in the number of nodes in the worst case. Instead one can resort approximate algorithms, e.g. loopy belief propagation (standard sum-product applied to non tree-structured graphs), variational mean field, or Gibbs sampling (all discussed in e.g. Murphy, 2012).

[7] Each single variable has a Gaussian distribution whose mean is a linear combination of the parent nodes. Consequently, also the joint distribution is Gaussian.

Chapter 3

Mechanistic models of binary cell fate choice: genetic toggle switches

As discussed in chapter 1, toggle switches serve as current paradigm of how binary fate decisions are implemented molecularly. In this chapter, we study the dynamics of a stochastic two stage toggle switch model, which explicitly accounts for mRNA synthesis and degradation. We find that, contrary to the expectation from a deterministic description, this switch shows complex multi-attractor dynamics even without autoactivation and cooperativity. Other stochastic models of toggle switches studied so far (Lipshtat et al., 2006; Kepler and Elston, 2001; Schultz et al., 2008; Warren and ten Wolde, 2004) focused on a one stage model of gene expression without explicitly considering mRNA as an intermediate stage. Here, we discover that when accounting for mRNA the toggle switch shows novel attractors which can be identified with committed and primed states in cell differentiation. Notably, we present a system with high protein abundance that nevertheless requires a probabilistic description to exhibit multistability and complex switching dynamics. In the second part of this chapter, we show how a toggle switch model can account for observed differentiation dynamics in granulocyte/monocyte progenitors by fitting the model using Approximate Bayesian Computation. Here, the fitted model predicts different timescales in the dynamics of granulocyte and monocyte differentiation.

Methods, results and figures of this chapter are based on Strasser et al. (2012) and Marr et al. (2012).

3.1 Dynamics of a genetic toggle switch based on a two-stage model of gene expression

Probabilistic models of the toggle switch account for low copy numbers and intrinsic fluctuations. Kepler and Elston (2001) discussed the dynamics of an exclusive switch, where two genes share the same promoter within a probabilistic framework. A comparison of simple switch circuitries is given by Warren and ten Wolde (2004). Contrary to deterministic models, transitions between the two macroscopic regimes where one of the two genes dominates are possible due to the inherently noisy gene transcription (Schultz et al., 2008; Walczak et al., 2005a), even without cooperative binding of transcription factors (Lipshtat

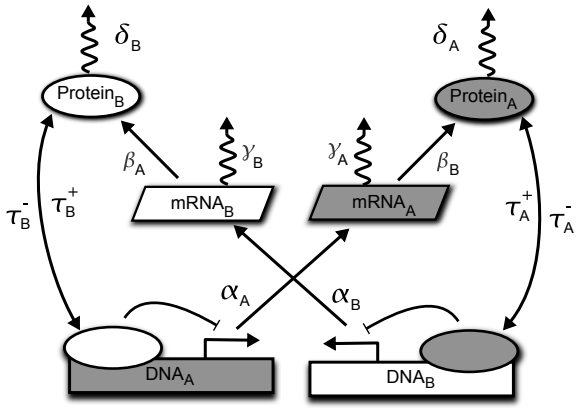


Figure 3.1: **Scheme of the two-stage switch.** Species associated with gene A are shown in gray, species associated with B are shown in white. Solid arrows indicate synthesis and binding, jagged arrows indicate degradation. $\mathrm{mRNA_A}$ is transcribed from $\mathrm{DNA_A}$ with rate α_A. It decays with rate γ_A and is translated into $\mathrm{Protein_A}$ with rate β_A. $\mathrm{Protein_A}$ decays with rate δ_A and can bind (unbind) $\mathrm{DNA_B}$ with rate τ_A^+ (τ_A^-). Protein-bound DNA leads to transcriptional arrest. The topology is symmetric with respect to the genes A and B, thus, the same reactions exist for B.

et al., 2006). More recent contributions focused on analytic descriptions (Walczak et al., 2005b; Schultz et al., 2007), the switching time between macroscopic regimes for different regulatory realizations (Loinger et al., 2007; Barzel and Biham, 2008; Schultz et al., 2008) or parameter regimes (Walczak et al., 2005a), boundaries for the switching time (Bialek, 2001), or delay effects (Zhu et al., 2007). Notably, all of these approaches are based on a one-stage model of gene expression, where DNA is directly processed into functional proteins. However, it has been shown that the characteristics of protein noise strongly depend on the underlying expression model (Thattai and van Oudenaarden, 2001; Shahrezaei and Swain, 2008).

In this section, we abstract the regulatory details of the prominent myeloid PU.1/Gata1 mutual inhibition. Contrary to common belief, which advocates the lumping of the two stages of expression, we show that the inclusion of both mRNA and protein leads to an interesting change in system dynamics. The probabilistic two-stage description exhibits complex multi-attractor dynamics without autoactivation and cooperativity.

Remarkably, a recent study reported low numbers of mRNAs in single murine blood cells: Warren et al. (2006) found around 10 transcripts of the PU.1 gene per cell in common myeloid progenitors. Furthermore, Schwanhäusser et al. (2011) measured mRNA abundances for 5000 genes in mouse fibroblasts and showed that the median number of mRNAs per gene per cell is 17. Filtering these 5000 genes for transcription factors (Gene ontology ID GO:0006355, Ashburner et al., 2000) results in 722 candidate genes and a median of 17 mRNAs per transcription factor per cell.

Based on these findings we study a probabilistic description of a toggle switch with low mRNA numbers, high protein abundance and in accordance with the known role of PU.1, monomeric transcription factor binding. We deliberately choose the simplest toggle switch model and neglect autoactivation due to our ignorance of the logic of activation and

inhibition at the promoter. However our results can easily be extended and are discussed for the case of dimeric regulation and exclusive autoactivation.

3.1.1 A toggle switch based on a two-stage model of gene expression

We describe the mutual inhibition of two genes, further on called A and B, using a two-stage model of gene expression (Thattai and van Oudenaarden, 2001; Shahrezaei and Swain, 2008) with mutual inhibition being realized as DNA-protein binding (see Fig. 3.1). This kind of switch has been implemented in vivo by Gardner et al. (2000). The model can be represented as a set of biochemical reactions for A and B, respectively, and a set of reaction rates α, β, etc. (see chapter 2):

$$\mathrm{DNA_A} \xrightarrow{\alpha_\mathrm{A}} \mathrm{DNA_A} + \mathrm{mRNA_A} \qquad \mathrm{DNA_B} \xrightarrow{\alpha_\mathrm{B}} \mathrm{DNA_B} + \mathrm{mRNA_B} \tag{3.1}$$

$$\mathrm{mRNA_A} \xrightarrow{\gamma_\mathrm{A}} \emptyset \qquad \mathrm{mRNA_B} \xrightarrow{\gamma_\mathrm{B}} \emptyset \tag{3.2}$$

$$\mathrm{mRNA_A} \xrightarrow{\beta_\mathrm{A}} \mathrm{mRNA_A} + \mathrm{Protein_A} \qquad \mathrm{mRNA_B} \xrightarrow{\beta_\mathrm{B}} \mathrm{mRNA_B} + \mathrm{Protein_B} \tag{3.3}$$

$$\mathrm{Protein_A} \xrightarrow{\delta_\mathrm{A}} \emptyset \qquad \mathrm{Protein_B} \xrightarrow{\delta_\mathrm{B}} \emptyset \tag{3.4}$$

$$\mathrm{Protein_A} + \mathrm{DNA_B} \xrightarrow{\tau_\mathrm{A}^+} \mathrm{DNA_B^{bound}} \qquad \mathrm{Protein_B} + \mathrm{DNA_A} \xrightarrow{\tau_\mathrm{B}^+} \mathrm{DNA_A^{bound}} \tag{3.5}$$

$$\mathrm{DNA_B^{bound}} \xrightarrow{\tau_\mathrm{A}^-} \mathrm{Protein_A} + \mathrm{DNA_B} \qquad \mathrm{DNA_A^{bound}} \xrightarrow{\tau_\mathrm{B}^-} \mathrm{Protein_B} + \mathrm{DNA_A} \tag{3.6}$$

Reactions (3.1) and (3.2) correspond to mRNA transcription from an unbound promoter and mRNA degradation, respectively. Reactions (3.3) and (3.4) resemble protein translation and degradation. The reactions (3.5) and (3.6) describe the binding and unbinding of a protein to the antagonistic gene and thereby the transition from an active to an inactive promoter and vice versa. Bound DNA lacks the ability to be transcribed. These two reactions subsume a more intricate mechanism of transcription-factor-DNA interaction (Gerland et al., 2002). Note that we assume monomeric transcription factor binding as the simplest of regulatory interaction (which can induce bimodal gene expression, Lipshtat et al., 2006). Our system's topology is symmetric with regard to the two genes, and so are the two columns of reactions (3.1)–(3.6) upon the exchange of gene labels A and B.

This model of gene expression is a highly simplified abstraction of the complex processes in the cell. Condensing transcription into a single biochemical reaction does not account for the various steps required to transcribe a gene, e.g. the assembly of the transcription initiation complex, unwinding of DNA or transition of the polymerase to elongation phase. Postprocessing and transport mechanisms are also neglected. However, simplified models of gene expression have successfully been applied to experimental data, supporting the validity of these simplifications (Harper et al., 2011; Raj et al., 2006; Huang et al., 2007).

3.1.2 Deterministic model

Most commonly one will study the properties of the system in a deterministic framework using ordinary differential equations (ODEs) that describe the time-evolution of species

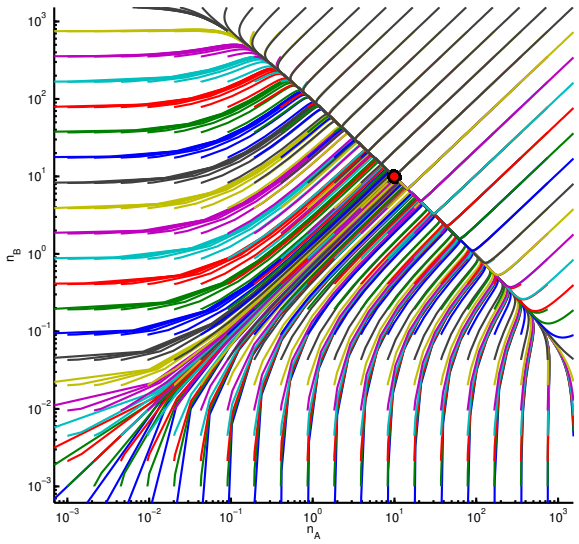


Figure 3.2: **Phase portrait of the deterministic two-stage toggle switch model projected onto the N_A and N_B dimensions.** The same parameters were used as in the probabilistic model of Fig. 3.3 and Fig. 3.4. The ODE was solved for different initial protein concentrations n_A, n_B and the mRNA concentrations where set to $m_A = 0.01 \cdot n_A$ and $m_B = 0.01 \cdot n_B$, resembling the fact that for the given parameters each mRNA will correspond to approximately 100 proteins. Trajectories are arbitrarily colored to improve readability. Note that trajectories can intersect as this in only a projection of the full system. Inspection of the phase portrait reveals the single steady state (red dot) as predicted by Eqs. (3.19)–(3.20). No limit cycles are visible in the phase portrait.

concentrations (Roeder and Glauche, 2006; Huang et al., 2007; Chickarmane et al., 2009; Cherry and Adler, 2000). The ODEs can directly be inferred from reactions (3.1)–(3.6) assuming mass action kinetics:

$$\frac{d}{dt}d_A = \tau_B^- \cdot (1 - d_A) - \tau_B^+ \cdot d_A \cdot n_B \tag{3.7}$$

$$\frac{d}{dt}d_B = \tau_A^- \cdot (1 - d_B) - \tau_A^+ \cdot d_B \cdot n_A \tag{3.8}$$

$$\frac{d}{dt}m_A = \alpha_A \cdot d_A - \gamma_A \cdot m_A \tag{3.9}$$

$$\frac{d}{dt}m_B = \alpha_B \cdot d_B - \gamma_B \cdot m_B \tag{3.10}$$

$$\frac{d}{dt}n_A = \beta_A \cdot m_A - \delta_A \cdot n_A + \tau_A^- \cdot (1 - d_B) - \tau_A^+ \cdot d_B \cdot n_A \tag{3.11}$$

$$\frac{d}{dt}n_B = \beta_B \cdot m_B - \delta_B \cdot n_B + \tau_B^- \cdot (1 - d_A) - \tau_B^+ \cdot d_A \cdot n_B \,, \tag{3.12}$$

where d_* is the abundance of unbound DNA$_*$, m_* is the abundance of mRNA$_*$ and n_* is the abundance of Protein$_*$ for $* \in \{A, B\}$. Bound DNA is expressed in terms of unbound DNA due to mass conservation. Note that these quantities are now continuous and have

to be interpreted as an average over a population of cells, i.e. with $d_{\mathrm{A}} = 0.5$, in half of the cells in the population, $\mathrm{DNA_A}$ is unbound.

We now solve Eqs. (3.7)–(3.12) at steady state by setting all time derivatives to zero. Using symmetric parameters for simplicity[1] we obtain the two following steady state solutions of Eqs. (3.7)–(3.12):

$$m_{\mathrm{A}}{}^{(1)} = m_{\mathrm{B}}{}^{(1)} = -\frac{\delta\tau^-}{2\beta\tau^+}(1-\eta) \tag{3.13}$$

$$n_{\mathrm{A}}{}^{(1)} = n_{\mathrm{B}}{}^{(1)} = -\frac{\tau^-}{2\tau^+}(1-\eta) \tag{3.14}$$

$$d_{\mathrm{A}}{}^{(1)} = d_{\mathrm{B}}{}^{(1)} = \frac{2}{1+\eta} \tag{3.15}$$

$$m_{\mathrm{A}}{}^{(2)} = m_{\mathrm{B}}{}^{(2)} = -\frac{\delta\tau^-}{2\beta\tau^+}(1+\eta) \tag{3.16}$$

$$n_{\mathrm{A}}{}^{(2)} = n_{\mathrm{B}}{}^{(2)} = -\frac{\tau^-}{2\tau^+}(1+\eta) \tag{3.17}$$

$$d_{\mathrm{A}}{}^{(2)} = d_{\mathrm{B}}{}^{(2)} = \frac{2}{1-\eta} \tag{3.18}$$

with $\eta = \sqrt{\frac{4\alpha\beta\tau^+}{\gamma\delta\tau^-} + 1}$. The first solution is positive, the second is negative (given all parameters are positive). Only the positive solution is of interest in biological systems, and given non-negative initial conditions the system will always converge towards the positive steady state solution (Müller-Herold, 1975).

Note that for small τ^+, Eqs. (3.13) and (3.14) reduce to the steady state solution of a simple two stage expression model (Thattai and van Oudenaarden, 2001)

$$n_{\mathrm{A}}{}^{(1)} = n_{\mathrm{B}}{}^{(1)} = \frac{\alpha\beta}{\gamma\delta}$$
$$m_{\mathrm{A}}{}^{(1)} = m_{\mathrm{B}}{}^{(1)} = \frac{\alpha}{\gamma}\,,$$

because

$$\eta \approx 1 + \frac{2\alpha\beta\tau^+}{\gamma\delta\tau^-}$$

for small τ^+ through the Taylor approximation

$$(1+x)^n \approx 1 + nx \text{ for } |x| \ll 1\,.$$

This is expected since setting τ^+ to 0 removes the interaction between both players, which will then evolve independently according to a two-stage expression model.

For decreasing τ^- the solution of the system approaches the origin

$$n_{\mathrm{A}}{}^{(1)} = n_{\mathrm{B}}{}^{(1)} = 0$$
$$m_{\mathrm{A}}{}^{(1)} = m_{\mathrm{B}}{}^{(1)} = 0$$

[1] $\alpha = \alpha_{\mathrm{A}} = \alpha_{\mathrm{B}}$, $\beta = \beta_{\mathrm{A}} = \beta_{\mathrm{B}}$, $\gamma = \gamma_{\mathrm{A}} = \gamma_{\mathrm{B}}$, $\delta = \delta_{\mathrm{A}} = \delta_{\mathrm{B}}$, $\tau^+ = \tau^+_{\mathrm{A}} = \tau^+_{\mathrm{B}}$ and $\tau^- = \tau^-_{\mathrm{A}} = \tau^-_{\mathrm{B}}$

because $\eta \approx \frac{\text{const}}{\sqrt{\tau^-}}$ and therefore the right hand sides of Eqs. (3.13) and (3.14) reduce to $\text{const} \cdot \frac{\tau^-}{\sqrt{\tau^-}}$. For decreasing τ^- this term will approach 0. Hence, if proteins never unbind the promoter, the system will be locked forever yielding 0 protein and mRNA levels in steady state.

We now assess the stability of the positive solution Eqs. (3.13)–(3.15) using standard linear stability analysis (see chapter 2). To reduce the complexity of our system for the stability analysis, we apply a quasi steady state approximation to the DNA binding/dissociation process ($\dot{d_\mathrm{A}} = \dot{d_\mathrm{B}} = 0$), reducing the dimensionality of our system to four equations:

$$\begin{aligned}
\frac{d}{dt} m_\mathrm{A} &= \alpha\psi(n_\mathrm{B}) - \gamma m_\mathrm{A} \\
\frac{d}{dt} m_\mathrm{B} &= \alpha\psi(n_\mathrm{A}) - \gamma m_\mathrm{B} \\
\frac{d}{dt} n_\mathrm{A} &= \beta m_\mathrm{A} - \delta n_\mathrm{A} + \tau^-(1 - \psi(n_\mathrm{A})) - \tau^+\psi(n_\mathrm{A}) n_\mathrm{A} \\
\frac{d}{dt} n_\mathrm{B} &= \beta m_\mathrm{B} - \delta n_\mathrm{B} + \tau^-(1 - \psi(n_\mathrm{B})) - \tau^+\psi(n_\mathrm{B}) n_\mathrm{B} \ ,
\end{aligned}$$

with $\psi(x) = \frac{\tau^-}{\tau^- + x \cdot \tau^+}$. The reduced system has the positive steady state solution

$$m_\mathrm{A}{}^{(ss)} = m_\mathrm{B}{}^{(ss)} = -\frac{\delta\tau^-}{2\beta\tau^+}(1 - \eta) \tag{3.19}$$

$$n_\mathrm{A}{}^{(ss)} = n_\mathrm{B}{}^{(ss)} = -\frac{\tau^-}{2\tau^+}(1 - \eta) \tag{3.20}$$

with $\eta = \sqrt{\frac{4\alpha\beta\tau^+}{\gamma\delta\tau^-} + 1}$. Notice that this is the same as the solution for mRNA and protein of the full system (Eqs. 3.13 and 3.14). We calculate the Jacobian matrix of the reduced system as

$$J = \begin{pmatrix}
-\gamma & 0 & 0 & \frac{\alpha\tau^+}{\tau^-} \cdot \psi^2(n_\mathrm{B}) \\
0 & -\gamma & \frac{\alpha\tau^+}{\tau^-} \cdot \psi^2(n_\mathrm{A}) & 0 \\
\beta & 0 & -\delta + \tau^+\psi^2(n_\mathrm{A})[1 + \tau^+ n_\mathrm{A} - \psi^{-1}(n_\mathrm{A})] & 0 \\
0 & \beta & 0 & -\delta + \tau^+\psi^2(n_\mathrm{B})[1 + \tau^+ n_\mathrm{B} - \psi^{-1}(n_\mathrm{B})]
\end{pmatrix} .$$

We evaluate the Jacobian at the steady state solution of the reduced system (Eqs. 3.19–3.20) and use the Hurwitz criterion to verify that all its eigenvalues have negative real part. We conclude that the system has one stable positive fixed point but we cannot analytically exclude the existence of limit cycles. However, inspection of the system's phase portrait (see Fig. 3.2) indicates that no limit cycles exist. Summarizing, we showed that the deterministic model has only one steady state solution and is thus monostable.

3.1.3 Stochastic model

Since the deterministic approach is only valid in the limit of large numbers, small molecule numbers of DNA, mRNA, and possibly proteins advocate a discrete probabilistic description of the toggle switch. We define the state of the system at time t as a vector $x(t)$,

where $x_i(t) \in \mathbb{N}_0$ is the abundance of species i at time t. Note that the state space is discrete as opposed to the deterministic model. To emphasize this difference we use the uppercase notation D_A, D_B, M_A, M_B, N_A, N_B for the number of molecules of the respective species. In particular, there is only one copy of DNA present which is either bound or unbound, such that $D_\mathrm{A} \in \{0,1\}$ and $D_\mathrm{B} \in \{0,1\}$ and the DNA state changes via reactions (3.5)–(3.6).

We can describe how the probability $\mathcal{P}(x,t)$ of being in a certain state x changes over time by using the Chemical Master Equation of the system (see chapter 2). We define $\mathcal{P}_{ij}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t)$ as the probability at time t to have M_A copies of $\mathrm{mRNA_A}$, M_B copies of $\mathrm{mRNA_B}$, N_A copies of $\mathrm{Protein_A}$, N_B copies of $\mathrm{Protein_B}$, and the corresponding promoter configuration ij where 0 (1) means an unbound (bound) promoter. The master equation of the reaction system (3.1)–(3.6) splits up into four coupled equations corresponding to the four promoter states and takes the explicit form:

$$
\begin{aligned}
\frac{d}{dt}\mathcal{P}_{00}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) &= \tau_\mathrm{A}^- E_{N_\mathrm{A}}^- \mathcal{P}_{01}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
&\quad + \tau_\mathrm{B}^- E_{N_\mathrm{B}}^- \mathcal{P}_{10}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
&\quad + [-\tau_\mathrm{B}^+ N_\mathrm{B} - \tau_\mathrm{A}^+ N_\mathrm{A} + \alpha_\mathrm{A}(E_{M_\mathrm{A}}^- - 1) \\
&\qquad + \alpha_\mathrm{B}(E_{M_\mathrm{B}}^- - 1) + \gamma_\mathrm{A}(E_{M_\mathrm{A}}^+ - 1)\cdot M_\mathrm{A} + \gamma_\mathrm{B}(E_{M_\mathrm{B}}^+ - 1)\cdot M_\mathrm{B} \\
&\qquad + \beta_\mathrm{A}(E_{N_\mathrm{A}}^- - 1)\cdot M_\mathrm{A} + \beta_\mathrm{B}(E_{N_\mathrm{B}}^- - 1)\cdot M_\mathrm{B} \\
&\qquad + \delta_\mathrm{A}(E_{N_\mathrm{A}}^+ - 1)\cdot N_\mathrm{A} + \delta_\mathrm{B}(E_{N_\mathrm{B}}^+ - 1)\cdot N_\mathrm{B}] \\
&\quad \cdot \mathcal{P}_{00}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
\frac{d}{dt}\mathcal{P}_{11}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) &= \tau_\mathrm{A}^+ E_{N_\mathrm{A}}^+ N_\mathrm{A} \mathcal{P}_{10}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
&\quad + \tau_\mathrm{B}^+ E_{N_\mathrm{B}}^+ N_\mathrm{B} \mathcal{P}_{01}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t)] \\
&\quad + [-\tau_\mathrm{A}^- - \tau_\mathrm{B}^- + \gamma_\mathrm{A}(E_{M_\mathrm{A}}^+ - 1)\cdot M_\mathrm{A} + \gamma_\mathrm{B}(E_{M_\mathrm{B}}^+ - 1)\cdot M_\mathrm{B} \\
&\qquad + \beta_\mathrm{A}(E_{N_\mathrm{A}}^- - 1)\cdot M_\mathrm{A} + \beta_\mathrm{B}(E_{N_\mathrm{B}}^- - 1)\cdot M_\mathrm{B} \\
&\qquad + \delta_\mathrm{A}(E_{N_\mathrm{A}}^+ - 1)\cdot N_\mathrm{A} + \delta_\mathrm{B}(E_{N_\mathrm{B}}^+ - 1)\cdot N_\mathrm{B}] \\
&\quad \cdot \mathcal{P}_{11}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
\frac{d}{dt}\mathcal{P}_{10}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) &= \tau_\mathrm{A}^- E_{N_\mathrm{A}}^- \mathcal{P}_{11}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
&\quad + \tau_\mathrm{B}^+ E_{N_\mathrm{B}}^+ N_\mathrm{B} \mathcal{P}_{00}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\
&\quad + [-\tau_\mathrm{B}^- - \tau_\mathrm{A}^+ N_\mathrm{A} + \alpha_\mathrm{B}(E_{M_\mathrm{B}}^- - 1) \\
&\qquad + \gamma_\mathrm{A}(E_{M_\mathrm{A}}^+ - 1)\cdot M_\mathrm{A} + \gamma_\mathrm{B}(E_{M_\mathrm{B}}^+ - 1)\cdot M_\mathrm{B} \\
&\qquad + \beta_\mathrm{A}(E_{N_\mathrm{A}}^- - 1)\cdot M_\mathrm{A} + \beta_\mathrm{B}(E_{N_\mathrm{B}}^- - 1)\cdot M_\mathrm{B} \\
&\qquad + \delta_\mathrm{A}(E_{N_\mathrm{A}}^+ - 1)\cdot N_\mathrm{A} + \delta_\mathrm{B}(E_{N_\mathrm{B}}^+ - 1)\cdot N_\mathrm{B}] \\
&\quad \cdot \mathcal{P}_{10}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t)
\end{aligned}
$$

$$\begin{aligned}\frac{d}{dt}\mathcal{P}_{01}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) &= \tau_\mathrm{A}^+ E_{N_\mathrm{A}}^+ N_\mathrm{A} \mathcal{P}_{00}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\ &+ \tau_\mathrm{B}^- E_{N_\mathrm{B}}^- \mathcal{P}_{11}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \\ &+ [-\tau_\mathrm{B}^+ N_\mathrm{B} - \tau_\mathrm{A}^- + \alpha_\mathrm{A}(E_{M_\mathrm{A}}^- - 1) \\ &\quad + \gamma_\mathrm{A}(E_{M_\mathrm{A}}^+ - 1) \cdot M_\mathrm{A} + \gamma_\mathrm{B}(E_{M_\mathrm{B}}^+ - 1) \cdot M_\mathrm{B} \\ &\quad + \beta_\mathrm{A}(E_{N_\mathrm{A}}^- - 1) \cdot M_\mathrm{A} + \beta_\mathrm{B}(E_{N_\mathrm{B}}^- - 1) \cdot M_\mathrm{B} \\ &\quad + \delta_\mathrm{A}(E_{N_\mathrm{A}}^+ - 1) \cdot N_\mathrm{A} + \delta_\mathrm{B}(E_{N_\mathrm{B}}^+ - 1) \cdot N_\mathrm{B}] \\ &\cdot \mathcal{P}_{01}(M_\mathrm{A}, M_\mathrm{B}, N_\mathrm{A}, N_\mathrm{B}, t) \; .\end{aligned}$$

The shift operators E_x^+ and E_x^- increase or decrease the function argument x by one, i.e. $E_x^\pm f(x) = f(x \pm 1)$. To our knowledge no results have yet been published on the solution of stochastic two-stage switches.

Since the master equation for the switch is analytically solvable only for a number of approximations (see e.g. Walczak et al., 2010) and not integrable for large molecule abundances, we simulate the system trajectories using Gillespie's algorithm (see chapter 2 and Gillespie, 1976). Each trajectory follows the master equation, and the set of infinite trajectories constitutes the distribution that solves the master equation.

3.1.4 Choice of parameters

To obtain appropriate parameters values for stochastic simulation, we delineate upper bounds for synthesis parameters from biophysical arguments and adapt degradation parameters to fit desired molecular levels. Table 3.1 lists the set of used parameter values.

First we derive upper boundaries for the transcription and translation rates. Transcription of DNA into mRNA is accomplished by the RNA-polymerase. One polymerase can process about 10-20 nucleotides (nt) per second in eukaryotes (Alberts et al., 2002; Dahlberg and Benkovic, 1991; Singh et al., 2007). As described by Alberts et al. (2002) the newly elongated RNA fragment is immediately released from the DNA, which enables other polymerases to follow up even before the first mRNA has been completed. The distance d between polymerases is estimated to be around 100 nt (Kennell and Riezman, 1977). The rate of transcription is independent of the sequence length l, since the longer the gene, the more polymerases can process it in parallel. Altogether we find the maximal transcription rate α by dividing the speed of transcription v with the sequence length l, multiplied with the number of transcribing polymerases,

$$\alpha = \frac{v}{l} \cdot \frac{n}{d} \approx \frac{10 \text{ nt/s}}{l} \cdot \frac{l}{100 \text{ nt}} = 0.1 s^{-1}$$

required that enough polymerases and nucleotides are present.

The maximal translation rate can be inferred in a similar way: Ribosomes, large complexes of proteins and rRNAs that translate mRNA into polypeptides, proceed with a speed v of 2 codons (= 6 nt = 2 amino acids) per second in eucaryotes (Alberts et al., 2002). One mRNA can be processed by many ribosomes (polyribosomes) at the same time (Alberts et al., 2002). The average space between two ribosomes is 80 nt or $\approx$ 27 amino

acids (AA) (Alberts et al., 2002). Therefore the overall translation rate for an mRNA of length n is

$$\beta \approx \frac{2\text{AA}/s}{l}\frac{l}{27\text{AA}} = 0.074s^{-}1 \ ,$$

again independent of the mRNA length l. This corresponds to the maximally possible translation rate. The actual rate will be smaller when not enough ribosomes or other involved molecules (tRNA, amino acids) are present. We estimate the minimal translation time as $1/0.074s^{-1} = 13.5s$, which is in good agreement with literature, where the time needed for one translation is said to be between 20 seconds and several minutes (Alberts et al., 2002). Notably, these transcription and translation rates only provide rough estimates of the relevant timescale. Throughout the manuscript, we use a transcription and translation rate of $\alpha = \beta = 0.05\ s^{-1}$, corresponding to an average time of 20 seconds per product, which seems to be reasonable in the context of the above considerations. The fact that both rates are equal is not expected to have influences on the results.

Interactions between proteins and DNA are mediated by specific regions of the proteins, called DNA-binding domains, which on the one hand can recognize specific DNA sequences and on the other hand maintain the interaction between DNA and protein. Zinc Fingers, Leucine Zippers or Helix-Turn-Helix motifs are prominent examples of DNA binding domains (Alberts et al., 2002). The binding between DNA and protein is maintained by hydrogen bonds, ionic bonds, and hydrophobic interactions. Single interactions are weak, but as many bonds are formed, the binding between DNA and protein becomes stronger. The binding rates are very fast compared to transcription and translation processes and according to Alon (2006) in the range of $1s^{-1}\text{Protein}^{-1}$ in *Eschericia coli*. The unbinding rate depends on the strength of the interaction and is assumed to be 10 times smaller ($0.1s^{-1}$) in our model, leading to strong binding of the protein to the DNA. All reaction rates of the models used in this work are summarized in Table 3.1.

As we showed above, the transcription and translation rates have upper bounds. The only way for a cellular system to further increase the abundance of proteins is to modulate the degradation rates of mRNA or proteins, giving longer lifetimes to mRNA and proteins. Thus, during this work we manipulate the degradations rates to adjust the system's protein level to a desired steady state. Notably, following the report of Warren et al. (2006), mRNA levels are set to 10 by adjusting the decay rate.

Reaction	Parameter value
Transcription α	$0.05s^{-1}$
Translation β	$0.05\ s^{-1}\text{mRNA}^{-1}$
mRNA degradation γ	$0.005\ s^{-1}$
Protein degradation δ	$5 \cdot 10^{-3}$ to $5 \cdot 10^{-6}s^{-1}$
DNA binding τ^{+}	$1\ s^{-1}\text{Protein}^{-1}$
DNA dissociation τ^{-}	$0.1\ s^{-1}$

Table 3.1: Parameters of the switch model used throughout this work. Protein degradation is chosen according to the desired protein level n. If not mentioned otherwise, all simulations and plots are based on this set of parameters.

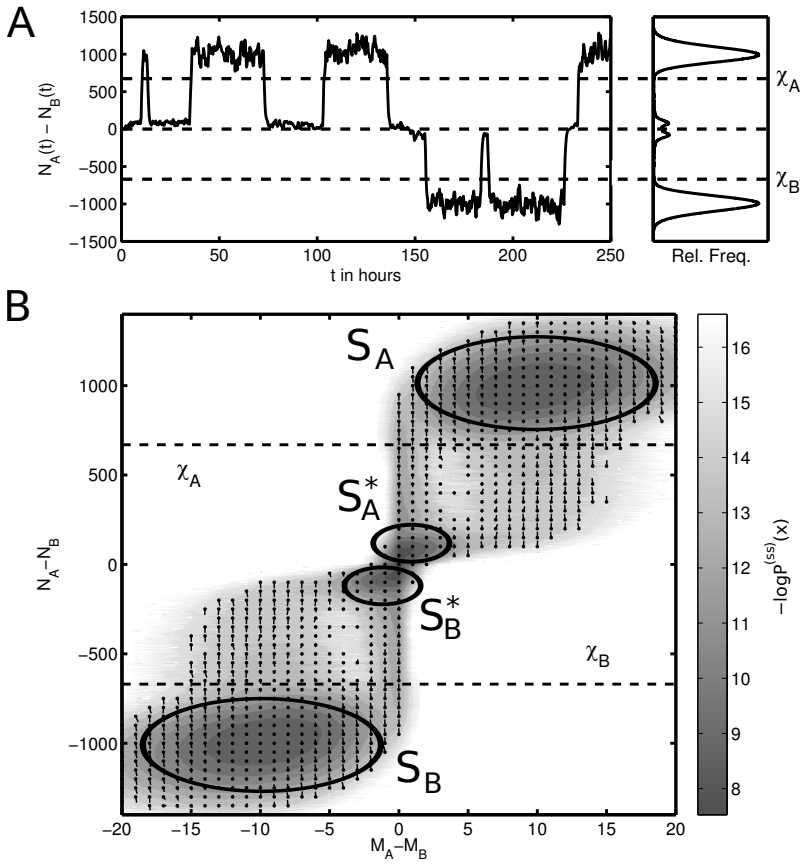


Figure 3.3: **Dynamics and quasi-potential of the switch showing the different attractors of the system.** A) The timecourse of $N_{\mathrm{A}}(t) - N_{\mathrm{B}}(t)$ clearly shows the dominating attractors, which can be separated in state space via the thresholds χ_{A} and χ_{B}. Either A dominates (attractor $\mathrm{S_A}$), or B dominates (attractor $\mathrm{S_B}$), or the system is temporarily locked by two bound promoters with only marginal protein expression of A or B (attractors S^*_{A} and S^*_{B}). A histogram of $N_{\mathrm{A}}(t) - N_{\mathrm{B}}(t)$ is shown on the right. B) The quasi-potential, defined as $\tilde{U}(x) = -\log\mathcal{P}^{(ss)}(x)$, includes the mRNA dimension of the system. It shows the four possible attractors as basins in a probability landscape. $\mathrm{S_A}$ and $\mathrm{S_B}$ are visible as basins at the lower left and upper right corners, whereas $\mathrm{S^*_A}$ and $\mathrm{S^*_B}$ are located around the origin ($N_{\mathrm{A}} - N_{\mathrm{B}} = M_{\mathrm{A}} - M_{\mathrm{B}} = 0$) of the landscape. Additionally, the outflux $J(x)$ acting on the system at the state x in state space are indicated as lines (circles correspond the the origin of the vector). Note that the outflux is different from concept of deterministic field lines. These vectors show that there are different paths for entering and leaving the dominating attractors. Parameters for the simulation are given in Table 3.1.

3.1.5 Quasi-potential

In this section we discuss the main features of the switch dynamics. Contrary to the deterministic model, time courses of the stochastic toggle switch model show multistable behavior (Fig. 3.3A). Given the parameters in Table 3.1 our toggle switch can adopt

different attractors (for an informal definition of a stochastic attractor, see section 2.2.5): The two attractors where one player dominates the other (called S_A and S_B depending on which player dominates) are clearly visible in Fig. 3.3A. A careful inspection of the timecourse and the probability distribution in Fig. 3.3A shows that there also exist two intermediate attractors where protein numbers are similar ($N_\mathrm{A} - N_\mathrm{B} \approx 0$). These attractors are called S^*_A and S^*_B from now on. In the timecourses of the system (Fig. 3.3A) one observes that the system frequently switches between the dominating and the intermediate attractors.

To get a deeper understanding of the complex dynamics of the system the notion of a quasi-potential can be used. Usually, one defines a potential $U(x)$ such that the forces $F(x)$ acting on the system correspond to the gradient of $U(x)$:

$$F = -\nabla U \ .$$

However, such a potential does not exist in general (if the curl $\nabla \times F$ is non-zero) and hence, one has to approximate it by a quasi-potential $\tilde{U}(x)$ such that

$$F = -\nabla \tilde{U} + F_r \ , \tag{3.21}$$

where F has been decomposed into a gradient of a potential and a remainder force F_r. Depending on the choice of F_r, different quasi-potentials can be constructed (for an overview, see Zhou et al., 2012). In the following, we pursue the construction proposed by Wang et al. (2011) such that $\tilde{U}(x) = -\log \mathcal{P}^{(ss)}(x)$, where $\mathcal{P}^{(ss)}(x)$ is the steady state distribution of the system. Note that for our purpose, the particular choice of quasi-potential is not important as it is used only as a visualization and no quantitative arguments are made.

The number of dimensions of the state space where the quasi-potential is defined equals the number of species in the system. Here the probability $\mathcal{P}^{(ss)}(x)$ of a state x in steady state is estimated from 15000 stochastic simulation runs obtained by the Stochkit software toolkit (Sanft et al., 2011). In Fig. 3.3B the projection of the quasi-potential on the $N_\mathrm{A} - N_\mathrm{B}$, $M_\mathrm{A} - M_\mathrm{B}$ plane is shown. The four attractors S_A, S_B, S^*_A and S^*_B can be seen clearly in the quasi-potential of the system. The two attractors S_A and S_B appear as basins at the lower left and upper right corner of Fig. 3.3 whereas the intermediate attractors S^*_A and S^*_B are located at the center and are not well separated.

The dominating attractors can easily be distinguished from the intermediate attractors via parameter dependent thresholds $\chi_\mathrm{A}, \chi_\mathrm{B}$ in the protein dimension, which we derive in the following: We approximate the protein number distribution in the attractors S_A and S_B using results from Thattai and van Oudenaarden (2001), who showed that for a simple two-state expression model, the mean and variance of protein numbers obey

$$\bar{N} = \frac{\alpha\beta}{\gamma\delta} \ \text{ and } \ \sigma^2 = \frac{\beta^2\alpha}{\gamma^2\delta + \delta^2\gamma} \ ,$$

respectively. Thereby, we assume that in the dominating attractors, the presence of the antagonist can be neglected due to its marginal transcription. We define the boundary χ_x of attractor S_x using a normal approximation of the dominating protein's distribution as

$$\chi_x = \bar{N}_x - Z_q \cdot \sigma_x \ , \tag{3.22}$$

where Z_q is the $q\%$ quantile of the standard normal distribution of the protein number with mean $\langle n \rangle_x$ and standard deviation σ_x. Using $q = 0.1$ throughout our study assures that 99.9% percent probability mass of the distribution lies beyond the lower boundary. Therefore we are certain to capture all relevant protein numbers belonging to S_A and S_B. Using these boundaries, we can define the attractors S_A and S_B accordingly:

$$S_\mathrm{A} = \{s \in S | N_\mathrm{A} > \chi_\mathrm{A} \wedge N_\mathrm{B} < \chi_\mathrm{B}\}$$
$$S_\mathrm{B} = \{s \in S | N_\mathrm{A} < \chi_\mathrm{A} \wedge N_\mathrm{B} > \chi_\mathrm{B}\} \, .$$

Importantly, one has to keep in mind that the system considered is out of equilibrium and that the dynamics of a non-equilibrium system are not entirely determined by the gradient of the quasi-potential $\tilde{U}$ but by the additional remainder force F_r stemming from the non-integrability of the system (see Eq. 3.21 and Wang et al., 2008). As a consequence barrier heights in the quasi-potential do not necessarily correlate with the probability of crossing the barrier.

To understand the dynamics of the switch in more detail we therefore consider for each state x in the state space the outflux $J(x)$ acting on the the system at this point (Schultz et al., 2008). We calculate the outflux as:

$$J(x) = \mathcal{P}^{(ss)}(x) \sum_y \mathcal{P}(y|x)(y - x) \, ,$$

where the probability $\mathcal{P}(y|x)$ of state y succeeding state x and the probability $\mathcal{P}^{(ss)}(x)$ are calculated from stochastic simulations. Note that the outflux is different from the concept of field lines used in phase portraits of ordinary differential equations. The outflux $J(x)$ is plotted as small arrows in Fig. 3.3 (vectors are normalized and circles correspond the origin of the vectors) for all states x with $\mathcal{P}^{(ss)}(x) > 2.5 \cdot 10^{-7}$. This indicates where the system will move from the current state on average. Due to this outflux the system enters and leaves the attractors S_A and S_B through different paths. This phenomenon has been described by Wang et al. (2010) and linked to the emergence of time directionality in non-equilibrium systems. In order to move from high (S_A or S_B) to low (S_0) protein numbers, at first the corresponding mRNA number has to drop. On the contrary, moving from low to high protein numbers requires the rise of mRNA numbers first.

A different view on the system's dynamics is provided by the quasi-potential landscape and outflux in the $N_\mathrm{A}^\mathrm{total}, N_\mathrm{B}^\mathrm{total}$ plane (Fig. 3.4), where $N_\mathrm{A}^\mathrm{total} = (1 - D_\mathrm{B}) + N_\mathrm{A}$ is the total number of $\mathrm{Protein_A}$ in the system, bound to DNA (first term) or free (second term). Choosing $N_\mathrm{A}^\mathrm{total}$ and $N_\mathrm{B}^\mathrm{total}$ as projected dimensions shows four distinct basins in the quasi-potential landscape. Two basins correspond to the attractors S_A and S_B. These are characterized by high amounts of the dominating protein and zero proteins of the repressed species. The attractors S_A^* and S_B^* are now clearly separated. In these two basins a single protein of one species is present and only a moderate protein number of the other species. In the following we show why these basins emerge and how the system moves between the attractors.

3.1.6 Dynamics in the quasi-potential

We explain the dynamics of the system with a typical trajectory of the system: Let us start with the trajectory in the attractor S_A (lower right) where $\mathrm{Protein_A}$ dominates $\mathrm{Protein_B}$.

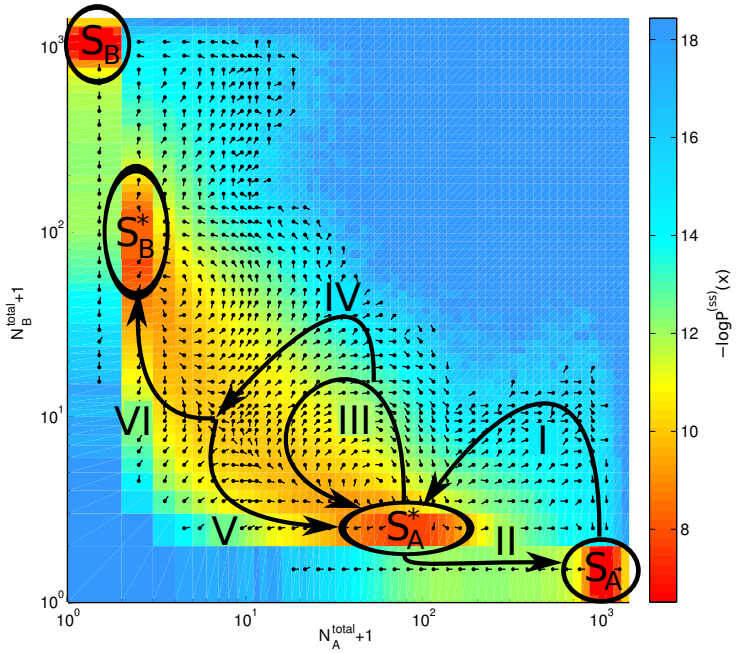


Figure 3.4: **Quasi-potential of the system projected onto the $N_{\mathrm{A}}^{\mathrm{total}}$ and $N_{\mathrm{B}}^{\mathrm{total}}$ dimensions.** Note that both axis are on logarithmic scale and are shifted by 1 in order to include $N_{\mathrm{A}}^{\mathrm{total}} = 0$ and $N_{\mathrm{B}}^{\mathrm{total}} = 0$. Therefore the lowest row in the plot corresponds to the case $N_{\mathrm{B}}^{\mathrm{total}} = 0$. The quasi-potential $\tilde{U}(x) = -\log \mathcal{P}^{(ss)}(x)$ is color coded where red areas reflect minima of the landscape. Visible are four minima corresponding to S_{A} (lower right), S_{B} (upper left), S_{A}^* (lower middle) and S_{B}^* (middle left). The vectors of the outflux at each point in state space are drawn as lines (circles correspond the the origin of the vector). Note that the outflux is different from concept of deterministic field lines. In contrast to Fig. 3.3 the vectors are normalized and therefore show only the direction, not the magnitude of the field. Bold arrows reflect typical trajectories (I-VI) of the system. For a discussion, see the main text.

Due to stochastic fluctuations in the promoter status, eventually a burst of proteins of B will occur and inhibit the promoter of A, whose protein numbers will drop (Fig. 3.4, trajectory I). While the formerly dominating $\mathrm{Protein_A}$'s are degraded, also the newly created $\mathrm{Protein_B}$ quickly decreases in numbers and only one bound $\mathrm{Protein_B}$ is saved from degradation. This drives the system towards the origin in the quasi-potential of Fig. 3.4. However, a single $\mathrm{Protein_B}$ cannot completely suppress the promoter of $\mathrm{DNA_A}$, leading to a small but constant synthesis of $\mathrm{Protein_A}$. The system settles into an intermediate state (S_{A}^*) defined by the presence of one $\mathrm{Protein_B}$ and an intermediate amount of $\mathrm{Protein_A}$ originating from the leaky inhibition of $\mathrm{DNA_A}$ and bursting. In order to leave this basin the system has two options: Either the single $\mathrm{Protein_B}$ is degraded when it momentarily is not bound to the promoter. Consequently the levels of $\mathrm{Protein_A}$ rise again and the system reaches S_{A}. The system is moved to the lower border of the quasi-potential where a strong

outflux pushes it towards S_A (Fig. 3.4, trajectory II). Alternatively, a burst of Protein_B displaces the system from S_A^* into regions where the vector field points strongly towards the diagonal ${N_A}^{\text{total}} = {N_B}^{\text{total}}$ (Fig. 3.4, trajectory III). However this burst is typically not strong enough to move the system onto the diagonal and it will fall back into the basin S_A^*. In order to enable a change from S_A^* to S_B^* the system has to reach the diagonal. This is accomplished if, while the system is moving towards the diagonal after the burst, additional bursts of Protein_B move it onto the diagonal (Fig. 3.4, trajectory IV). Once the system has hit the diagonal both protein levels will drop to very low numbers since non of the players has any significant advantage. Here by chance the system will move to any side of the diagonal and either towards S_A^* or S_B^* (Fig. 3.4, trajectories V, VI).

We find that leaving S_A^* towards S_A (Fig. 3.4, trajectory II) is much more probable than hitting the diagonal from S_A^* (Fig. 3.4, trajectory IV), which would provide the chance of switching. This is obvious from the mechanism described above: Even though the events triggering the two alternatives (degradation of Protein_B and an initial burst of Protein_B) have similar probabilities, the diagonal crossing requires additional events and is therefore much less probable. This cannot be deduced from the quasi-potential landscape alone: From Fig. 3.4 it can visually be inferred that the barrier separating S_A and S_A^* is higher than the barrier separating S_A^* and S_B^*. This wrongly suggests that moving between S_A^* and S_B^* occurs more frequently than moving between S_A and S_A^*.

Comparing the system dynamics of our switch with other descriptions we find that (i) deterministic one-stage and two-stage models show no bistability while (ii) a probabilistic one-stage model exhibits tristability with only one intermediate attractor (Lipshtat et al., 2006). We speculate that translational bursting destabilizes the intermediate attractor of the one-stage model, where none of the two players can overwhelm the other. Bursting provides an easy mechanism to escape this deadlock situation: It gives the player whichever bursts first a huge advantage over the other, giving rise not only to one protein (as in the one-stage model) but several proteins. As a result, the two-stage system is always quickly pushed away from the diagonal and stabilizes in the attractors S_A^* or S_B^*. Thus, only the combination of a probabilistic description with a two-stage model of gene expression leads to the complex multi-attractor dynamics described above.

3.1.7 Residence times

Genetic toggle switches are thought to be involved in the differentiation process of cells. A common idea is that different cell fates correspond to the different attractors of the system (Huang et al., 2005). Therefore it is of interest how long the system will stay in one of these attractors.

Residence time in S_A and S_B

Here , we focus on the time the system will stay in the attractors S_A or S_B. We assume that only in these two attractors the concentration of either player is sufficiently high to carry out a downstream biological function which resembles the switch's decision.

In previous contributions, such quantities have been calculated or determined by stochastic simulation for simpler switch models and were called spontaneous switching time (Bialek, 2001), switch lifetime (Warren and ten Wolde, 2004), mean first-passage

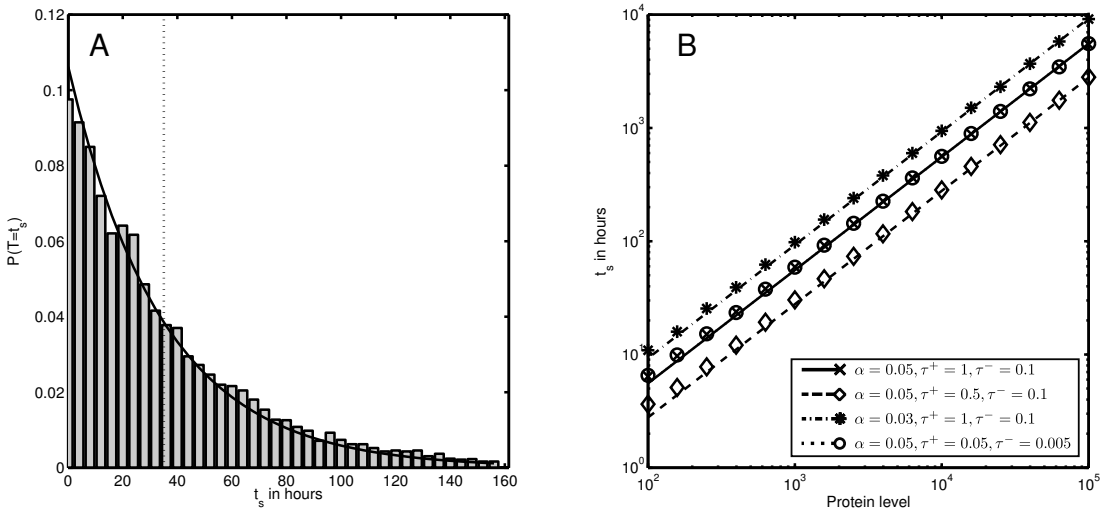


Figure 3.5: **Residence time t_s in the two-stage toggle switch.** A) The distribution for t_s obtained by stochastic simulation is in good agreement with the geometric distribution derived from our mean-field approximation. The mean of the distribution is indicated by a dashed line. The protein decay rate was set to $\delta = 8 \cdot 10^{-4} s^{-1}$. B) Mean residence time t_s versus mean protein level $\bar{N}$ derived from stochastic simulation (symbols) and our analytical approximation (lines) for four different parameter settings. Note that the analytical approximations as well as the simulation results of the first and fourth parameter set coincide. The exponent in the relation $t_s \propto (\bar{N}_\mathrm{A})^\nu$ is $\nu = 1$, in accordance with Eq. (3.24).

time (Kepler and Elston, 2001), or switching time (Barzel and Biham, 2008). Since the switch may flip from a dominating to an intermediate attractor, we choose residence time as the appropriate term for the quantity calculated below. In the following, we derive an analytical approximation for the time the switch stays in a dominating attractor, S_A or S_B, called the residence time t_s.

Let us assume that the system is in attractor S_A. Hence, the promoter of $\mathrm{DNA_B}$ is bound by $\mathrm{Protein_A}$ while the promoter of $\mathrm{DNA_A}$ is unbound. We assume that the protein levels in this attractor can be described with the simple two-stage model (Thattai and van Oudenaarden, 2001), resulting in a mean $\mathrm{Protein_A}$ level of $\bar{N}_\mathrm{A} = (\alpha_\mathrm{A}\beta_\mathrm{A})/(\gamma_\mathrm{A}\delta_\mathrm{A})$. Consequently, the protein level of $\mathrm{Protein_B}$ is $N_\mathrm{B} = 0$ as it is inhibited by the high levels of $\mathrm{Protein_A}$. In order to leave S_A it is crucial that one $\mathrm{Protein_B}$ is synthesized, which then can bind the promoter of $\mathrm{DNA_A}$ and shut down the synthesis of $\mathrm{Protein_A}$, ultimately driving the system out of S_A and into S_A^*. This trajectory (called trajectory I in Fig. 3.4) involves the following events: (i) unbinding of $\mathrm{Protein_A}$ from $\mathrm{DNA_B}$, (ii) synthesis of $\mathrm{Protein_B}$ during the unbound phase, and (iii) binding of $\mathrm{Protein_B}$ to the promoter of $\mathrm{DNA_A}$ before $\mathrm{Protein_B}$ is degraded.

First we describe the unbinding of $\mathrm{Protein_A}$ from $\mathrm{DNA_B}$. While the system is in S_A, $\mathrm{Protein_A}$ dissociates various times, leaving the promoter of $\mathrm{DNA_B}$ unbound. The average

time the promoter remains unbound, t_u, is equal to the average time until a binding reaction occurs, which is

$$t_u = \frac{1}{\tau_{\mathrm{A}}^{+} \cdot \bar{N}_{\mathrm{A}}} \, .$$

The time the promoter stays unbound is a random variable itself, but for simplicity we approximate it with its mean value. Note that t_u depends, somewhat counterintuitively, on τ^+ and not on τ^-, with $\tau_{\mathrm{A}}^{+}\bar{N}_{\mathrm{A}}$ being the propensity for a binding reaction.

To ultimately synthesize a $\mathrm{Protein_B}$, at least one $\mathrm{mRNA_B}$ has to be transcribed during t_u and translated before degradation. The probability of k transcription reactions to happen during t_u is

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

as the number of transcription reactions K during t_u is Poisson-distributed with mean $\alpha_{\mathrm{B}} \cdot t_u$. Thus, the probability of at least one transcription during the unbound phase is

$$q_s = 1 - P(K = 0) = 1 - \exp\left(-\frac{\alpha_{\mathrm{B}}}{\tau_{\mathrm{A}}^{+} \cdot \bar{N}_{\mathrm{A}}}\right) \, .$$

The probability of translation during an average mRNA lifetime $1/\gamma_{\mathrm{B}}$ is accordingly $q_t = 1 - \exp\left(-\beta_{\mathrm{B}}/\gamma_{\mathrm{B}}\right)$. Finally the probability for a binding reaction during average protein lifetime $1/\delta_{\mathrm{B}}$ is $q_b = 1 - \exp\left(-\tau_{\mathrm{B}}^{+}/\delta_{\mathrm{B}}\right)$.

However, not only one but several unbound phases may occur before $\mathrm{Protein_B}$ is successfully synthesized. The number L of unbound phases until and including successful synthesis follows a geometric distribution, $P(L = l) = (1-q)^{l-1}q$ with parameter $q = q_s \cdot q_t \cdot q_b$. The average number of unbound phases during a time interval Δt is $\tau_{\mathrm{A}}^{-} \cdot \Delta t$. Thus, we can convert the random variable L into $T = L/\tau_{\mathrm{A}}^{-}$ via a linear transformation of a random variable, giving the actual time until successful synthesis of $\mathrm{Protein_B}$. Notably, the derivation of the distribution for residence times goes beyond previous mean-field approximations. Using the properties of the geometric distribution for the random variable T, we end up with the mean and the variance of the residence time:

$$t_s = \frac{1}{\tau_{\mathrm{A}}^{-} \cdot q_s q_t q_b} \quad \text{and} \quad \sigma_{t_s}^2 = \frac{1}{(\tau_{\mathrm{A}}^{-})^2} \cdot \frac{1 - q_s q_t q_b}{(q_s q_t q_b)^2} \, . \tag{3.23}$$

An important approximation for the residence time can be derived under the assumption of rapid translation and slow mRNA degradation, $\beta \gg \gamma$, leading to $q_t \approx 1$. This implies that it is quite certain that an mRNA will be translated at least once before degradation. In the regime of rapid transcription factor binding ($\tau^+ \gg \delta, \alpha$) the probability for a binding reaction is close to one, $q_b \approx 1$, while the probability for at least one transcription can be approximated with $q_s \approx \alpha_{\mathrm{B}}/(\tau_{\mathrm{A}}^{+}\bar{N}_{\mathrm{A}})$. Taken together, this leads to a linear dependence of the residence time on the protein number,

$$t_s \approx (\tau_{\mathrm{A}}^{+}/\tau_{\mathrm{A}}^{-}) \cdot (\bar{N}_{\mathrm{A}}/\alpha_{\mathrm{B}}) \, . \tag{3.24}$$

We now compare our analytical approximation with the residence time derived from simulations. To that end, we have to infer the dominating attractors from the simulated

time courses. Recall that we can identify the dominating attractors via thresholds χ_A, χ_B at protein levels. The residence time of attractor S_A (S_B) is estimated as the consecutive time in a trajectory where $N_A > \chi_A$ ($N_B > \chi_B$). We compare the analytically derived geometric distribution for the residence times (see Eq. 3.23) with numerical results by simulating the switch with a given parameter set and estimating the residence times from 10000 stochastic simulations. Fig. 3.5A shows excellent agreement between the geometric distributed residence time and the simulations for a protein degradation rate of $\delta = 8 \cdot 10^{-4} s^{-1}$. This legitimates the approximations and assumptions made above for the parameter regime of rapid transcription factor binding. From the analysis of the mean residence time for different protein half-lives, we find again a good agreement between the simulation and the approximation (see Fig. 3.5B). Moreover, the slope of the log-log curve of the simulation is 1 – confirming a linear dependence of the residence time from the mean protein level.

With the result from Eq. (3.23) we can compare the mean residence time of different switch models. First we consider a gene expression model where transcription and translation are condensed into a single protein synthesis reaction. In analogy to the two-stage model of gene expression (Shahrezaei and Swain, 2008), this can be called a one-stage model of gene expression. To achieve the same amount of proteins at similar degradation rates, the synthesis rate in the one-stage model needs to be larger compared to the transcription and translation rates in the two-stage model. The probability q_t which accounts for translation during mRNA lifetime can be set to 1, since there is no mRNA stage and proteins are produced immediately. The binding probability q_b remains unchanged. However, because of the increased synthesis rate, the probability q_s of synthesis during the unbound phase will be larger than in the two-stage model. Therefore, the mean residence time will be decreased in the one-stage model as compared to the two-stage model, leading to more frequent attractor changes. This finding is in accordance with the previously reported stabilizing effect of bursts in an exclusive switch (Schultz et al., 2008).

A second modification of the switch includes not only mutual inhibition but also autoactivation of both genes. If the promoter of the gene is unbound it will be transcribed with a small basal rate κ. If the promoter is bound by its own protein product the gene will be transcribed with full rate $\alpha \gg \kappa$. Repressor bound promoters are inactive. For simplicity we assume that either activators or repressors are bound but not both at the same time. Note that in this case also the deterministic ODE model is bistable (Siegal-Gaskins et al., 2011). Considering the mean residence time in a two-stage switch with autoactivation, we find that the probability q_s of mRNA synthesis during the unbound phase is smaller than in the ordinary two-stage model. Since no activator is present in this attractor, mRNA has to be transcribed with the small basal rate κ, making the transcription more improbable. The probability q_t for translation remains unchanged. However, the probability q_b of protein binding to the antagonistic promoter is also decreased since this promoter is occupied by the abundant activator most of the time. Therefore, repressor binding to this promoter requires an additional dissociation reaction of the activator during repressor lifetime. As both q_s and q_b are decreased the mean residence time in switch models with autoactivation will be strongly increased compared to the ordinary two-stage model.

Summarizing, we find that the residence time is (i) geometrically distributed, (ii) the

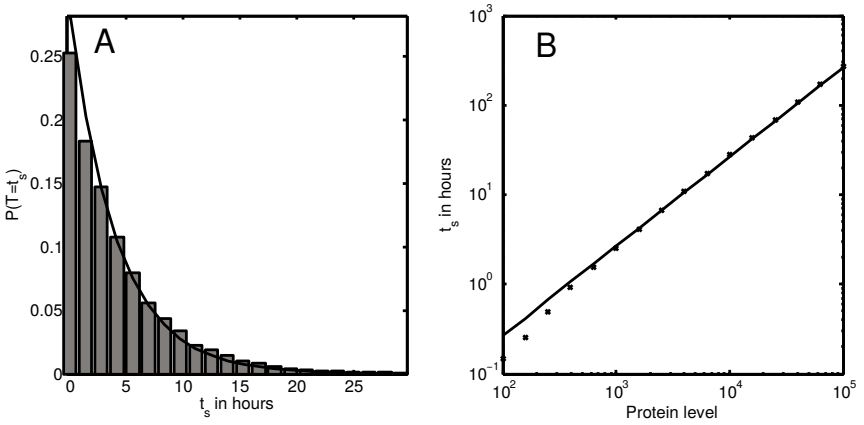


Figure 3.6: **Residence time of the system in the intermediate attractors ($S^*_{\mathbf{A}}$ or $S^*_{\mathbf{B}}$).** A) The histogram of the residence times in S^*_{A} obtained from stochastic simulation suggests that the residence time in the intermediate attractors follows a geometric distribution. The solid line corresponds to the distribution predicted by Eq. (3.26). B) Similar to Fig. 3.5 one observes a linear scaling of the mean residence time and the protein abundance. The solid line corresponds to the mean residence time predicted by Eq. (3.26), whereas crosses correspond to mean residence time derived from stochastic simulation. The prediction fits well to the simulated data for protein levels beyond 10^3 but deviates at low protein levels. This deviation is due to the fact that at low protein levels the attractors S^*_{A} and S^*_{B} are indistinguishable from S_{A} and S_{B} in simulated data because of strong intrinsic noise.

mean of the distribution grows linearly with the number of proteins for slow mRNA degradation, and (iii) both the intermediate step of mRNA production and the autoactivation of transcription factors increase the residence time.

Residence time in $S^*_{\mathbf{A}}$ and $S^*_{\mathbf{B}}$

Analogous to the residence times in S_{A} and S_{B} we derive an analytical expression for the residence time in the intermediate attractors S^*_{A} and S^*_{B}. Two mechanisms to escape from S^*_{A} are possible: Degradation of the single $\mathrm{Protein_B}$ when it is not bound to the promoter (Trajectory II in Fig. 3.4) or a burst of $\mathrm{Protein_B}$ (Trajectory III in Fig. 3.4). We find that the probability of degradation of $\mathrm{Protein_B}$ while unbound is:

$$p_{deg} = 1 - \exp\left(-\frac{\delta_{\mathrm{B}}}{\tau^+_{\mathrm{B}}}\right)$$

The probability for a burst of $\mathrm{Protein_B}$ is:

$$p_{burst} = 1 - \exp\left(-\frac{\alpha_{\mathrm{B}}}{\tau^+_{\mathrm{A}} \cdot \bar{N}^*_{\mathrm{A}}}\right)$$

where $\bar{N}^*_{\mathrm{A}}$ is the average amount of $\mathrm{Protein_A}$ in the attractor S^*_{A} [2]:

$$\bar{N}^*_{\mathrm{A}} = \frac{\tau^-}{\tau^- + \tau^+} \cdot \frac{\alpha\beta}{\gamma\delta} \,. \tag{3.25}$$

As delineated before, the above probabilities are the parameters of the geometric random variables describing the time until such an event happens. We are interested in the residence time in the attractor and therefore have to take the minimum of the two geometric random variables, since whichever event happens first leads to an escape from the attractor. The minimum of the two geometric random variables is again a geometric random variable but with an adjusted parameter

$$p_{min} = 1 - (1 - p_{deg}) \cdot (1 - p_{burst}) \,.$$

Finally we calculate the mean and variance of the residence time in S^*_{A} as:

$$t_s = \frac{1}{\tau^- \cdot p_{min}} \tag{3.26}$$

$$\sigma^2_{t_s} = \frac{1}{(\tau^-)^2} \cdot \frac{1 - p_{min}}{(p_{min})^2} \,. \tag{3.27}$$

A simulation study for $S^*_{\mathrm{A}}/S^*_{\mathrm{B}}$ confirms this calculations (see Fig. 3.6). Again, the residence time in the attractor is exponentially distributed and scales with the number of proteins in the system.

What is still missing to describe the entire dynamics of the system, but remains elusive at this point is an analytical expression for the ratio of probabilities of the transitions $S^*_{\mathrm{A}} \to S_{\mathrm{A}}$ and $S^*_{\mathrm{A}} \to S^*_{\mathrm{B}}$.

3.1.8 Discussion

Implications for cell differentiation We now discuss the implications of our findings in the context of cell differentiation driven by the toggle switch. We find that the residence time in S_{A} and S_{B}, a key property of the system, is geometrically distributed. Previous contributions (Bialek, 2001; Warren and ten Wolde, 2005; Barzel and Biham, 2008) focused only on the mean residence time and did not consider its underlying distribution. What does a geometric distribution for the residence time imply for the differentiation process dependent on the state of a genetic switch? To discuss this question, let us first reason on how a differentiation decision could be established with the toggle switch lined out in the previous sections.

We discriminate two scenarios for the differentiation of a cell: In the first scenario, the state of the switch completely determines the cell fate. Starting in the progenitor attractors S^*_{A} or S^*_{B}, after a certain amount of time, the switch will move to a committed attractor. We assume that the high numbers of proteins of the dominating player will trigger the differentiation program of the associated lineage and establish the mature cell type. However due to stochasticity, the switch will drop out of the committed attractors

[2] The promoter of the dominating player A is free only $100 \cdot \frac{\tau^-}{\tau^+ + \tau^-}$ % of the time, because of the presence of one protein of B, leading to reduced mRNA synthesis.

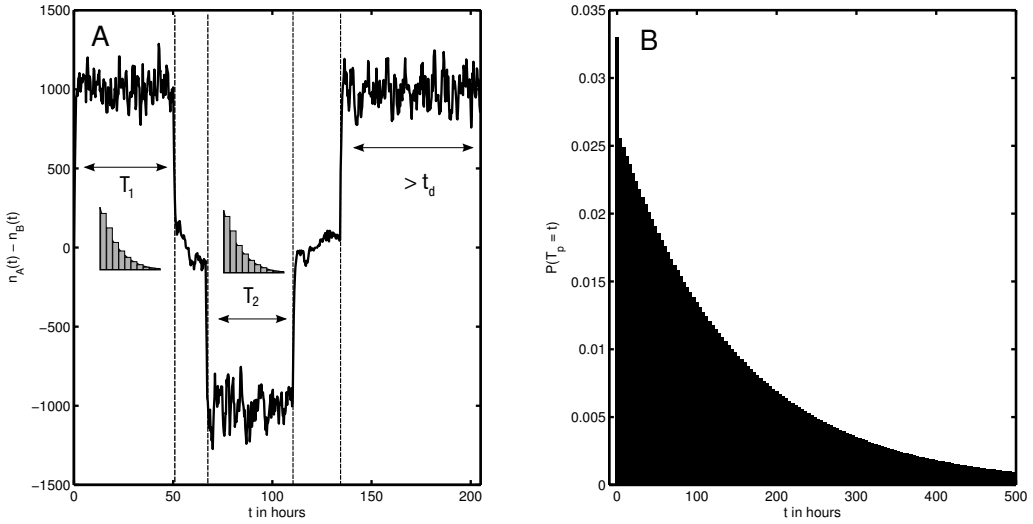


Figure 3.7: A: Scheme of the differentiation mechanism. The system changes the dominating attractor twice before it settles into S_A for a time greater than t_d, which induces differentiation. The residence times in S_A and S_B follow a (truncated) geometric distribution. The overall time until differentiation is therefore (neglecting the intermediate attractors) the sum of these geometric random variables. B: Distribution of the differentiation time T_p for a residence time of $1h$ and $t_d = 5h$. Small differentiation times are more probable than higher bigger differentiation times. Note the strong peak at $t = 0$ indicating that it is most probable for a cell to differentiate immediately .

and the cell will not only lose the current lineage decision, but possibly even switch to the opposing cell fate. In order to establish stable lineages in this scenario, the cell has to assure that the residence time of the switch is much longer than all relevant biological processes of the cell, especially cell lifetime. This guarantees that the cell will keep a lineage decision once it has obtained one. Yet the geometric distribution of the residence time imposes difficulties in this scenario: Even if the mean residence time is high, short residence times will always be more probable than longer residence times. The toggle switch could either be stabilized with the aforementioned autoactivation of the players, or with very high protein numbers so that the geometric distribution flattens and transforms to an almost uniform distribution. Both means would assure that only a very small percentage of a population of cells forgets its lineage decision during lifetime.

In the second scenario we assume that the cell gets locked in one fate by changing the shape of the underlying potential so that further transitions between attractors become less possible. Such a change of the potential could for example be facilitated by chromatin changes, as proposed by Akashi et al. (2003). In the following, we assume that only if one state dominates the other for a long enough fixation time t_d, downstream genes necessary for the decision process are activated (e.g. leading to chromatin remodeling), and the cell differentiates. Such a time depending property could be implemented with low-pass filters (see (Narula et al., 2010) for an example in hematopoietic stem cells) and would

allow for an integration of external signals (see Rieger et al. (2009) for the instructive power of hematopoietic cytokines). In this scenario, the residence time determines when differentiation will occur: The switch will constantly move into and out of the dominating attractors, until the residence time is finally long enough so that the dominating player can activate the downstream differentiation machinery (Fig. 3.7A). Ignoring the time the system spends in the intermediate attractors and just summing up the residence times in S_{A} and S_{B} until a long enough residence time for differentiation occurs, we find that this time follows a geometric-like distribution (see Fig. 3.7B and Supplementary Information of Strasser et al., 2012). Under this differentiation mechanism, most cells will differentiate very fast and only few cells will take more time. Experiments that measure the time for single cells needed to go from the primed to the committed state (as an extension to the 2-day threshold reported by Heyworth et al., 2002 for GM-CTC cells) in order to support or reject these hypotheses remain to be done.

In previous studies (Roeder and Glauche, 2006; Huang et al., 2007; Wang et al., 2010), attractors where either one or the other player is dominating, thereby repressing the antagonist (S_{A}, S_{B}) corresponded to committed cells. We also find analogs for the intermediate states S_{A}^* and S_{B}^*. In these attractors the system has a strong preference towards one specific dominating attractor, but is not fully committed yet. A similar behavior is known as lineage priming in stem cell biology (Graf and Stadtfeld, 2008). Two different studies (Müller-Sieburg et al., 2002; Chang et al., 2008) showed that a population of stem and progenitor cells, respectively, can be divided into subpopulations that mainly give rise to only one of two possible cell types. In our simple model this would correspond to stem cells that reside either in S_{A}^* or S_{B}^*. These stem cells can still give rise to both cell fates but have a strong tendency towards one of them.

Remarkably only a two-stage probabilistic model of the toggle switch shows dynamics reminiscent of lineage priming. Although a progenitor state exists in one-stage models of the toggle switch, cells in this state will move to either the one or the other committed state with equal probability.

Comparison to previous work Finally we discuss how our findings relate to previous studies on the toggle switch. We found that the mean of the residence time distribution scales linearly with the number of proteins in the system. The more proteins are present, the longer the average residence time in S_{A} or S_{B}. However, shorter residence times are still most probable due to the geometric distribution. This holds for the one-stage, the two-stage, and the auto-activating scenario.

This linear scaling differs from the exponential (Bialek, 2001) or near exponential (Warren and ten Wolde, 2005) scaling described previously in the one-stage scenario. In contrast to our model, the model of Warren and ten Wolde (2005) considers dimerization of the transcription factors, motivated by the fact that cooperative binding is necessary to achieve bistability in a deterministic framework (Chickarmane et al., 2009). We showed that, as soon as stochastic fluctuations are introduced, a system with multiple attractors is achieved that can act as a proper switch with additional states of low co-expression. Including dimerization as a prerequisite for inhibition in a one-stage model will strongly increase the stability of the attractors $S_{\mathrm{A/B}}$ (Warren and ten Wolde, 2005). This is consistent with our findings: Instead of requiring translation of one protein of the suppressed

species, we now require this rare event to happen twice during a short time frame, which is much less probable. However, the inclusion of dimerization will have less effect on the two-stage switch: Since proteins are typically synthesized in bursts (in our model the average burst size is $\beta/\gamma = 10$) and dimerization is a fast process (Warren and ten Wolde, 2005), as soon as one burst occurs almost certainly a dimer is formed and can inhibit the currently dominating player. Therefore the probability of leaving the attractors $S_{\mathrm{A/B}}$ is similar to a non-dimeric inhibition.

Contrary to our results, Warren and ten Wolde (2005) report that introduction of mRNAs reduces the stability of the switch. This discrepancy can be understood in the light of dimerization. In their one-stage model dimerization is a key ingredient of stability, which is lost when introducing translational bursts ("shot noise"). As we considered monomeric transcription factor binding, stability does not rely on dimerization. Therefore mRNAs increase the stability of the system, because they introduce additional conditions required for switching.

Due to these differences in the model it is hard to resolve the discrepancy between our linear and the exponential scaling of residence time found by Bialek (2001) and Warren and ten Wolde (2005). However we want to emphasize that the theoretical results shown by Warren and ten Wolde (2005) only consider protein numbers up to 30. In this region our simulation results show slight deviations from the analytical linear dependence (Fig. 3.5). At such low protein numbers the system does not only leave the dominating attractor according to the mechanism described in our results. It is also likely that just due to fluctuations in the gene expression (not fluctuations in the promoter) the dominating attractor is left. This mechanism operates only at very small protein numbers and its probability rapidly decreases with rising protein numbers. Therefore our results do not contradict the findings of Warren and ten Wolde (2005), but consider a different parameter regime with higher protein numbers. Interestingly, the noise-driven attractor changes are also described by Kashiwagi et al. (2006) where the authors link this mechanism to the selection of a favorable, less noisy attractor in *E. Coli* populations.

In another contribution, Morelli et al. (2008) use the forward flux sampling algorithm to assess the stability of a one-stage genetic toggle switch with dimeric transcription factor binding. They find a similar mechanism of attractor flipping which is based on the synthesis of the suppressed species due to promoter fluctuations. Using the forward flux sampling, they obtain estimates of the switching rate (the inverse of the mean residence time) for different amounts of fluctuations in DNA-protein interaction and dimerization. Morelli et al. (2008) modulate the size of fluctuations at the promoter by varying the ratio of binding rate and synthesis rate, the adiabaticity parameter $\omega = \tau^+/\alpha$ (τ^- is adjusted to keep τ^+/τ^- constant). Small ω leads to strong fluctuations, whereas large ω reduces fluctuations. They find that increasing ω decreases the average switching rate and therefore stabilizes the switch. This dependency vanishes for $\omega > 5$, where the average switching rate remains constant. The latter is in accordance with our results in Eq. (3.24), where the mean residence time depends only on the ratio of τ^+ and τ^-, not on the absolute values and is therefore independent of ω. The dependency of the average switching rate for $\omega < 5$ is not predicted by Eqs. (3.23) and (3.24). It is also not visible in the stochastic simulations, where mean residence times of systems with $\omega = 1$ and $\omega = 20$ coincide (Fig. 3.5). The results of Morelli et al. (2008) were simulated for an average number of

proteins $\bar{N}_{\mathrm{A}} = \bar{N}_{\mathrm{B}} = 27$. As mentioned above, in regions of very small protein numbers the system might leave the dominating attractor by a mechanism not captured by Eqs. (3.23) and (3.24), probably causing the difference of the results of Morelli et al. (2008) and our results for small ω.

3.2 GMP differentiation dynamics explained by a toggle switch

In the previous section, we analyzed the dynamics of a toggle switch model and focused on the residence times of the system in the attractors and their dependence on parameters, such as protein degradation rates. In this section, we analyze single-cell time-lapse microscopy data of differentiating granulocyte/monocyte progenitors (Rieger et al., 2009), which report the timing of differentiation in individual cells. Assuming that this cell fate decision is implemented as a toggle switch, we fit a toggle switch model to the observed differentiation dynamics and infer the parameters of this switch.

Following the current paradigm, hematopoiesis can be pictured as a tree, made up from a concatenation of branching decisions (Orkin and Zon, 2008). Starting from a hematopoietic stem cell (HSC), all mature blood cell types are generated via progenitor states, where the lineage potential is reduced in each differentiation step. To replenish e.g. granulocytes (G) and monocytes (M) – blood cells with important function in immune response and phagocytosis (Dahl, 2009) – a HSC differentiates to a multipotent progenitor (MPP), to a common myeloid progenitor (CMP), and to a granulocyte-monocyte progenitor (GMP), before the final lineage decision between G and M is made (see Fig. 3.8). While details of the hierarchical differentiation have been revealed along with the increasing specificity of cell state markers, the tight balance between blood cell numbers and the inter-regulation of the involved processes is far from understood.

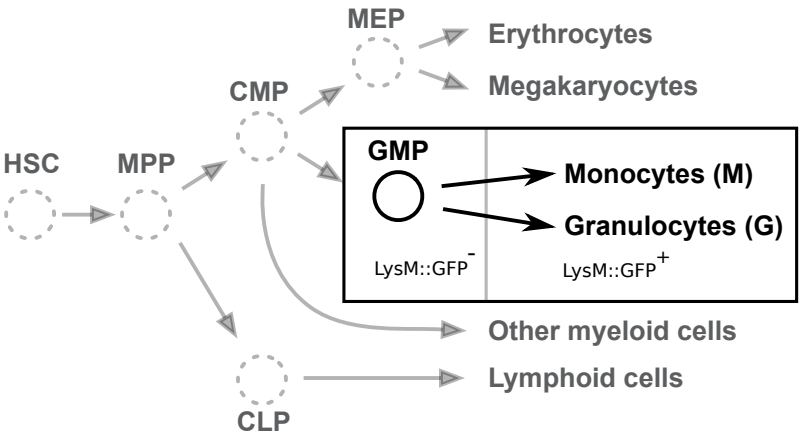


Figure 3.8: **The G/M lineage decision within hematopoiesis.** Granulocyte-monocyte progenitors (GMP) are bipotent and can differentiate into granulocytes (G) and monocytes (M). The LysM::GFP reporter indicates the loss of bipotency, i.e. GMPs are LysM::GFP negative, whereas both granulocytes and monocytes are LysM::GFP positive.

3.2.1 GMP differentiation probability from time-lapse microscopy data

In previous experiments (Rieger et al., 2009), GMPs have been sorted with a purity of over 95%. Sorted cells have been imaged, tracked, and analyzed for up to five days under

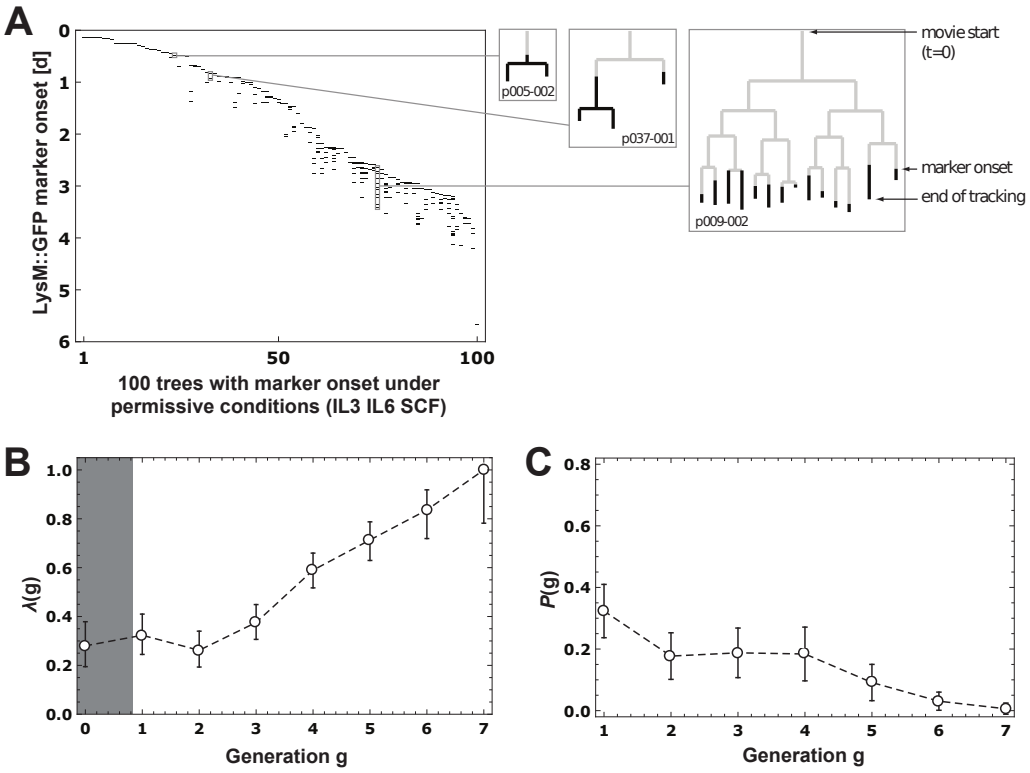


Figure 3.9: **Analysis of GMP genealogies from single-cell time-lapse microscopy.** A) LysM::GFP marks the loss of bi-potency. We show the onset time in days for 100 trees under permissive (IL3, IL6, SCF) culture conditions, allowing the differentiation of both G and M cells (data from Rieger et al., 2009). In each tree, multiple marker onsets can be observed, depending on the number of leaves. We observe more late and synchronous onsets than expected from the branching process with the parameters inferred from the colony assay data. In the three tree examples shown on the right, marker onset is indicated by the change from gray to black lines. Each tree is identified by its position in the movie and its tree number. B) We assume that LysM::GFP is an instant marker of differentiation. The differentiation probability $\lambda(g)$ from the genealogy data (open discs) depends on the generation within a tree. Error bars denote 95% confidence intervals and have been determined with the Clopper-Pearson method (Clopper and Pearson, 1934). C) The probability density for a cell to differentiate exactly in generation g, $P(g)$. For calculating the errors, we propagate the larger of the two confidence intervals. Note that cells in generation 0 have not been tracked from birth and a bias towards undifferentiated cells might apply. Thus, this data point is disregarded.

permissive conditions (IL3, IL6, SCF), allowing for both granulocytes and monocytes. To monitor the loss of the GMP state (see trees in Fig. 3.9), LysM::GFP mice (Faust et al., 2000), which express enhanced green fluorescent protein (GFP) from the lysozymeM gene locus as a marker for unilineage commitment, have been used in the experiments. The onset of a fluorescent signal in a cell is annotated in the movie data and marks its irreversible loss of bi-potency. All cells are tracked until marker onset, if they have not died or been lost during tracking before. After marker onset, tracking is normally stopped after a couple of timepoints, but occasionally is continued to the next generations (see Fig. 3.9A for three examples of tracked trees with marker annotation). The marker onset times for all cells in all trees with fluorescent signal (100 out of 143 trees) are shown in Fig. 3.9A.

To quantify the dynamics of the LysM::GFP marker and – assuming that LysM::GFP is an instantaneous marker of differentiation – the dynamics of differentiation, we estimate the probability of marker onset $\lambda(g)$ in generation g as

$$\lambda(g) = \frac{\text{LysM::GFP onsets in generation g}}{\text{Total number of cells in generation g}} \,.$$

We observe an increase of λ with generation g (Fig. 3.9B). If we disregard generation 0 (where cells have not been tracked from birth but from the movie start and thus a bias towards LysM::GFP negative cells might apply) λ rises from about 35% in generations 1, 2 and 3 to 100% in generation 7, where all cells differentiate (see Fig. 3.9B).

$\lambda(g)$ is the instantaneous probability to differentiate in generation g. From $\lambda(g)$, we can also calculate the overall probability mass function $P(g)$ of observing a differentiating cell in generation g:

$$P(g) = \lambda(g) \prod_{g'=0}^{g-1} [1 - \lambda(g')] \,.$$

The observed differentiation probability $\lambda(g)$ results in a broad probability density $P(g)$ for a cell to differentiate in g (Fig. 3.9C). For example, the probability to differentiate in generation 5 is still well above 5%.

The differentiation probability $\lambda(g)$ defined above describes the probability of a cell to differentiate in generation g given it reaches generation g. In biomedically motivated survival analysis and reliability theory in engineering, analogous concepts are called hazard function and failure rate, respectively (see, e.g. Lee and Go, 1997 for a review). There, time-dependent hazard functions and failure rates emerge quite naturally from, e.g. aging or erosion. In the case of GMP differentiation, a time-dependent differentiation probability can occur for a number of reasons: The medium conditions might change over time, cell-cell signaling might impact as the cell density grows, or an inherent program might increasingly force the cells to differentiate. In the following we study a mechanistic, molecular model of the differentiation process and show that a time-dependent differentiation probability emerges naturally in this model.

3.2.2 Molecular toggle switch model

The molecular details of GMP differentiation are still under debate. The most detailed contribution comes, to the best of our knowledge, from Laslo et al. (2006). Here, the

authors proposed a mutual antagonism between Gfi-1 and the integrated monocytic factor EgrNab (consisting of the genes Egr-1, Egr-2 and Nab which have redundant molecular functions) to mediate the lineage choice of GMPs. Previous analyses suggested a pivotal role of PU.1 and C/EBPα, which are both required for the generation of GMPs (Iwasaki et al., 2005; Dakic et al., 2005). One hypothesis links the ratio between PU.1 and C/EBPα to a primary cell fate decision (Dahl et al., 2003), and there is also evidence for other factors being involved in the differentiation process (see Dahl, 2009 for a comprehensive review and Krumsiek et al., 2011 for a meso-scale model). While it has been unambiguously shown that cytokines can instruct the lineage decision (Rieger et al., 2009), the intrinsic toggle switch seems to be determined by the antagonistic players Gfi-1 and EgrNab (Laslo et al., 2008).

Assuming that two antagonistic players control the intrinsic GMP lineage decision that drives differentiation towards granulocytes and monocytes under permissive conditions, we set up a chemical reaction kinetics model of a toggle switch involving the antagonists G (potentially Gfi-1) and M (potentially EgrNab), that promote the granulocyte and monocyte lineage, respectively. We describe the mutual inhibition of these two genes using a one-stage model of gene expression with mutual inhibition being realized as DNA-protein binding (see Fig. 3.10A). Analogous to the previous section, the model can be represented as a set of biochemical reactions for G and M, respectively, and a set of reaction rates. The seven reactions describing the synthesis and binding of the player G, with symmetric reactions (but potentially different rates) for the player M are:

$$\mathrm{DNA_G} \xrightarrow{\alpha_\mathrm{G}} \mathrm{DNA_G} + \mathrm{Protein_G} \tag{3.28}$$

$$\mathrm{Protein_G} \xrightarrow{\delta_\mathrm{G}} \emptyset \tag{3.29}$$

$$2 \cdot \mathrm{Protein_G} \xrightarrow{\kappa_\mathrm{G}^+} \mathrm{Dimer_G} \tag{3.30}$$

$$\mathrm{Dimer_G} \xrightarrow{\kappa_\mathrm{G}^-} 2 \cdot \mathrm{Protein_G} \tag{3.31}$$

$$\mathrm{Dimer_G} \xrightarrow{\delta_\mathrm{G}} \emptyset \tag{3.32}$$

$$\mathrm{Dimer_G} + \mathrm{DNA_M} \xrightarrow{\tau_\mathrm{G}^+} \mathrm{DNA_M^{bound}} \tag{3.33}$$

$$\mathrm{DNA_M^{bound}} \xrightarrow{\tau_\mathrm{G}^-} \mathrm{Dimer_G} + \mathrm{DNA_M} \tag{3.34}$$

Reaction (3.28) corresponds to protein synthesis from an unbound promoter. Transcription and translation are aggregated into a single reaction. In general, one should consider extending the above system by explicit transcription and translation reactions, if no clear separation of time-scales applies. In addition to being more detailed, such a two-stage gene expression model (Shahrezaei and Swain, 2008) can induce two undecided and two decided attractors (Strasser et al., 2012). Here, we choose the one-stage model not only for the sake of simplicity but also because the simpler model reflects the type of observed data more naturally, where we are unable to discriminate between one or two undecided GMP states.

Proteins can either degrade (reaction 3.29) or form homodimers (reaction 3.30). Homodimers dissociate into two monomers (reaction 3.31) or are degraded (reaction 3.32).

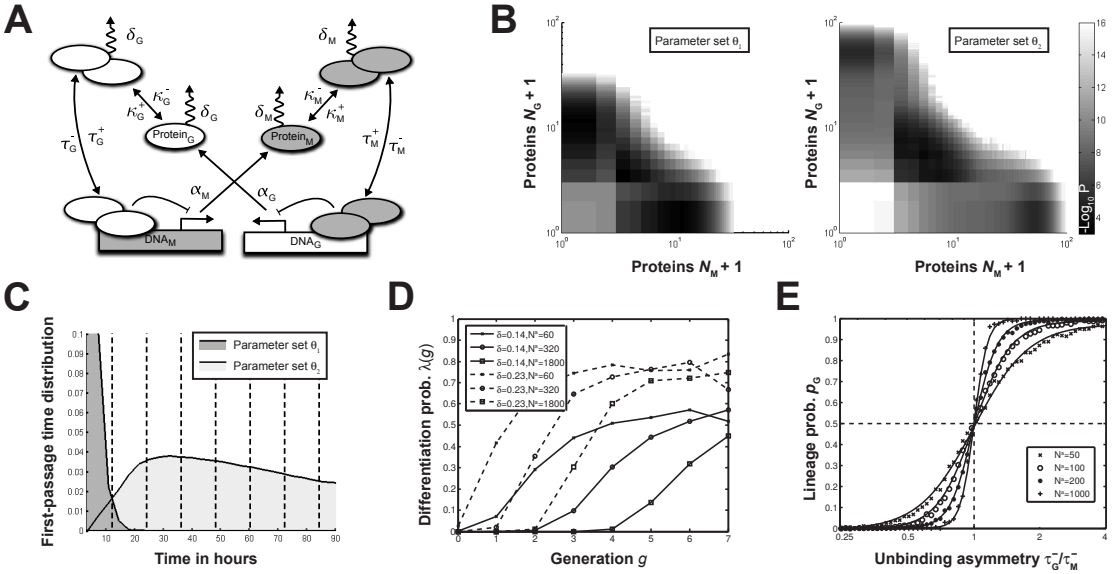


Figure 3.10: **Molecular toggle switch model.** A) We abstract the complex interactions between monocytic and granulocytic factors to two mutually inhibiting players, in the following referred by the indices G and M. $\mathrm{DNA_G}$ is transcribed to $\mathrm{Protein_G}$, which can bind in a dimerized form to the promoter at $\mathrm{DNA_M}$ and inhibit its transcription. The same reactions apply for M. B) We plot the resulting state space with three clearly discernible attractors appearing as regions with high probability (black) for the two parameter sets θ_1 and θ_2. The eight-dimensional state space is projected onto the N_G, N_M plane, showing the total amount of either protein in the system. The central attractor represents the undifferentiated GMP, while the distant attractors represent differentiated cells. Color coded is the $-\mathrm{Log}_{10}$ probability of the steady state. C) Probability distributions of the first-passage time from the central to one of the two decided attractors for the two parameter sets θ_1 and θ_2 used in B). Dashed vertical lines represent the average cell cycle length of GMPs, $t_\mathrm{cc} \approx 12h$, as inferred from the genealogies. D) Generation dependence of the differentiation probability $\lambda(g)$ for different parameters of the molecular model. The protein decay rate δ and the total protein amount of dominating player in the decided attractor, N^a, determine the position and the slope of the onset of λ. E) Differentiation bias p_G as a function of the binding strength of the granulocytic homodimer G. For readability, Hill-functions were fitted to the data. The sensitivity of the switch to asymmetries in the binding rates increases with the number of proteins in the system. All parameter sets used in the figure can be found in Table 3.2.

Note that we assume that the homodimers are degraded with the same rate as monomers. The last two reactions (3.33) and (3.34) describe the binding and unbinding of a homodimer to the antagonistic gene and thereby the transition from an active to an inactive promoter and vice versa. Bound DNA lacks the ability to be transcribed and only dimers can inactivate the promoter. Note that cooperative binding via dimerization is solely included to stabilize the system: Non-cooperative one-stage switches quickly forget a differentiation decision (Loinger et al., 2007), leave the decided attractor and are therefore inadequate in the context of lineage choice.

The above reactions specify the possible transitions between all states $x \in (\mathbb{N}_0)^8$ in the state space of species abundances. The master equation (Van Kampen, 1992) describes the time evolution of the probability for being in state x, $\mathcal{P}(x)$ via

$$\frac{d\mathcal{P}(x)}{dt} = \sum_{x'} \left[w_{xx'}\mathcal{P}(x') - w_{x'x}\mathcal{P}(x) \right] \ ,$$

with transition rates $w_{xx'}$ between states x and x'. Again, the master equation corresponding to the reaction system (3.28)–(3.34) cannot be solved analytically, but it can be simulated using Gillespie's algorithm (Gillespie, 1976). After a transient time, the simulation leads to a quasi-steady state, where the probabilities in state space do no longer change, $\frac{d\mathcal{P}(x)}{dt} = 0$. For appropriate parameters (see Table 3.2), we find that the molecular model (Fig. 3.10A) induces three regions of high probability in state space, the attractors of the system (see Fig. 3.10B). In the spirit of Waddington's landscape (see chapter 1 and Waddington, 1957) one can associate different cell fates with these attractors. The central attractor represents the undifferentiated GMP state, as both lineage determining factors are present only in small amounts, neutralizing each other. The two distant attractors represent the differentiated granulocyte and monocyte cells fates, where the corresponding lineage determining factor is abundant, whereas its antagonist is completely absent. Henceforth, these are called decided attractors.

The two parameter sets θ_1 and θ_2 (Table 3.2) result in different steady state distributions, as shown in Fig. 3.10B, and accordingly different first-passage times, as shown in Fig. 3.10C. The first-passage time is defined as the time required to leave the central attractor and reach any of the decided attractors of the system (Walczak et al., 2005a), and can be calculated from the simulations of the toggle switch model , where the attractor boundaries are defined analogous to the previous section (see Eq. 3.22). Fig. 3.10C depicts the first-passage time distribution for the two parameter sets θ_1 and θ_2, resembling an exponential distribution and a Γ-distribution, respectively. These are two common distributions for first-passage times in stochastic systems (Bel et al., 2010). Let us study the origin of these different distributions with respect to the parameters of the system: For the parameter set θ_1 the central attractor is narrow and the two decided attractors are close by in terms of protein numbers (see Fig. 3.10C). Through fluctuations in protein numbers the system is quickly pushed out of the central attractor. Once it has left the central attractor, only few synthesis reactions are required to settle in one or the other decided attractor. Therefore, the transition time between attractors is negligible, leading to an overall exponential first-passage time distribution, whose parameter depends on the high rate of escape from the central attractor. For parameter set θ_2 the central attractor is wider and therefore more time passes until the system by chance leaves the central

attractor. Additionally, significant time is needed to bridge the distance to the decided attractors, which is larger than in parameter set θ_1 (note the log-scale in Fig. 3.10B). The small rate of escape leads to the long tail of the first-passage time distribution of parameter set θ_2 in Fig. 3.10C, whereas the time spent to move in between attractors causes the shift of the distribution to the right. The distribution in this case resembles a Γ-distribution, which is characteristic for random walks over long distances with a bias in one direction (Bel et al., 2010).

Let us interpret the first-passage time distributions observed in Fig. 3.10C in the context of differentiation probability: The exponentially distributed first-passage time corresponds to a time-independent differentiation probability $\lambda(g) = \lambda$, as the exponential distribution is memoryless. If however the first-passage time follows a non-exponential distribution (as seen in Fig. 3.10C) the differentiation probability is time-dependent and the simple branching process cannot reflect this property. Our model therefore can induce either time-dependent or independent differentiation probabilities, as determined by the molecular rates of the switch. It shows how these two very different scenarios of differentiation can be traced down to the same molecular origin, the first-passage time.

Next we investigate more systematically how the time-dependence of λ relies on the choice of parameters of the molecular model. Note that in the following we will only consider symmetric systems, that is, the synthesis, degradation, binding and unbinding rates for both player are the same ($\delta_G = \delta_M = \delta$, etc.). Fig. 3.10D shows $\lambda(g)$ for varying degradation rates δ and $N^a = \alpha/\delta$, representing the protein levels of the dominating player in the decided attractor (see supporting information for a detailed discussion). All curves show similar characteristics: After some transient time, the curves grow almost linearly before asymptotically approaching an upper bound. The higher the protein levels, the later is the onset of growth in the curves. This is intuitive as higher protein levels imply larger distances between the attractors, which sets a minimal time before the system is able to reach the decided attractors. The probability to observe a decided system before this time is 0. The different asymptotics of the curves can be attributed to the degradation rate, which sets the timescale of the system. For constant protein level N^a, a high degradation rate δ implies a high synthesis rate α as $N^a = \alpha/\delta$. This speeds up the dynamics of the whole system, therefore increasing the fraction of simulations that escape from the central attractor per unit time, which is in fact $\lambda(g)$.

Note that we disregard the tree structure of the genealogies in our molecular model, and instead simulate single branches, corresponding to the time series of a single cell and its ancestors. However, by calculating the fraction of cells reaching a decided attractor within the time window of one generation, we can calculate the differentiation probability λ as a function of the generation g and thus establish a one to one correspondence between the first-passage time in the molecular model and the probability density $P(g)$ in the branching process.

By adjusting the parameters of the model, we can also simulate a biased differentiation towards the one or the other lineage. To test this, we systematically changed the binding strength of the G homodimer ($\tau_{\mathrm{G}}^{+}/\tau_{\mathrm{G}}^{-}$) while keeping the binding strength of the M homodimer constant. All remaining parameters are kept symmetric. In Fig. 3.10E, we show how the probability p_{G} for the granulocyte lineage changes in response to varying binding strength. Intuitively, since both binding strengths are equal, the system has equal

probability of differentiating towards the G or the M attractor. Increasing the binding strength of G leads to stronger repression of M, giving G a slight advantage in the battle of the two factors. Similarly, a decrease of binding strength leads to a disadvantage of G and and the probability to commit towards M is increased. Interestingly the response of p_{G} to binding strength is influenced by the amount of proteins in the decided states N. For low protein numbers small differences in binding strength of the two players still result in a balanced decision towards either G or M ($p_{\mathrm{G}} \approx 0.5$). For higher protein numbers even small differences in binding strength will destroy the balance between the two factors and the favored player will prevail almost certainly. This sensitive response is interesting in the light of lineage instruction: In cytokine medium, G-CSF and M-CSF are able to instruct differentiation with a high reliability (Rieger et al., 2009).

Parameter	α_{G}, α_{M}	δ_{G}, δ_{M}	κ^+_{G}, κ^+_{M}	κ^-_{G}, κ^-_{M}	τ^+_{G}, τ^+_{M}	τ^-_{G}, τ^-_{M}
Unit	s^{-1}	s^{-1}	$s^{-1}\mathrm{Protein}^{-1}$	s^{-1}	$s^{-1}\mathrm{Protein}^{-1}$	s^{-1}
Set θ_1	$8.0 \cdot 10^{-3}$	$1.0 \cdot 10^{-3}$	1.0	0	1.0	0.1
Set θ_2	$2.5 \cdot 10^{-3}$	$5.0 \cdot 10^{-5}$	1.0	0	1.0	0.1

Table 3.2: Parameter sets of the molecular model used in Fig. 3.10B,C. Subscripts identify the respective player (G or M) in the toggle switch model, superscripts specify binding (+) and unbinding (-) reactions. While the dimerization rates κ and the DNA-protein binding rates τ are identical in both sets, we varied protein synthesis rates α and protein decay rates δ. Note that here, all rates are symmetric with regard to the player, e.g. $\delta_{\mathrm{G}} = \delta_{\mathrm{M}}$.

3.2.3 Bayesian parameter inference identifies a scale separation of decay rates

Having shown how different first-passage times and probabilities of commitment towards one or the other lineage emerge from a simple toy model of a toggle switch, one can now fit the model to observed quantities in order to estimate molecular rates. In the following we will show what can be learned from these quantities in terms of molecular parameters in a proof-of-concept way.

Analytic expressions of the first-passage time distribution and the lineage probabilities for the toggle switch are hard to derive and remain elusive. Therefore, we have to resort to Approximate Bayesian Computation (see chapter 2 and Toni et al., 2009) to infer molecular parameters from the observed differentiation and commitment probabilities. Approximate Bayesian Computation allows parameter inference even though no analytical expression for the likelihood of the data given the parameters is available. Instead of maximizing the intractable likelihood, Approximate Bayesian Computation searches for parameters that minimize a chosen distance function, thereby best fitting the data.

We fit the model parameters θ to the data with respect to the L_1 distance function

$$d((p^{\theta}_{\mathrm{G}}, \lambda^{\theta}), (p^{obs}_{\mathrm{G}}, \lambda^{obs})) = \frac{1}{2}\left(\frac{1}{7}\sum_{g=1}^{7}|\lambda^{\theta}(g) - \lambda^{obs}(g)| + |p^{\theta}_{\mathrm{G}} - p^{obs}_{\mathrm{G}}|\right) \tag{3.35}$$

where $\lambda^{obs}(g)$ is the observed differentiation probability in generation $g = 1 \dots 7$ (depicted in Fig. 3.9B). Similarly, $\lambda^{\theta}(g)$ is the differentiation probability obtained from simulations

with parameters θ. Finally, p_{G}^{obs} is the observed probability to differentiate to granulocytes and p_{G}^{θ} is its simulated counterpart. p_{G}^{obs} was obtained from colony assays (see Marr et al., 2012 for details) and is estimated as $p_{\mathrm{G}}^{obs} = 0.67$. The first term on the right side of Eq. (3.35) quantifies how closely the parameter θ can reproduce the observed differentiation probabilities, whereas the second term quantifies its match to the observed lineage bias. Note that the factors of $\frac{1}{2}, \frac{1}{7}$ in Eq. (3.35) are used to scale the distance to the interval $[0, 1]$. We used a standard SMC-ABC algorithm implemented by a customized version of the abc-sysbio-toolkit (Liepe et al., 2010) to fit the toggle switch model to the observed data.

We made a quasi-steady state assumption for the dimerisation and DNA-protein binding reactions to reduce the number of parameters. Assuming that dimerisation, protein-DNA and their respective reverse reactions are much faster than the other reactions of the system, one can condense the full system consisting of 14 reactions and eight species (Eqs. 3.28–3.34) into four reactions and two species:

$$\emptyset \xrightarrow{\hat{\alpha}_{\mathrm{G}}} \mathrm{Protein_G} \tag{3.36}$$

$$\mathrm{Protein_G} \xrightarrow{\delta_{\mathrm{G}}} \emptyset \tag{3.37}$$

$$\emptyset \xrightarrow{\hat{\alpha}_{\mathrm{M}}} \mathrm{Protein_M} \tag{3.38}$$

$$\mathrm{Protein_M} \xrightarrow{\delta_{\mathrm{M}}} \emptyset\,, \tag{3.39}$$

where the synthesis rates now are state dependent:

$$\hat{\alpha}_{\mathrm{G}}([N_{\mathrm{G}}, N_{\mathrm{M}}]) = \frac{\alpha_{\mathrm{G}}}{1 + (N_{\mathrm{M}}/K_{\mathrm{M}})^2}$$
$$\hat{\alpha}_{\mathrm{M}}([N_{\mathrm{G}}, N_{\mathrm{M}}]) = \frac{\alpha_{\mathrm{M}}}{1 + (N_{\mathrm{G}}/K_{\mathrm{G}})^2}$$

K_{G} and K_{M} are the dissociation constants of the DNA-protein interactions. K_{G} corresponds to the amount of $\mathrm{Protein_G}$ where $\mathrm{DNA_M}$ is bound half of the time (analogously for K_{M}). This reduces the number of unknown parameters to be estimated to six: two synthesis rates (α_{G},α_{M}), two degradation rates (δ_{G},δ_{M}) and two dissociation constants ($K_{\mathrm{G}} = \tau_{\mathrm{G}}^{-}/\tau_{\mathrm{G}}^{+}$, $K_{\mathrm{M}} = \tau_{\mathrm{M}}^{-}/\tau_{\mathrm{M}}^{+}$), all assigned with flat prior distributions.

We iterated 10 populations consisting of 200 parameter sets, where λ^{θ} and p_{G}^{θ} were estimated from 1000 repeated simulations for each parameter set θ. The last population contained only parameter sets with $d(\theta) < 0.1$, giving already a good fit to the data (see Fig. 3.11A). Afterwards, we calculate the probability density $P(g)$ from the simulated $\lambda(g)$ and find that apparently, moderate deviations in $\lambda(g)$ result in considerable deviations in $P(g)$ (see Fig. 3.11B). Thus, we calculate the distance $d(P^{\theta}, P^{obs}) = \sum_{g=1}^{7} |P^{\theta}(g) - P^{obs}(g)|$ (bounded to $[0, 2]$)for each parameter set to quantify its goodness of fit within the obtained posterior distribution.

We find asymmetric protein degradation rates for the best fits to the experimental data (black dots in Fig. 3.11C): a high δ_{G} implies a low δ_{M}, and vice versa. In contrast to the protein degradation, synthesis rates appear mostly correlated (see Fig. 3.11D). Therefore, the best fits result in systems where the degradation rate of one player might

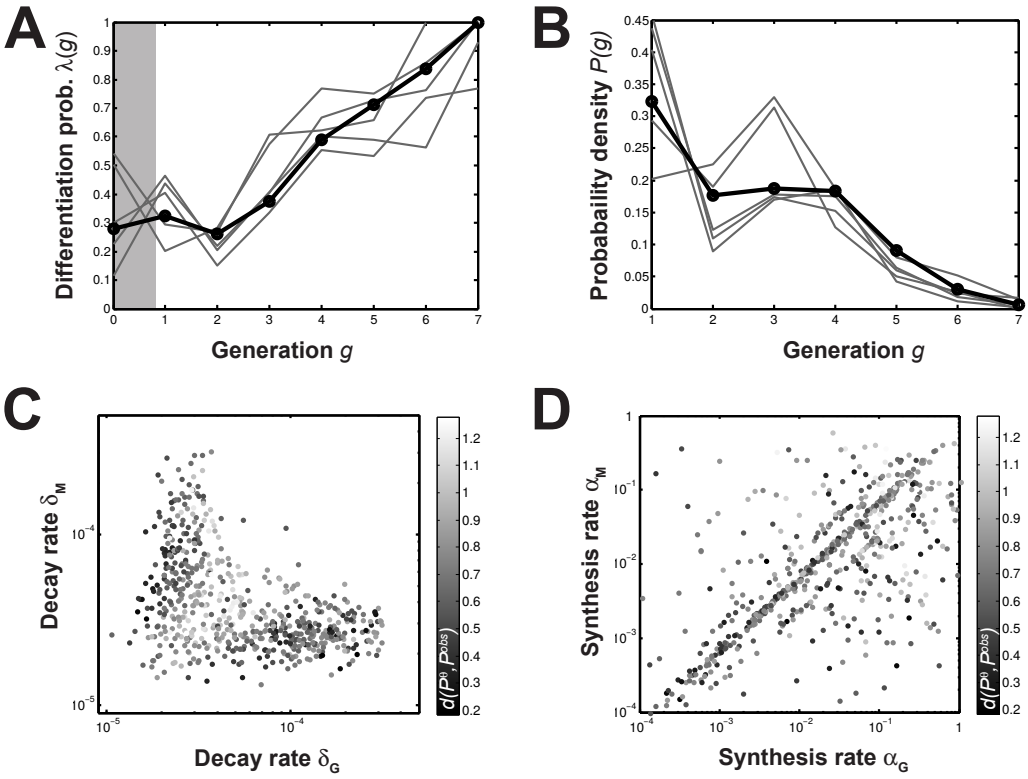


Figure 3.11: **Inference of model parameters with Approximate Bayesian Computing.** A) Fit of the best five parameter sets to the experimentally observed $\lambda(g)$ (black). B) The corresponding probability densities $P(g)$ still show larger deviations from the observed data due to their high sensitivity on small errors in the fit to λ. C,D) Scatter plots of the degradation (C) and synthesis (D) rates present in the last iteration of ABC-SMC (population size was increased from 200 to 600 via resampling). The distance $d(P^\theta, P^{obs})$ of each particle θ to the observed probability densities $P(g)$ as shown in B) is color coded, where black corresponds to small distance. Interestingly, best fits show a split of the decay rates of the two players, implying that one differentiates quickly while the other player needs more time to reach the decided attractor.

be c times that of the other player (e.g. $\delta_{\mathrm{G}} = c\ \delta_{\mathrm{M}}$). At the same time, the number of proteins separates in an inverse fashion ($N^a_{\mathrm{G}} = 1/c\ N^a_{\mathrm{M}}$). This separation of scales induces qualitatively different $\lambda(g)$ curves for the two players: with regard to the above example, for player G not only the slope in $\lambda(g)$ is decreased due to a lower δ_{G}, but also the onset of λ is shifted due to the higher N^a_{G}.

Note that, as shown in Fig. 3.10D , the degradation rate δ but not the synthesis rate α determines the dynamics of the system, which can also be shown theoretically: Assuming a simple gene expression system (a birth death process), we want to calculate the time to reach its protein steady state level given some initial condition $p(t = 0)$ (which corresponds to the edge of the undecided attractor in the toggle switch).

We ask for the speed of convergence of this process to the decided attractor. The speed of convergence, r, of a stochastic process X can be measured via (Coolen-Schrijner and Van Doorn, 2001):

$$r(X) = \inf\left\{r > 0 \mid M - E[X(t)] = O\left(e^{-\frac{t}{r}}\right)\right\} ,$$

where $E[X(t)]$ is the expectation value of the process at time t and $M = \lim_{t\to\infty} E[X(t)]$. For the simple birth death process considered here:

$$\begin{aligned} M &= \alpha/\gamma \\ E[X(t)] &= \frac{\alpha}{\delta} + \left(E[X(0)] - \frac{\alpha}{\delta}\right) \cdot e^{-\delta\cdot t} \\ M - E[X(t)] &= (\alpha/\delta - E[X(0)]) \cdot \exp(-\delta t) . \end{aligned}$$

Therefore,

$$r(X) = \delta^{-1} .$$

Another measure for convergence (Coolen-Schrijner and Van Doorn, 2001) is

$$m(X) = \int_0^\infty \left(1 - \frac{E[X(t)]}{M}\right) dt ,$$

which for the simple birth death process evaluates to

$$m(X) = \frac{1 - C \cdot (\frac{\alpha}{\delta})^{-1}}{\delta} .$$

Since $\frac{\alpha}{\delta} \gg C$,

$$m(X) \approx \delta^{-1} .$$

Overall, we find that the transition time from a low expression state (undecided attractor) to steady state expression (decided attractors) scales with the inverse of the decay rate in a stochastic description of the system.

Summarizing, the differentiation probability $\lambda(g)$, as observed in the time-lapse data, is thus best fitted by a system where one player decides quickly, while the other takes longer to execute the differentiation decision. Interestingly, both players can act as the slow or the fast species in the system, implying that the asymmetry in degradation rates is independent of the lineage bias. That is, the preference of one cell fate ($p_{\mathrm{G}} > p_{\mathrm{M}}$) does not induce the separation of scales in the differentiation dynamics, but only the shape of $\lambda(g)$.

Chapter 4

Phenomenological models of cell fate choice

In chapter 3, we studied the regulatory motives of genetic toggle switches as a mechanistic implementation of cell fate choice in individual cells. Additionally, we observed how the phenomenon of a time (generation) dependent differentiation probability λ emerges from these internal mechanisms, thus providing further abstraction of the underlying system.

In the present chapter, we continue at this level of abstraction and thus depart from mechanistic models of cell fate choice. For many cell fate decisions, regulatory mechanisms are still unknown. Furthermore, cell fate decisions may not be cell autonomous, but depend on external signals e.g. the micro-environment ("niche", Wang and Wagers, 2011), cell-cell communication (Shalek et al., 2014) or cytokines (Rieger et al., 2009). Hence, it is useful to study cell fate decisions on a phenomenological level, i.e. without mechanistic details, and in the context of an entire cell population.

Cell fate decisions, in the following referred to as "cell state transitions" can be observed in their spatiotemporal context using single cell time-lapse microscopy combined with cell tracking and image processing. However, the experimental data cannot immediately provide an explanation why the state transition occurs. For example, the differentiation rate of a stem cell towards a more mature cell type may depend on time (chapter 3, Marr et al., 2012), the makeup of surrounding niche cells (Morrison and Spradling, 2008), cell density (Lorincz, 2006), or on a combination of these features. While it is possible to quantify the emergence of cellular patterns in colonies (Scherf et al., 2012; Shivanandan et al., 2013), it is impossible to tell from the mere observation if the simultaneous differentiation of multiple cells is a random event or if it is triggered by, e.g., the increased density in the colony. The inference of features influencing the differentiation rate requires sufficient statistics for analysis, and thus a large number of cellular genealogies, which is in particular for mammalian systems still a challenging and labor intensive task (Schroeder, 2008; Amat et al., 2014). Thus, careful experimental design is indispensable.

In the following, we present a model and analysis framework that can infer the spatiotemporal features influencing state transitions and also allows to estimate the number of cellular genealogies required for this analysis. To validate the performance of our framework, we first simulate cellular genealogies from a generative spatiotemporal model for different scenarios of transition rate dependencies. We then develop an inference method

based on generalized linear models and feature selection with L_1 regularization. We show that our method is able to correctly identify the differentiation rate as a multi-feature function and determine the number of required genealogies and allowed tracking errors for different scenarios. Finally, we use the correlations between cell siblings to validate the chosen approach and detect shortcomings – either due to non-considered features, or due to cell-internal effects that drive cell state transitions.

4.1 A generative framework for synthetic spatiotemporal cellular genealogies

In the following, we present the basic model for cell state transitions and introduce four different scenarios of transition rate dependencies.

4.1.1 General model

Throughout this chapter, we use a simple model of cell state transition with two cellular states A and B (Fig. 4.1A). A single cell is defined by its 2D spatial coordinates $x \in \mathbb{R}^2$, its state $s \in [0, 1]$, where $s = 0$ ($s = 1$) if the cell is in state A (B) and its age τ, i.e. the time since the last division. A single cell in state A (black circle in Fig. 4.1A) can divide into two cells in state A, or transition into another state B (cyan circle), where it can only divide. The transition rate $\lambda(t, F_i(t))$ of a cell i depends on the features F_i of the cell. Notably, the features F, like time, cell cycle state, position or local cell density, can change over time. Specific examples of the function $\lambda(t, F_i(t))$ are introduced later on. Furthermore, the cell moves in 2D space modeled by Brownian motion. The division rate $\gamma(\tau)$ is age dependent to account for non-exponential lifetime of cells[1]. Note that these non-exponential waiting times usually render the system non-Markovian. However, since we introduced cell age as state variable of our system, the process remains Markovian.

The system evolves probabilistically in time and has to be described by a Master Equation, whose derivation we now briefly sketch. If there were only a single cell present and no cell division possible, the probability distribution $\mathcal{P}(x, s, \tau, t)$ of finding at time t a cell at location x, state s and age τ evolves as:

$$\begin{aligned}\dot{\mathcal{P}}(x, s, \tau, t) = \nabla^2 \mathcal{P}(x, s, \tau, t) + \frac{\partial}{\partial \tau}\mathcal{P}(x, s, \tau, t) \\ - \delta_{s,0} \cdot \lambda(x, t, \tau) \cdot \mathcal{P}(x, s, \tau, t) \\ + \delta_{s,1} \cdot \lambda(x, t, \tau) \cdot \mathcal{P}(x, s - 1, \tau, t) \ ,\end{aligned}$$

where $\delta_{n,m}$ is the Kronecker delta and $\nabla^2 = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2}$. The first term on the right hand side accounts for spatial diffusion of the cell in and out of location x, the second term accounts for aging, the third term accounts for cells transitioning out of state A ($s = 0$) and the fourth term describes cells transitioning into state B ($s = 1$). The transition rate can depend on features of the cell (e.g. its spatial position). As initial conditions we choose $\mathcal{P}(x, s, \tau, 0) = \delta_{x,x_0} \cdot \delta_{s,0} \cdot \delta_{\tau,0}$, i.e. a cell at location x_0 in state a and age 0.

[1] constant γ would yield unrealistic exponential lifetimes

In order to model a growing population of cells that interact with each other, we must derive evolution equations not only for a single cell, but for pairs of cells, triples of cells, etc., resulting in an infinite set of equations if the cell population size is unlimited. These equations will be coupled: For example, a division event in the single cell equation will add to the probability of the cell-pair equation (see Appendix A for the single and pair equations). Although solving these equations is beyond the scope of this thesis, we note that similar equations arise in Quantum Field Theory and are subject to theoretical investigation also in biological context (Birch and Young, 2006; Dodd and Ferguson, 2009).

Analogous to the previous chapter, instead of solving the equation, we simulated realizations of the underlying stochastic process (Fig. 4.1B): Since the system has continuous (space) and discrete (cell state) variables, a standard stochastic simulation algorithm cannot be applied and a hybrid simulation method must be used (see e.g. Haseltine and Rawlings, 2002). Cell position is treated as Brownian motion (potentially with drift) and is updated via an Euler-Maruyama scheme (Fuchs, 2013). To evolve the cell state in time for a single cell in state A, the simulation proceeds in small time steps Δt, during which a state transition event takes place with probability

$$P_i(t) = 1 - e^{-\int_t^{t+\Delta t} \lambda(t', F_i(t'))dt'} \approx 1 - e^{-\lambda(t, F_i(t))\cdot\Delta t}$$

for some arbitrary, state and time-dependent transition rate $\lambda\left(t, F_i(t)\right)$. The rate λ is evaluated at the beginning of each iteration, and the time step Δt is chosen sufficiently small[2]. The cell divides after 12 hours on average, corresponding to the typical lifetime of mammalian stem and progenitor cells (Buggenthin et al., 2013; Rieger et al., 2009) (for simplicity, but without loss of generality, we assumed cell lifetime to follow a uniform distribution: $t_{\text{div}} \sim \text{Uniform}([10h, 14h])$). The cell division replaces the dividing cell by two daughter cells, with positions close to that of the mother cell and with the same cell state: e.g. a mother cell in state A gives rise to two daughters in state A. These cells are then simulated in parallel. Over the course of the simulation, a cellular genealogy with a distinct cell state pattern emerges (Fig. 4.1C). Genealogies are simulated for 100 hours ($8 - 9$ generations of cells) corresponding to the typical observation periods of long term time-lapse microscopy (Costa et al., 2011; Eilken et al., 2009; Rieger et al., 2009).

4.1.2 Local cell density

Local cell density is estimated using a kernel f that determines how much each cell contributes to the local density at a certain point x in space as a function of intercellular distance. We define the local cell density $\rho_i^f(t)$ of cell i at time t with respect to a kernel $f : \mathbb{R} \to [0, \infty]$:

$$\rho_i^f(t) = \sum_{j \neq i} f\left[d(x_i(t), x_j(t))\right] \ , \tag{4.1}$$

where $x_i(t)$ is the spatial coordinate of cell i at time t and $d(x_i, x_j)$ denotes Euclidean distance. In the simulations we use either a tophat kernel (Fig. 4.1E, upper panel) with

$$f(r) = I(r < R) \ , \tag{4.2}$$

[2]Such that no appreciable change in cell locations occurs and the rate λ is approximately constant.

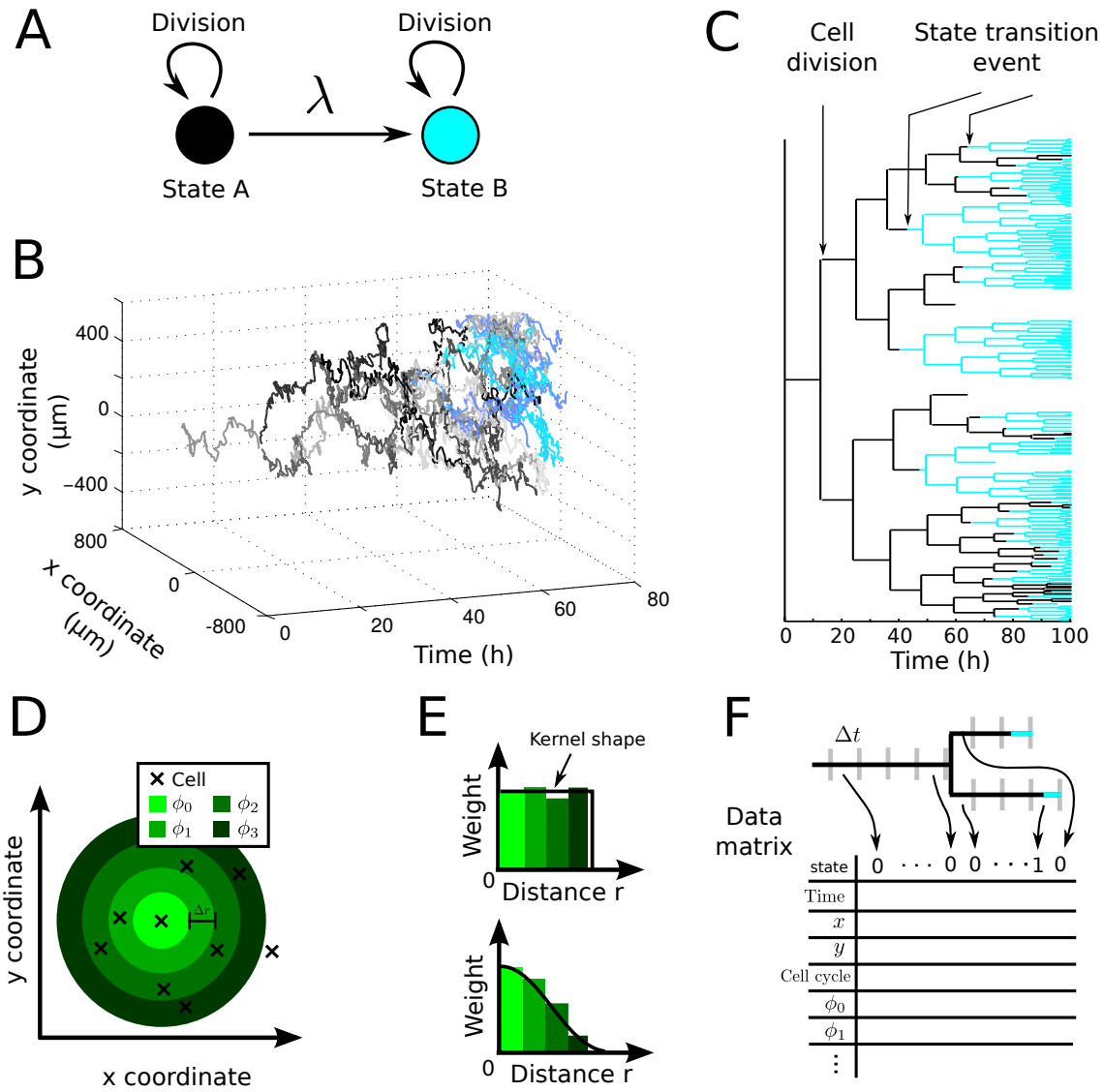


Figure 4.1: **Spatiotemporal simulation and analysis of cell state transitions.** A) In our model, a cell in state A (black) can divide or transition into state B (cyan). The transition is governed by the transition rate λ, which can depend on features like time, position, cell cycle, or the local cell density. B) Visualization of a cellular genealogy in space and time with cells in state A (black to gray) and state B (cyan to blue). C) Tree view of the genealogy depicted in B (coloring as in A). D) Local cell density is modeled via a set of annular basis functions ϕ_k with inner radii $k\Delta r$ and constant thickness Δr (green circles). Cells are indicated as crosses. E) Linear combinations of the ϕ_k can approximate any density dependence (e.g. a tophat kernel, upper panel, or a Gaussian kernel, lower panel). F) The tree structured data is transformed into a data matrix by discretizing time and creating one sample for each cell at each time interval, simulating a measurement process.

where $I(\ldots)$ is the indicator function of $[0, 1]$, or a Gaussian kernel (Fig. 4.1E, lower panel) with

$$f(r) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{r^2}{2\sigma^2}} \ . \tag{4.3}$$

For the tophat kernel each cell within distance R contributes equally to the local density experienced by cell i, whereas cells with distance larger than R do not contribute at all. For the Gaussian kernel the contribution to the local cell density decreases smoothly with distance.

4.1.3 Cell state transition scenarios

We create four datasets corresponding to different scenarios of cell state transition:

1. We consider a scenario where the transition rate is constant (λ constant), resembling spontaneous transitions independent of other effects:

$$\lambda(t, F_i(t)) = c \ , \tag{4.4}$$

with $c = 0.01\ h^{-1}$. Thus, at state transition in a cell with a 12 h lifetime will occur with $p = 0.11$.

2. For a time-dependent scenario ($\lambda \propto$ time), the transition rate is chosen as

$$\lambda(t, F_i(t)) = a \cdot t \ , \tag{4.5}$$

i.e. linearly increasing with time ($a = 3 \cdot 10^{-4}\ h^{-2}$). Note that λ does not depend on any other feature F of the cell.

3. For a density-dependent scenario ($\lambda \propto$ density), the local density of a cell i at time t is mediated by a tophat kernel (Eq. (4.2)) with $R = 300\ \mu m$ (roughly the distance a cell moves in its lifetime). The transition rate λ is then defined by

$$\lambda(t, \rho_i^{\text{tophat}}(t)) = b \cdot \rho_i^{\text{tophat}}(t) \ , \tag{4.6}$$

with $b = 0.002\ h^{-1}$.

4. For a time and density-dependent scenario ($\lambda \propto$ density + time), the contributions of the previous two factors are summed, using a Gaussian kernel (Eq. (4.3) with $\sigma = 30$) to define cell density:

$$\lambda(t, \rho_i^{\text{Gauss}}(t)) = a \cdot t + b \cdot \rho_i^{\text{Gauss}}(t) \ . \tag{4.7}$$

4.2 Inference framework

In this section, the methods to infer the transition rate from observed genealogies are presented.

4.2.1 Non-parametric estimation of the transition rate

The transition rate λ can be estimated non-parametrically by considering the definition of the rate as the probability of a transition in an infinitesimal time dt:

$$P(t, t + dt | F_i(t)) = \lambda(t, F_i(t)) \cdot dt \ , \tag{4.8}$$

where $P(t, t+dt|F_i(t))$ is the probability for a transition in the interval $[t, t+dt]$ in the presence of the features F.

We estimate the probability $P(t, t+dt|F)$ of a state transition in $[t, t+dt]$ given features F as

$$\hat{P}(t, t+dt|F) = \frac{\text{Number of transition events}|(t, F)}{\text{Number of cells in state A}|(t, F)} , \tag{4.9}$$

which is the fraction of candidate cells (in state A) that transit into state B in $[t, t+dt]$ having features F. After rearranging Eq. (4.8), we obtain

$$\hat{\lambda}(t, F) = \frac{1}{\Delta t} \cdot \frac{\text{Number of transition events}|(t, F)}{\text{Number of cells in state A}|(t, F)} \tag{4.10}$$

To measure the uncertainty of the estimates, we calculate Bayesian credibility intervals. When estimating the probability $P(t, t+dt|F)$ in Eq. (4.9), we in fact we estimate the parameter of a binomial distribution: we observed that k successes (events) occurred out of n trials and we are interested in the probability p of the success. Clearly, k follows a binomial distribution with parameters n, p:

$$k \sim \text{Binomial}(n, p)$$

Having observed a particular k and n we want to infer the parameter p of the underlying binomial distribution. In a maximum likelihood setting, one can show that this is just $\hat{p} = k/n$. Confidence intervals for this maximum likelihood estimator can be constructed according to various methods (Wald-, Wilson-, or Clopper-Pearson confidence intervals, Kendall and Stuart, 1967).

We invoke a Bayesian approach instead, calculating the posterior distribution $P(p|D)$ of p given the observed data $D = (n, k)$ which is related to the likelihood $\mathcal{L}(D|p)$ via Bayes theorem:

$$P(p|D) = \frac{\mathcal{L}(D|p) \cdot \pi(p)}{P(D)} , \tag{4.11}$$

where $\pi(p)$ is a prior distribution of p and $P(D) = \int \mathcal{L}(D|p) \cdot \pi(p)\ dp$ is the marginal distribution of the data. Here, the likelihood is the probability mass function of a binomial distribution:

$$\mathcal{L}((n, k)|p) = \binom{n}{k} p^k (1-p)^{n-k} .$$

We use a Beta-distribution as a prior, such that

$$\pi(p) = p^{\alpha-1}(1-p)^{\beta-1} \frac{1}{\text{B}(\alpha, \beta)} ,$$

where $\text{B}(\alpha, \beta)$ is the Beta-function and α, β are parameters that determine the shape of the prior distribution.

As the Beta-distribution is the conjugate prior of the binomial likelihood, the posterior in Eq. (4.11) can be calculated in closed form as:

$$P(p|D) = p^{\alpha+k-1}(1-p)^{\beta+n-k-1} \frac{1}{\text{B}(\alpha+k, \beta+n-k)}$$

This is again a Beta-distribution, but with new parameters $\alpha' = k + \alpha$ and $\beta' = n - k + \beta$, which shows how the observed data updates our prior knowledge about p in the posterior.

As no prior information on p is available, we choose an uninformative Jeffreys prior corresponding to $\alpha = \beta = 1/2$ (see Jeffreys, 1939), resulting in the following posterior distribution of p:

$$P(p|k,n) = p^{k-1/2}(1-p)^{n-k-1/2}\frac{1}{\mathrm{B}(k+1/2, n-k+1/2)} \,. \tag{4.12}$$

If we consider the posterior mean $\bar{p}$, we find

$$\bar{p} = \frac{k+1/2}{n+1} \,,$$

which is approximately the same as the maximum likelihood estimate (k/n) if k, n are large. We obtain 95% credibility intervals by calculating the 5% and 95% quantiles of the posterior in Eq. (4.12).

A particularly appealing property of Bayesian credibility intervals is that they will strictly be within $[0, 1]$, unlike their frequentist counterparts which are not constrained to the $[0, 1]$ domain of probabilities. This is especially prevalent if the estimated probability itself is close to 0 or 1.

4.2.2 Estimating the transition rate via GLMs

Estimating the transition rate via Eq. (4.10) becomes infeasible when the number of considered features becomes large and the number of observed data is limited: The uncertainty of the estimator $\hat{\lambda}$ in Eq. (4.10) becomes large as only few datapoints will be observed for each feature combination.

Instead, one can infer the transition rate parametrically using a machine-learning framework. We consider every timepoint of each cell as an observed sample $(F^{(i)}, Y^{(i)})$, where $F^{(i)}$ is a set of features measured for this sample (absolute time, time since last division, absolute spatial coordinates, and different measures of local cell density ϕ_k). We use superscripts to index the samples to clearly distinguish it from the per-cell indexing via subscripts used previously. $Y^{(i)} \in \{0, 1\}$ denotes the class label of the sample being either "state A" $(Y^{(i)} = 0)$ or "transition into B" $(Y^{(i)} = 1)$. A sample is considered as $Y^{(i)} = 1$ if a state transition occurred in the time interval of the sample. Timepoints after the state transition (either of the cell itself or its progeny) are discarded (Fig. 4.1F) since we are interested in what actually triggers the transition of cells, not the state of the cell itself. Counter-intuitively, all samples $(F^{(i)}, Y^{(i)})$ are independent, even though, e.g. adjacent samples typically are strongly correlated with respect to their features , as can be derived as follows:

Consider a single cell that is created at time t_0 and undergoes a state transition at time t_N. The likelihood of this whole cell in terms of the transition rate is

$$L = \lambda(t_N) \cdot e^{-\int_{t_0}^{t_N} d\tau \lambda(\tau)} \,. \tag{4.13}$$

If we however decide to split up this cell into N individual observations O_i we get the following likelihood for each observation:

$$\tilde{L}(O_i) = \begin{cases} e^{-\int_{t_{i-1}}^{t_i} d\tau\lambda(\tau)} & i \neq N \\ \lambda(t_N)e^{-\int_{t_{N-1}}^{t_N} d\tau\lambda(\tau)} & i = N \end{cases}$$

Assuming independence, the overall likelihood is

$$\begin{aligned} \tilde{L} = \prod_i L(O_i) &= e^{-\int_{t_0}^{t_1} d\tau\lambda(\tau)} \cdot e^{-\int_{t_1}^{t_2} d\tau\lambda(\tau)} \cdot \ldots \\ &\ldots \cdot e^{-\int_{t_{N-2}}^{t_{N-1}} d\tau\lambda(\tau)} \cdot \lambda(t_N)e^{-\int_{t_{N-1}}^{t_N} d\tau\lambda(\tau)} \\ &= e^{-\int_{t_0}^{t_1} d\tau\lambda(\tau) - \ldots - \int_{t_{N-1}}^{t_N} d\tau\lambda(\tau)} \cdot \lambda(t_N) \\ &= e^{-\int_{t_0}^{t_N} d\tau\lambda(\tau)} \cdot \lambda(t_N) \end{aligned}$$

Comparing this expression to Eq. 4.13, we find that they are actually the same. Therefore the assumption of independence is correct.

We use generalized linear models (GLMs, (McCullagh, J. Nelder, 1989)) to learn the relation between features $F^{(i)}$ and class labels $Y^{(i)}$ as

$$\mathbb{E}(Y^{(i)}|F^{(i)}, w) = \mu^{(i)} = g^{-1}(w^T F^{(i)}) \,,$$

where $\mu^{(i)}$ is the expected value of an exponential family distribution, g^{-1} is called the mean function, and w is the weights vector that has to be learned from the data.

Choosing a Bernoulli distribution and an exponential mean function would exactly correspond to our data generating process: In our simulations a single sample $(Y^{(i)}, F^{(i)})$ is created according to

$$\begin{aligned} Y^{(i)} &\sim \text{Bernoulli}(p^{(i)}) \\ p^{(i)} &= 1 - e^{-\lambda(F^{(i)})\cdot\Delta t} \\ &= 1 - e^{-w^T F^{(i)}\cdot\Delta t} \end{aligned}$$

If we switch class labels such that

$$\begin{aligned} V^{(i)} &= \begin{cases} 1, & Y^{(i)} = 0 \\ 0, & Y^{(i)} = 1 \end{cases} \\ V^{(i)} &\sim \text{Bernoulli}(q^{(i)}) \\ q^{(i)} &= 1 - p^{(i)} = e^{-w^T F^{(i)}\cdot\Delta t} \,, \end{aligned}$$

we see that the generative model corresponds to a GLM with a Bernoulli distribution and an exponential mean function

$$\mathbb{E}[V^{(i)}] = q^{(i)} = e^{-w^T F^{(i)}\Delta t} \,.$$

However, this specific GLM (known as logbinomial regression) has unfavorable numerical properties leading to convergence issues (Zou, 2004). Therefore, we resort to a GLM that has the desired exponential mean function but a Poisson instead of a Bernoulli distribution (also known as Poisson regression) and has better numerical properties. Note that Poisson regression is generally used to model count data (where $Y^{(i)} \in \mathbb{N}_0$), but is a good approximation to binary data ($Y^{(i)} \in \{0, 1\}$) in the case of rare events, since the probability mass function of Bernoulli and Poisson distribution are very similar if $p \ll 1$:

$$\begin{aligned} P_{\text{Ber}}(Y = 0|p) &= 1 - p \\ P_{\text{Ber}}(Y = 1|p) &= p \\ P_{\text{Poi}}(Y = 0|p) &= e^{-p} \approx 1 - p \\ P_{\text{Poi}}(Y = 1|p) &= p \cdot e^{-p} \approx p(1 - p) \approx p \end{aligned}$$

Thus, we obtain the following log-likelihood from Poisson regression:

$$\log p(Y|F, w) = \sum_i \left[Y^{(i)} w^T F^{(i)} - e^{w^T F^{(i)}} - \log(Y^{(i)}!) \right] .$$

4.2.3 Feature selection via L_1 regularization

To determine the relevant features of the transition rate and to exclude features that only indirectly influence the state transition (as e.g. for scenario 3 with a density dependent λ, where however λ also indirectly depends on time; see Fig. 4.2C, D and section 4.3.1), we apply L_1 regularization to the GLM, also known as Lasso (Tibshirani, 1996), where one minimizes the following function with respect to the weights w:

$$\begin{aligned} g(w) &= -\log p(Y|F, w) + \kappa \cdot \|w\|_1 \\ &= \sum_i \left[Y^{(i)} w^T F^{(i)} - e^{w^T F^{(i)}} - \log(Y^{(i)}!) \right] + \kappa \cdot \|w\|_1 , \end{aligned} \tag{4.14}$$

with $\|w\|_1 = \sum_i |w_i|$. This regularization is equivalent to placing a Laplace prior with location parameter $m = 0$ and scale parameter $b = \kappa^{-1}$ on the weights (Murphy, 2012), resembling our knowledge that most of the weights should be zero and the resulting model should be sparse.

Note that ideally, one should use L_0 regularization, which penalizes only the presence of a feature, but not the magnitude of its coefficient as does the L_1 regularization. However, L_0 regularization is intractable computationally, as the objective function is non-smooth and hence difficult to minimize. Instead one has to resort to L_1 regularization as an approximation (for a detailed discussion on sparsity, see Murphy, 2012, chapter 13).

Depending on the chosen regularization strength κ, one obtains models of differing sparsity (Fig. 4.3A). We follow the standard approach to determine the optimal regularization parameter κ^*: for each κ, we perform ten-fold cross validation using the deviance of the model as the error criterion and choose κ^* based on the 1SE rule (Hastie et al., 2009): We select the largest κ (hence the simplest model) that in terms of its deviance is still within one standard error of the best κ. Optimization and cross validation of Lasso is

performed using the function `lassoglm()` from the Matlab Statistics Toolbox, which uses a coordinate descend algorithm for optimization (Friedman et al., 2010).

Additionally, we have to account for the fact that the classes in our dataset are severely imbalanced with more non-events than events (at a ratio of 1:200 in our simulations). Such class imbalance can lead to problems for machine learning algorithms (He and Garcia, 2009). Therefore, we down-sample the majority class ($Y^{(i)} = 0$) to achieve a ratio of 1:3, yielding a good trade-off between class balance and number of overall samples. Feature selection using Lasso is applied to this down-sampled dataset via Eq. (4.14). Since down-sampling intentionally discards data and Lasso feature selection is sensitive to data perturbation (Murphy, 2012), we repeat the procedure $N = 50$ times, each time using a different sample of the majority class, combining it with the minority class and fitting the Lasso to this dataset. This approach is adapted from rare event logistic regression with replication (Guns et al., 2012) and is reminiscent of bootstrap Lasso (Bach, 2008). Finally, for each feature, we record the probability of inclusion in the model (the percentage of the N iterations that included the feature into the model at the optimal regularization strength κ^*). We consider those features to be relevant that have an inclusion probability larger than 90% (Bach, 2008). We now fit this sparse model to the full data without the L_1 penalty (a process called "debiasing", Murphy, 2012), since L_1 regularization is biased towards too small weights. We thus obtain our final model, its associated weights $\hat{w}$ and the corresponding transition rate

$$\hat{\lambda}(t, F) = -\hat{w}^T F \cdot \Delta t \ . \tag{4.15}$$

4.2.4 Local cell density as a linear combination of basis functions

To assess the influence of local cell density upon state transitions, we have to assume a specific kernel f, which allows us to evaluate the local density ρ via Eq. (4.1), which is then used as a feature in the GLM. However, this kernel is typically unknown a priori and has to be inferred in parallel with the GLM.

To that end, we model the unknown (radially symmetric) density kernel f as a linear combination of basis functions ϕ_k, $k = 0, 1, \ldots$

$$f \approx \sum_k \omega_k \cdot \phi_k \ , \tag{4.16}$$

where the ϕ_k are defined as

$$\phi_k(r) = \sum_{j \neq i} I\left[k\Delta r < r \leq (k+1)\Delta r\right] \ ,$$

and $I(\ldots)$ is the indicator function. ϕ_k resembles a ring of inner radius $k\Delta r$ and thickness Δr (Fig. 4.1D). For example, we can approximate the tophat kernel with radius R (Eq. 4.2) by choosing the coefficients ω_k as

$$\omega_k = \begin{cases} 1, & k\Delta r < R \\ 0, & k\Delta r \geq R \end{cases} \ .$$

In order to infer the kernel f from the data, the basis functions ϕ_k are evaluated for each sample cell by counting the number of cells within ϕ_k (see Fig. 4.1D), and are then included as individual features into the GLM. By fitting the GLM, the coefficients ω_k are determined and are used to reconstruct the shape of the kernel (see Fig. 4.1E).

4.2.5 Expected frequencies of subtree patterns

Having estimated the transition rate $\hat{\lambda}$ via the regularized GLM, we calculate the number of subtree patterns expected under this transition rate. The expected frequencies of sister cell pairs where in either both cells, one cell, or none of the two cells state transition occurs, can be used to validate the inferred transition rate. We define the random variable C_i to indicate whether cell i underwent a state transition within its lifetime ($C_i = 1$) or stayed in state A ($C_i = 0$). Note that the C_i describe the state of a cell over its entire lifetime, as opposed to the $Y^{(i)}$ used in the previous section, which denote the state of a cell at a small time interval Δt. Using the estimated transition rate $\hat{\lambda}$, we calculate the probability of a state transition in a single cell i as

$$P(C_i = 1) = p_i = 1 - e^{-\int_{\zeta_i}^{\eta_i} \hat{\lambda}(\tau, F_i(\tau))d\tau} \tag{4.17}$$

where $\hat{\lambda}(\tau, F_i(\tau))$ is the estimate of the transition rate the cell experiences throughout its lifetime $[\zeta_i, \eta_i]$ based on its features $F_i(\tau)$. Similarly, we derive the probability of a state transition in its sister cell i' as $P(C_{i'} = 1)$. Considering the whole dataset containing M pairs of sister cells $(i, i'), i = 1 \dots M$, the expected number of pairs where both sister undergo a state transition is:

$$E_2 = \sum_{i=1}^{M} P(C_i = 1, C_{i'} = 1) \ ,$$

where $P(C_i = 1, C_{i'} = 1)$ is the joint probability of these events. However, assuming independence between sisters, this factorizes to

$$E_2 = \sum_{i=1}^{M} P(C_i = 1) \cdot P(C_{i'} = 1) = \sum_{i=1}^{M} p_i \cdot p_{i'} \ . \tag{4.18}$$

The expected number of pairs where a state transition occurs in only one sister (E_1) and in none of the sisters (E_0) are:

$$E_0 = \sum_{i=1}^{M} (1 - p_i) \cdot (1 - p_{i'}) \tag{4.19}$$

$$E_1 = \sum_{i=1}^{M} (1 - p_i) \cdot p_{i'} + p_i \cdot (1 - p_{i'}) \ . \tag{4.20}$$

Applying Eq. (4.17), we can evaluate (E_0, E_1, E_2) in terms of the estimated transition rate $\hat{\lambda}$.

In order to test whether our observed data matches the expected frequencies (E_0, E_1, E_2) we count the observed frequencies (O_0, O_1, O_2) in the data and perform a χ^2 test with two degrees of freedom and

$$\chi^2 = \sum_{j=1}^{3} \frac{(E_j - O_j)^2}{E_j} \ .$$

4.2.6 Summary of the inference methods

Before applying the proposed methods, let us shortly recapitulate the procedure: For each data point i in the observed genealogies (for each cell at each timepoint), features $F^{(i)}$ of interest and the corresponding cellular state $Y^{(i)}$ (state A or transition into B) are extracted.

Now, an estimate $\hat{\lambda}(t, F)$ of the transition rate can be obtained non-parametrically using Eq. (4.8). Alternatively, one can estimate the transition rate parametrically via regularized GLMs, where Eq. (4.14) is optimized with respect to the weights w and the transition rate $\hat{\lambda}(t, F)$ is reconstructed from these weights via Eq. (4.15). To counteract class imbalance, the GLM is trained several times on bootstrapped subsamples of the data.

Finally, after estimation of the transition rate with either of the proposed methods, we validate the estimate using the tree structure of the data. Plugging in our estimate $\hat{\lambda}(t, F)$ into Eq. (4.17) we compare the expected frequencies of subtree patterns (Eqs. 4.18–4.20) to the observed frequencies via a χ^2-test.

4.3 Application to simulated data

In the following, we apply the proposed methods to four different scenarios, show how to recover the features regulating the transition rate from the data and assess the required sample size and tolerable tracking error in cellular genealogies for our analysis.

4.3.1 Estimation of constant and time-dependent transition rates from cellular genealogies

In the simplest scenario the rate λ is constant during the whole time of observation (λ constant, Eq. 4.4). This corresponds to state transitions occurring spontaneously independent of other influences. Using the simulation framework for cellular genealogies (see section 4.1.1), we generate a sample of 100 genealogies with constant rate λ. We then reconstruct the rate $\hat{\lambda}$ from the data via Eq. (4.10) (black curve in Fig. 4.2A) as a function of time. The underlying true rate λ (red curve in Fig. 4.2A) is well contained within the Bayesian 95% credibility intervals of our estimate (gray areas in Fig. 4.2A). Additionally, when reconstructing $\hat{\lambda}$ from the data as a function of local cell density ρ, we again observe a constant transition rate (inset of Fig. 4.2A).

Next, we simulate 100 genealogies with a linear time-dependent transition rate ($\lambda \propto$ time, Eq. 4.5). With the same approach we estimate $\hat{\lambda}$ (see Fig. 4.2B) and again, we observe good agreement between the estimated (black curve in Fig. 4.2B) and the true transition rate (red curve in Fig. 4.2B).

We now account for cell-cell communication and consider a transition rate depending on local cell density ($\lambda \propto$ density, Eq. 4.6): the more cells present in the vicinity of the cell of interest, the more likely it is that a state transition occurs. We estimate the density dependent rate from 100 simulated genealogies, assuming we already know the underlying density kernel (this assumption will be relaxed later on). The estimated rate $\hat{\lambda}(\rho)$ (black curve in Fig. 4.2C) linearly increases with local cell density and the true rate is well contained within the credibility intervals (gray area in Fig. 4.2C), showing that one can identify the influence of local cell density on the transition rate.

However, if we instead estimate the rate as a function of time from the same dataset, we would conclude that the it is time-dependent, since the rate strongly increases over time (see Fig. 4.2C, inset). This is an indirect influence: as time increases, local cell density grows exponentially and as a result, cells are more prone to undergo a state transition (see Fig. 4.2D inset)[3]. We can resolve this by estimating the rate simultaneously as a function of time and local density, $\hat{\lambda}(t, \rho)$ (Fig. 4.2D). For fixed local density ρ, the rate is almost constant across different times (black arrow in Fig. 4.2D). However, the transition rate changes considerably if the local density changes. Therefore, we can conclude that the true transition rate depends only on local cell density. Notice however that this conclusion relies on having sufficiently many samples, yielding a good coverage of the (t, ρ) space, and knowledge of the range (R) and nature of the spatial interaction. If R is chosen too small, any dependence of λ on the local cell density is hidden by the dominating indirect time-dependence. Moreover, analyzing $\hat{\lambda}$ visually becomes infeasible for higher feature dimensions.

4.3.2 Learning the transition rate with generalized linear models

To approach the aforementioned issues, we infer the transition rate more systematically using the machine-learning framework of generalized linear models (GLM, see Methods for details). Instead of considering only one feature at a time, we include all features at once and apply feature selection to determine the relevant ones. An additional advantage of this approach is that it is not necessary to assume any density kernel a priori (as in the previous section). Instead, we use a set of spatial features ϕ_k, whose linear combination can approximate any kernel (Eq. 4.16). We then use the proposed GLM equipped with L_1 regularization to learn the relationship between features and class label and to obtain those features that directly influence the state transition rate.

We apply this approach to the density-dependent dataset ($\lambda \propto$ density, Eq. 4.6). Starting with strong regularization (that is, a large κ and consequently a sparse model) only the most relevant features have non-zero weights and are included (Fig. 4.3A). By decreasing the regularization parameter, the weights of the features gradually increase, making the model more complex. The optimal regularization κ^* (the black line in Fig. 4.3A corresponds to the mean of κ across the 50 bootstraps) is determined by cross validation (see section 4.2.3). All features with non-zero weights at κ^* are included in the model. The ground truth of features used to simulate the dataset is indicated by solid (relevant) and dashed (irrelevant) lines in Fig. 4.3A.

[3] A similar effect is observed in the time-dependent scenario ($\lambda \propto$ time, Eq. 4.5), where the transition rate increases with local cell density due to increasing density over time (Fig. 4.2B inset).

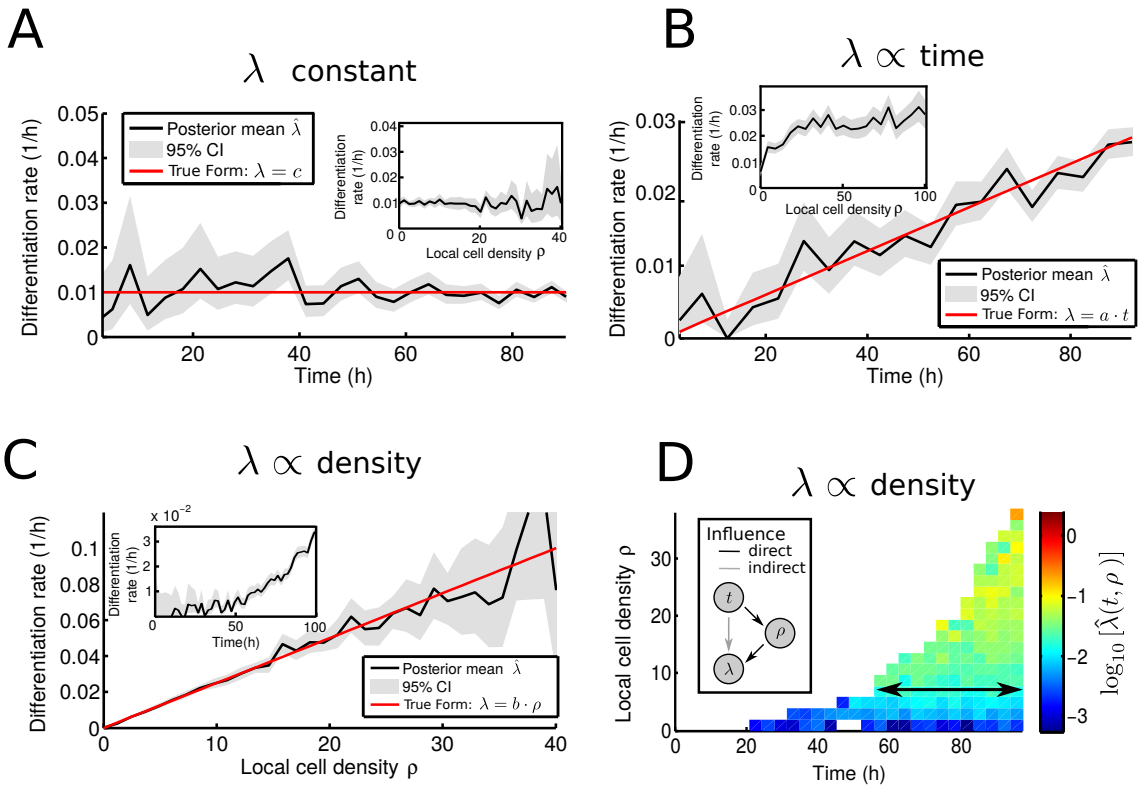


Figure 4.2: **Features regulating the transition rate can be estimated non-parametrically from cellular genealogies with annotated state transition events.** A) The transition rate estimated from 100 genealogies (posterior mean, black line) agrees well with the true constant transition rate (red line). Gray areas indicate the 95% credibility region of the estimate. Inset: Estimated transition rate as a function of local cell density. B) The transition rate estimated from 100 genealogies simulated with linear time-dependent rate agrees well with the true rate (red solid line). Inset: Estimated transition rate as a function of local cell density. C) The transition rate as a function of local cell density ρ for 100 genealogies simulated with density-dependent rate. The estimated transition rate seems to depend on both local density ρ (in line with the simulated form $\lambda = b \cdot \rho$) and time (see inset). D) The estimated transition rate $\hat{\lambda}$ as a function of both density and time reveals that the time-dependence observed in the inset in C) is an indirect influence (density increases with time, see inset). Instead, the transition rate depends only on local cell density ρ (as seen by the mostly uniform pattern of $\hat{\lambda}$ in time for fixed ρ, indicated by arrow).

We estimate the inclusion probability of a feature as the fraction of the 50 bootstraps that selected the feature (Fig. 4.3B). For example, the features $\phi_2, \ldots, \phi_6$ (representing local cell densities at different radii, see Fig. 4.1D) are present in all bootstraps, ϕ_8 is present in 70% of the bootstraps, and all other features have low inclusion probabilities. In particular, time is included in only 18% of the bootstraps and spatial location (x,y) and time since last division (cell cycle) have zero inclusion probability. Choosing a cutoff

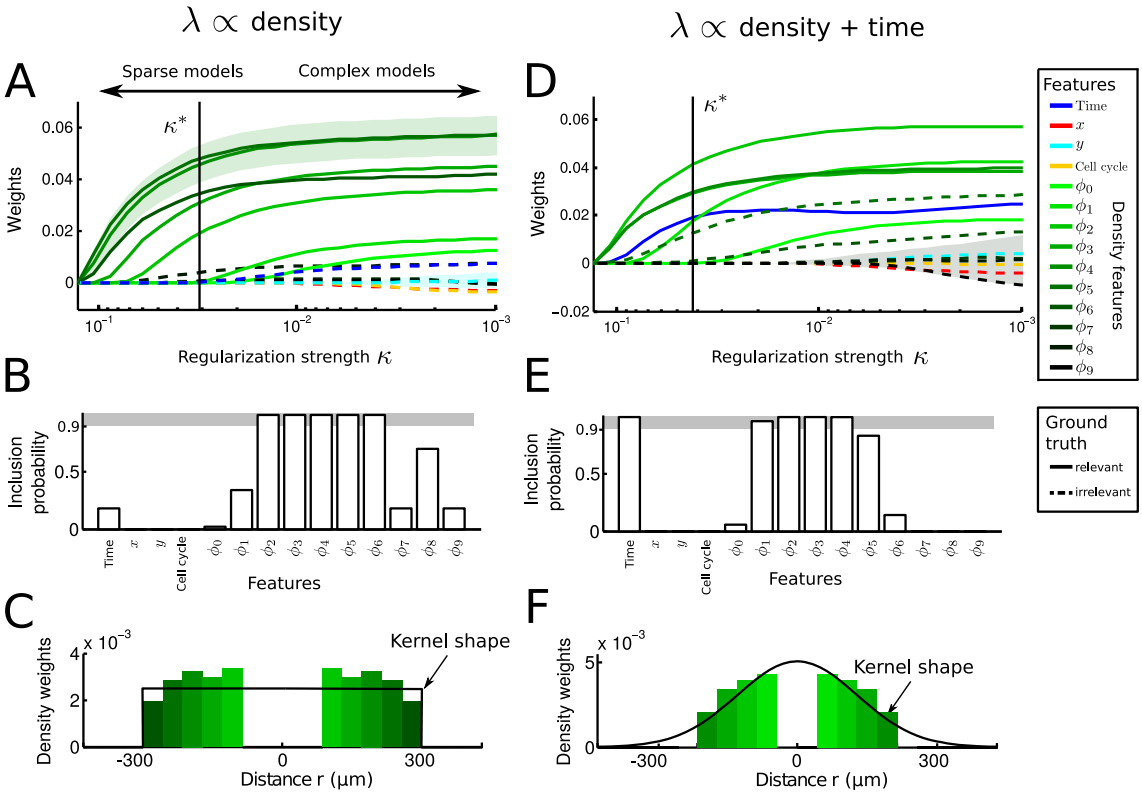


Figure 4.3: **Regularized generalized linear models (GLM) select the relevant features for cell state transitions.** A) Regularization path of the GLMs applied to the density dependent dataset. The means (lines) and standard deviations (shaded regions) of the regression weights w are plotted against the regularization strength κ across 50 bootstrap samples (see section 4.2.3). The mean of the optimal regularization strength κ^* determined by cross validation is shown as a vertical black line. Solid (dashed) lines correspond to relevant (irrelevant) features in the respective scenario. B) Percentage of bootstrap samples that included the respective features. Included features were determined as those with non zero weights at κ^*. Enforcing a 90% threshold (gray area) on the inclusion probability for each feature, we select the relevant features of the model. C) Reconstructed kernel of local cell density (bars) from the selected features in B. The true underlying tophat kernel shape is shown in black. D-F) Analogous to A-C, but for a dataset where the transition rate depends on time and local cell density with a Gaussian kernel. Both features are correctly identified and the density kernel is correctly estimated.

at 90% (gray area in Fig. 4.3B) for a feature to be included in the final model, we recover all features (except ϕ_0, ϕ_1) that were used to generate that dataset. We miss ϕ_0 and ϕ_1 since their contribution to the overall transition rate is effectively very low: the average number of cells within ϕ_1 is approximately 0.2, whereas the average number of cells within e.g. ϕ_7 is approximately 1. Hence, leaving out ϕ_1 will not change the overall result, and the algorithm chooses to neglect the feature in favor of sparsity.

After feature selection, we can reconstruct the density kernel as a weighted sum of the

basis functions ϕ_k via Eq. (4.16) (shown as green bars in Fig. 4.3C). Here, we observe that the reconstructed kernel closely resembles the true underlying tophat kernel that was used to simulate the data (shown as a black curve in Fig. 4.3C).

We extend the set of relevant features and now consider a scenario where the transition rate depends on time and on local cell density ($\lambda \propto$ density + time, Eq. 4.7), modeled via a Gaussian kernel (with $\sigma = 130\ \mu m$) instead of a tophat kernel. The regularization path and the feature inclusion probabilities (Fig. 4.3D,E) show that the GLM correctly selects both time and local cell density ($\phi_1, \ldots, \phi_4$) with inclusion probabilities close to 1. Finally, using the weights associated with the selected density features we reconstruct the kernel of local cell density and find that it indeed matches a Gaussian kernel (Fig. 4.3F). As before (Fig. 4.3A-C), the feature selection procedure misses ϕ_0 due to its relatively small contribution to the overall transition rate. We conclude that our proposed method is capable of identifying the features that directly influence the transition rate and faithfully filters out indirect influences. Furthermore, we can estimate the shape of the density kernel from the data.

4.3.3 Sample size estimation

Accurate single-cell identification and tracking in time-lapse movies is still a challenging task and requires, at least in mammalian systems manual data curation (Schroeder, 2008; Amat et al., 2014). Thus estimating the required sample size for any given effect size is necessary for efficient experimental design.

To assess the impact of sample size on the performance of the feature selection, we systematically reduce the number of observed state transition events (by reducing the number of genealogies) and calculate the inclusion probabilities as a function of sample size (Fig. 4.4A,B, averaged across 10 replicates of the respective sample size). Starting at the original sample size of 3000 onsets (using all 100 genealogies), we find the same features above the threshold as before (Fig. 4.3B,E). Decreasing the sample size, the inclusion probability of certain features gradually drops below 0.9 (e.g. ϕ_2 in Fig. 4.4A): The data no more contains sufficient statistical information to identify the feature as relevant. At a sample size below 500 events, all features are considered irrelevant. However, a sample size of 1500 (corresponding to 35 genealogies) is sufficient to faithfully detect the underlying features influencing the transition rate and to distinguish a direct time-dependence (Fig. 4.4A) from an indirect one (Fig. 4.4B).

4.3.4 Influence of tracking error

To obtain genealogies from time-lapse microscopy data, manual (Schwarzfischer et al., submitted) or automatic tracking (for an overview of current methods, see Maska et al., 2014) is required. Neither automatic nor manual tracking can produce perfect genealogies, but will introduce errors especially when local cell density is high or cells move fast as compared to the time resolution of the imaging. To test the influence of tracking errors on the our method, we introduce artificial tracking errors into the simulated datasets by interchanging the identity of randomly selected cells of the same generation and hence swapping entire subtrees of the genealogies. The amount of tracking error is defined as the percentage of all cells in the dataset where an artificial tracking error was introduced. We

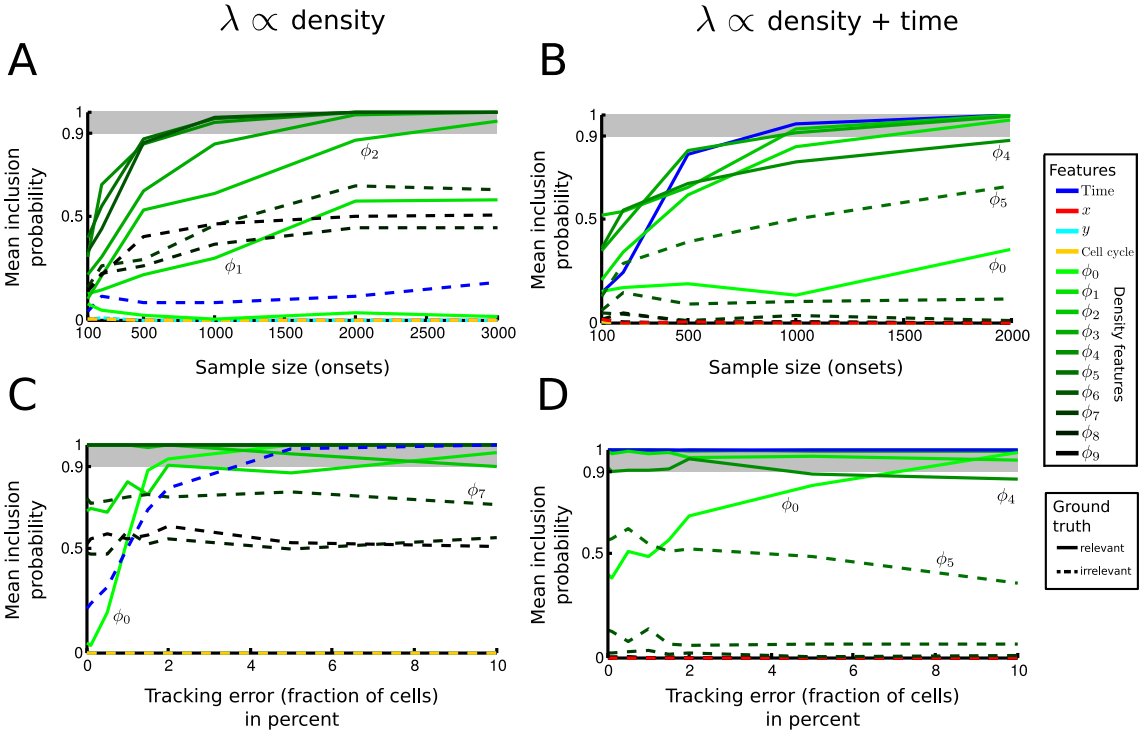


Figure 4.4: **The method's performance is robust for different sample sizes and moderate amount of tracking error.** A) Mean inclusion probability for each feature plotted against sample size. Relevant features are correctly selected (inclusion probability > 0.9, gray area) when 1500 or more transition events (corresponding to approximately 35 genealogies) are used for the analysis. Solid (dashed) lines correspond to relevant (irrelevant) features in the respective scenario. B) Analogous to A, but for a dataset used in Fig. 4.3D-F, where the transition rate depends on time and local cell density with a Gaussian kernel. C) Mean inclusion probability plotted against the amount of tracking error for the density dependent scenario from Fig. 4.3A-C (100 genealogies). The correct features are identified (inclusion probability > 0.9, gray area) up to a tracking error of 5%. For larger tracking error, time (blue curve) is incorrectly identified as a direct instead of an indirect influence. D) Analogous to C, but for the dataset where the transition rate depends on time and local cell density with a Gaussian kernel (100 genealogies).

simulate different amounts of tracking error with up to 10% of all cells in the experiment containing a tracking error. We now evaluate the previous results on these erroneous datasets.

We find that for both the density dependent ($\lambda \propto$ density, Fig. 4.4C) and the time and density dependent scenarios ($\lambda \propto$ density + time, Fig. 4.4D) our method reliable identifies the underlying features (using the inclusion probability threshold of 0.9 as before) for up to 5% of tracking error. For higher amounts of tracking error, we erroneously identify time as a relevant feature in the first scenario (blue line in Fig. 4.4C) and we fail to identify ϕ_4 (Fig. 4.4D) as relevant feature in the second scenario. Note that tracking errors impact this analysis only by the creation of spurious state transitions (a cell in state A is at some

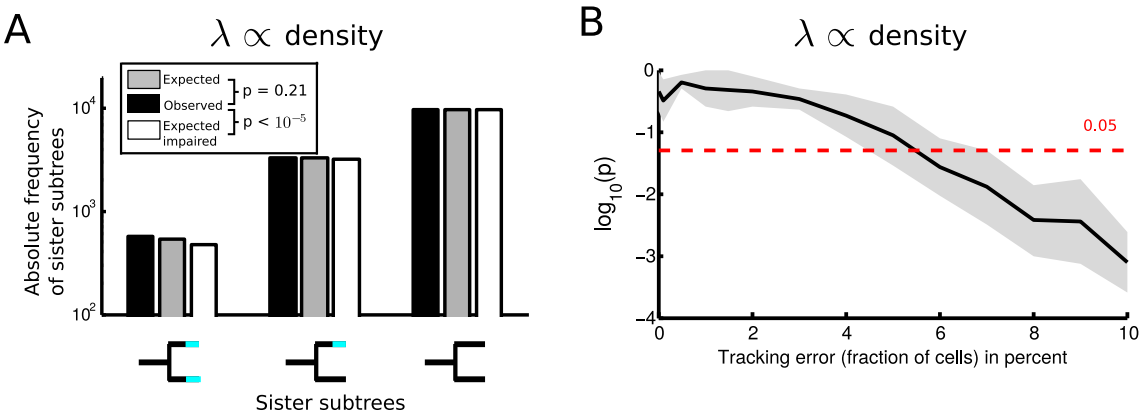


Figure 4.5: **Expected frequencies of sister pairs reveal if the model can account for the observed genealogical correlations.** A) Comparison of the observed and expected frequencies of sister pairs (both, one, or none undergoing a state transition) of the dataset used throughout Fig. 4.3A-C shows no significant difference ($p = 0.21$, χ^2-test, see section 4.2.5). Fitting the same data, but not accounting for the ϕ_5, ϕ_6 features causes significant deviations from the expected frequencies ($p = 1.3 \cdot 10^{-6}$). B) P-values of the χ^2-test (average and standard deviation over 10 replicates) to compare the observed and expected frequencies of sister pairs against amount of tracking error for the density dependent scenario. For tracking errors $< 5\%$, the method correctly concludes that the frequencies of observed sister pairs are in agreement with the model (applying a significance threshold of $\alpha = 0.05$, red dashed line).

point accidentally interchanged with a cell in state B) and the method does not rely on extensive, long trackings.

4.3.5 Model validation using sister correlations

Apparently, our method is able to infer state transition mechanisms by identifying relevant features even in the presence of moderate tracking errors. However, what if we miss to include relevant features in the GLM, e.g. unobserved influences like nutrient concentrations? In this section, we use the tree structure to validate the chosen model by investigating whether the transition rate λ estimated by the GLM is capable of explaining the observed correlated transition events within the cellular genealogies. We focus here on correlations between sister cells, but the approach easily generalizes to higher order relationships within a genealogy, like cousin-quartets. Suppose that we obtained a reasonable estimate $\hat{\lambda}$ of the transition rate. Then, the state transition of one sister cell is independent of the other and just determined by the transition rate that might differ due to the spatial context in both cells. With this independence assumption, we can calculate the probability to observe sister subtree patterns (where both, one or none of the sister cells change state) just as the product of the individual probabilities (see section 4.2.5). Note that these probabilities are calculated over the entire lifetime of each cell finally resulting in the expected number of sister subtree patterns for the entire dataset.

Using these frequencies, we assess if the transition rate learned by the GLM (agnostic of the tree structure) is capable of explaining the observed correlations in the genealogies and therefore is an adequate description of the data. For the dataset where the state transition depends only on local cell density ($\lambda \propto$ density, Eq. 4.6), we calculate the expected frequencies of sister subtrees given the previously estimated transition rate (Fig. 4.5A, gray bars) and compare these to the observed frequencies in the data (Fig. 4.5A, black bars). No significant differences are observed ($p = 0.21$, χ^2-test, see section 4.2.5), and hence, there is no indication of correlations beyond what we expect from the density dependent transition rate, in agreement with the generative framework.

Next, we show how this idea can be used to determine if all relevant features have been included in the GLM. To that end, we now deliberately neglect the spatial features ϕ_5, ϕ_6 when fitting the transition rate via the GLM. Since these two features influence the transition rate in the chosen scenario, fitting the impaired GLM yields a different $\hat{\lambda}$ and hence also different expected frequencies of sister subtrees (Fig. 4.5A white bars). The frequencies are significantly different ($p = 1.3 \cdot 10^{-6}$), indicating the model is inappropriate, as there is more correlation in the trees than the model can account for (due to the missing ϕ_5, ϕ_6).

Our approach to validate the model using sister correlations (Fig. 4.5A) relies on entire correct trackings of both sister cells, as we integrate over the entire lifetime of these cells in Eq. (4.17). Analogous to Fig. 4.5A, we evaluate whether we observed frequencies of sister subtrees match the expectations of the model (which was also fitted to the dataset containing the tracking errors) via a χ^2-test for different amounts of tracking error. For the density dependent scenario, we find that up to 5% of tracking error, we do not observe significant differences between observed and expected frequencies ($\alpha = 0.05$), correctly indicating that the density dependent transition rate can explain the observed frequencies (Fig. 4.5B). However, more than 5% of tracking error result in substantial changes of the sister correlations, which cannot be explained by the model of the transition rate (shown by the significant differences in frequencies).

4.4 Discussion

In this chapter, we have presented a method to infer the mechanisms driving cell state transition events from observed cellular genealogies. As two features explicitly regulating the transition rate, we have here considered time and local cell density. Our method extends the approach by Snijder et al. (2009) who showed that the response of a cell to a certain stimulus (in their case, a virus infection) strongly depends on each cell's "population context", that is, its localization within the colony, its cell density and cell cycle stage. This approach, which has been applied to the analysis of high-content screens by Knapp et al. (2011), is designed for static data and a single, controlled perturbation. The cells are subject to a treatment at a defined timepoint and their response is recorded by a single image. For our purpose a static approach, where the timepoint of the event is predetermined, is not applicable. Instead, we assume that cells undergo state transitions spontaneously, and hence transition events can happen at any point in time but their probability chances over time due to the changing environment the cells experience.

Methodologically, our method is extendable to detect more general, non-linear dependencies in the data. To that end, we can either explicitly perform a basis function expansion of the features F, or combine the regularized GLM with kernel methods, e.g. relevance vector machines (Tipping, 2001). Note that our approach shares certain aspects with proportional hazard models (Cox, 1972). However, these models cannot handle tree-structured data and thus are not applicable to cellular genealogies.

With respect to regulating features, our method can be extended to any other parameter that is experimentally accessible. In terms of tumor growth for example, the presence (or local density) of distinct cancer cell subtypes might influence transitions between states of different proliferative potential (Stingl and Caldas, 2007). For blood progenitor cells, including the expression levels of Pu.1 (Kueh et al., 2013), a pivotal fate determining factor (Krumsiek et al., 2011), as a feature will allow to compare extrinsic and intrinsic (Strasser et al., 2012) effects on cellular plasticity. For pluripotent murine embryonic stem cells, we would like to discuss more specifically, how our method can be applied. Embryonic stem cells transit from a Nanog-high state, where cells are safeguarded from differentiation, to a Nanog-low state, where cells are more prone to differentiate (Chambers et al., 2007). Several models have been proposed to explain the transition from Nanog-high to Nanog-low states. For some models (Chickarmane and Peterson, 2008; Kalmar et al., 2009; Chickarmane et al., 2012), the transition emerges from the entirely intrinsic dynamics of a small transcription factor network and external signals are neglected. Other models augment the intrinsic dynamics of a transcription factor network by external differentiation signals (Glauche et al., 2010; Chickarmane et al., 2006) and autocrine signaling (Herberg et al., 2014), both of which could depend on other factors, such as local cell density. To study Nanog-high to Nanog-low state transitions and its potential dependence on e.g. local cell density with our method, one would first set up our simulation framework to generate genealogies similar to embryonic stem cell genealogies observed in experiments in terms of cell lifetime, movement, seeding density, etc. Next one has to implement the hypothesized mechanism either in a simplified way (a simple dependence of the transition on local density) or in full detail (e.g. taking the proposed model by Herberg et al., 2014, and modulating the strength of the autocrine feedback with local density). Using our method, one can then estimate the number of samples and the allowable tracking error required to discover the hypothesized mechanism in the data. Finally, one can design experiments and post-processing according to these requirements, and decide e.g. if automatic tracking algorithms (yielding many but potentially wrong genealogies) or careful manual tracking (to obtain accurate but fewer genealogies) should be used, what platform and experimental techniques to use (e.g. microfluidics Blagovic et al., 2011), etc.

Summarizing, our approach is designed for dynamic data provided by time-lapse microscopy, which allows to observe state transitions in their spatiotemporal and genealogical context. The requirements for an appropriate dataset are (i) single-cell genealogies obtained from automatic or manual cell tracking, (ii) at least as many annotated state transitions as determined by our analysis, and (iii) the identification of all cells surrounding a transition event in an sufficiently large radius. To the best of our knowledge, no such dataset exist up to now, but manual and automated tracking tools increase accuracy and efficiency (Chenouard et al., 2014; Amat et al., 2014; Schwarzfischer et al., submitted). Moreover, our method relies only on short trackings of one cell cycle to quantify sister

correlations (Fig. 4.5). Since fluorescent fate markers exist for various systems, morphological quantification has been shown to be usable for fate recognition (Cohen et al., 2010), and robust cell segmentation algorithms work on full time-lapse movies (Buggenthin et al., 2013), we believe that adequate datasets from various cell systems will emerge in the near future. Due to the method's generality, many different types of cell state transitions can be investigated in their spatiotemporal context, from apoptosis over stem cell differentiation to epithelial-mesenchymal transitions and tumorogenesis.

Chapter 5

Inferring lineage decisions from genealogies

Up to now, we have discussed separately mechanistic cell-intrinsic models (chapter 3) as well as coarse-grained cell-extrinsic models of cell fate decision (chapter 4).

In this chapter, we now combine these two approaches into a single model, which ameliorates previous limitations, such as the assumption of instantaneous marker onset after differentiation (chapters 3 and 4) and neglect of the tree structure (chapter 3). Finally, we apply this method to genealogies of differentiating blood stem cells and assess whether the long standing paradigm of the PU.1/Gata1 toggle switch in myeloid/erythroid lineage decisions hold true.

Time-lapse microscopy can not only be used to observe cell state transitions such as differentiation single cells, but can put these transitions into their genealogical context via cell tracking. This lead to the striking observations that cells which are descendants of the same ancestor cell tend to behave similar: For example, in yeast, sister cells switch gene expression of a simple regulatory circuit in a correlated fashion (Kaufmann et al., 2007). In hematopoiesis, it has been observed that differentiation events are highly correlated within trees (Hoppe et al., in revision; Rieger et al., 2009), that is, very often sister cells show similar timing in fate choice. (see Fig. 5.1).

In chapter 4 we showed how external features, such as local cell density can lead to correlated differentiation events (see Fig. 4.5A) and proposed a statistical test which determines if those external features are sufficient to explain the observed correlated events. However, given the rapid movement of blood cells and the extent of correlations across several generations, common external features as origin of these correlations is unlikely. Instead, we propose a different mechanism: Differentiation events are typically read out via the expression of some marker gene. For example, expression of the CD16/32 membrane receptor (detected by in culture antibody staining) marks differentiation of a early blood progenitor cell into a granulocyte-monocyte progenitor Hoppe et al., in revision, and expression of LysM::GFP (fusion protein) marks differentiation of a granulocyte-monocyte progenitor into either granulocytes or monocytes (Rieger et al., 2009). However, these markers report the event of differentiation only indirectly, because they are a downstream consequence of differentiation (e.g. LysM is upregulated because this bacteriolytic enzyme is essential for the immune action of mature monocytes and granulocytes). Hence, there

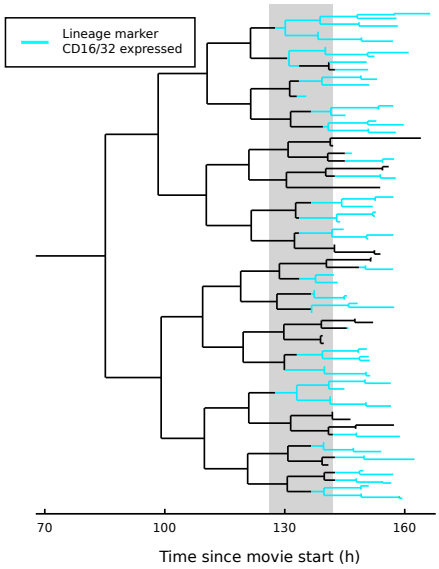


Figure 5.1: **Observed marker correlation in genealogies.** A genealogy originating from a single hematopoietic stem cell (only part of the genealogy is shown) is observed to differentiate into the granulocyte-monocyte lineage as indicated by the onset of CD16/32 expression (cyan). Most marker onsets are observed in a narrow time-window around 130 hours after movie start (gray box). Data taken from Hoppe et al., in revision.

is a cascade from the differentiation decision towards the observed marker expression, resulting in a delay between actual differentiation and the timepoint where it can be read out experimentally. Here, we propose that the correlations in trees arise due to a delay between a cellular decision and its observation. We develop at a computational method that estimates and removes this delay to obtain the true timepoint of decision. We first study the performance of the method on a simple model of linearly time dependent differentiation hazard and a delay caused by gene expression and then apply the method to a mechanistic toggle switch model of differentiation.

The blood stem cell genealogies used in section 5.3.4 are contained in the manuscript by Hoppe et al., in revision.

5.1 A differentiation model on genealogies

Time lapse microscopy combined with cell tracking and fluorescence signal quantification (Schwarzfischer et al., submitted) provides lineage trees of single cells as well as their corresponding fate. In Fig. 5.2A, we show an example of a lineage tree schematically. Each tree starts with a single cell at $t = 0$ (the start of time lapse microscopy). These cells are assumed to be all equivalent, synchronized and undifferentiated, because they underwent stringent flow cytometry purification before entering the time-lapse microscopy pipeline. During the course of the experiment, the cell will eventually divide and give rise to two daughter cells, which is indicated by the branching events in Fig. 5.2A. The length of the

segments corresponds to the cell's lifetime. These cells will eventually also divide, giving rise to further progeny, but at some point onset of the differentiation marker is observed. This is annotated in the tree in Fig. 5.2A by gray circles. Note that cells are tracked and keep dividing after the point of marker onset, but as the tree beyond this point is irrelevant for our purpose, we terminate it at the points of marker onset.

5.1.1 Genealogies as tree structures

The representation of genealogies shown in Fig. 5.2A is useful to visualize the genealogies, but now we shall introduce a formal representation of the genealogies as rooted, unordered, binary and labeled trees. A tree is a acyclic, connected graph, $T = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a set of nodes, and $\boldsymbol{E}$ is a set of directed edges between nodes. For each cell in the genealogy we introduce a single node v. Two nodes $v, w \in \boldsymbol{V}$ are connected by a directed edge $e = (v, w)$ from v to w if v is the parent cell of w. The node w is called the child of v. Nodes $v, w \in \boldsymbol{V}$ are sisters/siblings if they share the same parent node. The node $r \in \boldsymbol{V}$ is called the root of tree T if it has no parent node. The set of nodes $\boldsymbol{L} \subset \boldsymbol{V}$ without children are called leaves of T. A node v is called an ancestor of node w if there exists a path $(e_1, \ldots, e_n)$ that starts at v and stops at w, which is then called a descendant of v.. The set of ancestors and descendants of a node v are denoted by $ac(v)$ and $de(v)$, respectively. A tree is termed binary if every node has at most two children and unordered if no particular ordering is assigned to the two children nodes.

A tree $T = (\boldsymbol{V}, \boldsymbol{E}, \sigma)$ is said to be labeled, if we define a function $\sigma : V \to \Sigma$ which assigns every node $v \in \boldsymbol{V}$ to an element of an alphabet Σ. For cellular genealogies, each node $v \in \boldsymbol{V}$ is labeled by its marker expression $m_v \in \mathbb{B}$, where $\mathbb{B} = \{0, 1\}$, i.e. whether the cell does ($m_v = 1$) or does not ($m_v = 0$) expresses the differentiation marker. Furthermore each node v is labeled by the time of division $\varsigma_v \in \mathbb{R}^+$ which gave rise to cell v (its "birth time") and its observation period $\tau_v \in \mathbb{R}^+$. Observation of the current cell is terminated if (i) the cell divides, (ii) it expresses the marker, (iii) the cell is lost in tracking, dies, or the end of the experiment is reached. As we stop observation at marker expression, we have $m_i = 0$ for all non-leave nodes $v_i \in \boldsymbol{V} \setminus \boldsymbol{L}$. For convenience to enumerate the nodes, we label the nodes $v \in \boldsymbol{V}$ by positive integers $c_v \in \mathbb{N}^+$: The root node r of the tree has label $c_r = 1$. Children v, w of node u with label $c_u = i$ are labeled $c_v = 2 \cdot i, c_w = 2 \cdot i + 1$, where ordering of v, w is irrelevant. Overall, we obtain the labeling

$$\begin{aligned} \sigma : \boldsymbol{V} &\to \mathbb{B} \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{N}^+ \\ v &\mapsto \left(m_v, \quad \varsigma_v, \quad \tau_v, \quad c_v\right) . \end{aligned}$$

5.1.2 Latent state transitions: hidden trees

Often, one will observe that, for example, sister cells (but also more distantly related cells) behave similarly in terms of marker expression: Either both sister cells lack the marker and are therefore thought to be undifferentiated or in both sisters a marker onset is detected, often even at similar times (see section 4.3.5 for a statistical test of sister correlations and section 5.3.1 for an application thereof). We propose that the observed correlations emerge because of a delay between the differentiation event and the observation of marker onset.

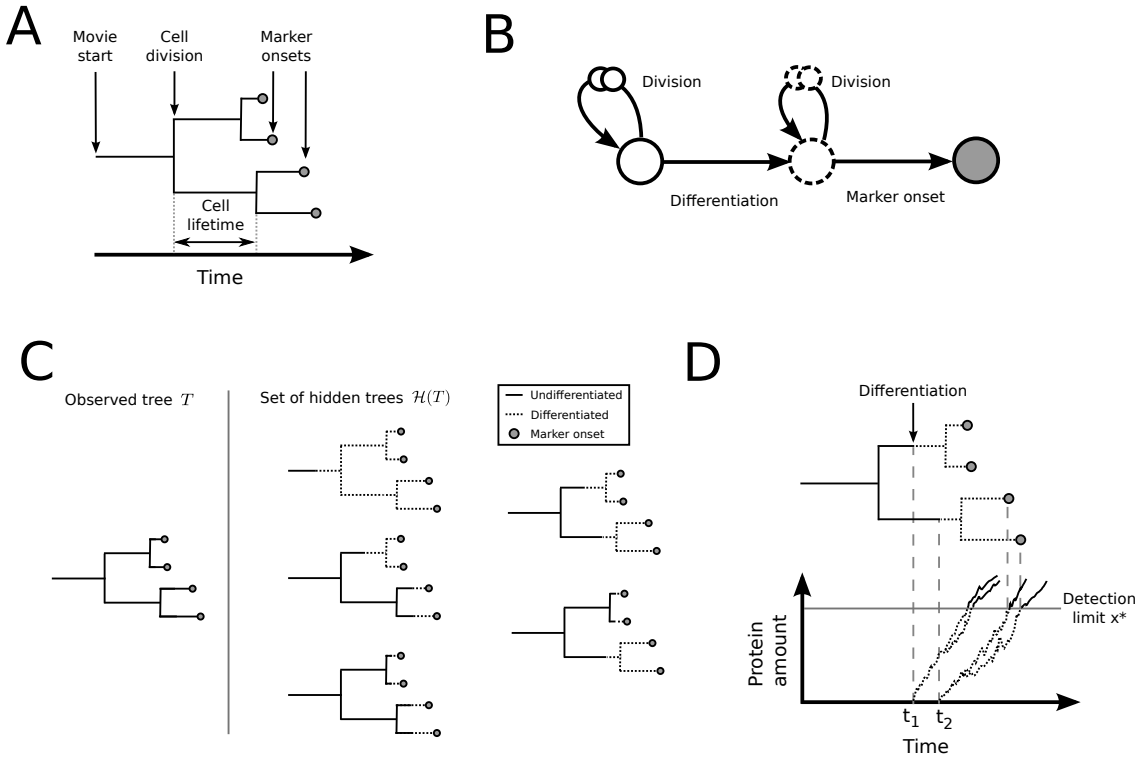


Figure 5.2: **Marker delay causes correlated onsets in lineage trees.** A) Tracking a single cell from movie start gives rise to a lineage tree via multiple cell divisions. Once the expression of a marker is detected in a specific cell (indicated by gray circles). Frequently, one observes synchronous marker onsets between sister cells in real data (Rieger et al., 2009; Kaufmann et al., 2007),Hoppe et al., in revision. B) In a simple model of differentiation and delayed marker onset, an undifferentiated cell (white circle) can either divide into two undifferentiated cells or differentiate. A differentiated cell (dashed white circle) can divide into two differentiated cells or become positive for the marker (gray circle). C) The marker onsets observed in tree T of A) can originate from different possible scenarios of differentiation, termed hidden trees $\mathcal{H}(T)$, as the underlying dynamics are unknown. D) One possible realization of the proposed model with outcome A). Cell 2 and 3 independently differentiate at time t_1 and t_2 and start expressing the marker, but before reaching the detection threshold (gray line) both cells divide. The offspring inherits the state of the mother cell and hence sister cells will reach the threshold at similar, but due to stochasticity in gene expression, not identical times.

We define the differentiation event as the irreversible commitment of the cell into its future fate. For example, in the myeloid/erythroid cell fate decision of the CMP into either GMP or MEP (see Fig. 1.6B), we consider the CMP as undifferentiated, because it is bipotent, i.e. it can give rise to both GM- and MegE-lineages. As soon as the cell differentiates and commits to one or the other fate, it looses bipotency, which is characteristic for undifferentiated cells. For simplicity, in the remainder of the chapter we will consider only one of these two possible cell fate choices. In context of hematopoiesis, we focus only on CMPs committing towards the GM-lineage. An extension of the method towards to distinct cell fates is discussed in section 5.4.

In our simply model (Fig. 5.2B), an undifferentiated cell (shown as solid white circle) can either divide (giving rise to two undifferentiated daughter cells) or progress into the differentiated state (shown as dashed circle). However, this cell does not immediately show marker onset, but can still divide several times before it progresses into the final state where marker onset occurs (gray circle in Fig. 5.2B). According to this generic model, one observed tree $T = (\boldsymbol{V}, \boldsymbol{E}, \sigma)$ can be explained by several scenarios of differentiation (Fig. 5.2C), that we will call "hidden trees".

Formally, we define a hidden tree as a rooted, unordered, binary and labeled tree, $H = (\boldsymbol{V}, \boldsymbol{E}, \sigma')$, which in comparison to the observed tree T has additional labels associated to each node, such that

$$\begin{aligned} \sigma' : \boldsymbol{V} &\to \mathbb{B} \times \mathbb{R}^+ \times \mathbb{R}^+ \times \mathbb{N}^+ \times \mathbb{B} \times \mathbb{B} \\ v &\mapsto \left(m_v, \quad \varsigma_v, \quad \tau_v, \quad c_v, \quad u_v, \quad d_v\right) , \end{aligned}$$

where u_v indicates if v is undifferentiated and d_v indicates v is differentiated. Note that labels u_v, d_v, m_v are not mutually exclusive: For a cell v that differentiated $u_v = d_v = 1$, as it is undifferentiated in the beginning but then changes into the differentiated state. Similarly, for a cell v that starts expressing the marker, $d_v = m_v = 1$. Due to the directionality of the process depicted in Fig. 5.2B, certain constrains are imposed on the labeling:

$$\begin{aligned} \forall_{v:m_v=1} \forall_{w \in ac(v)} m_w &= 0 \\ \forall_{v:u_v=1} \forall_{w \in ac(v)} u_w &= 1 \\ \forall_{v:d_v=1} \forall_{w \in de(v)} d_w &= 1 \\ \forall_{v:m_v=1} d_w &= 1 \ . \end{aligned}$$

That is, ancestors of marker positive cells are marker negative, ancestors of undifferentiated cells are also undifferentiated, descendants of differentiated cells remain differentiated, and marker positive cells must be differentiated.

To relate observed and hidden trees, we note that an observed tree $T = (V, E, \sigma)$ is obtained from a hidden tree $H = (\boldsymbol{V}, \boldsymbol{E}, \sigma')$ by applying a projection to the labeling σ', that is, dropping the labels d_v, u_v, but conserving the nodes and edges of H. On the other hand, an observed tree T corresponds to a set of hidden trees denoted by $\mathcal{H}(T)$ (Fig. 5.2C).

Our goal is to judge which of these alternatives $H \in \mathcal{H}(T)$ is the most likely one given experimental data. Therefore, we have to make assumptions about the differentiation and delay process. Most importantly, we assume that the differentiation decisions are

independent between cells: No internal (unobserved) information is passed from mother to daughter cell that has influence on the timing of differentiation. This independence requirement is a natural definition of a decision event: For example, if one detects any correlation in differentiation in sister cells, one can argue that the actual decision to differentiate was initiated previously in the mother cell. Therefore, the probability to differentiate must only depend on external factors, that are not inherited during cell division. In the following, we will simply assume that the probability to differentiate is a function of time.

For the marker delay process, we assume that immediately after the decision to differentiate, the cell starts expressing the marker protein (Fig. 5.2D). Due to a detection limit of the experimental technique, the expression of the marker cannot be detected instantaneously, but only when the amount of protein exceeds this limit x^* (which is typically at a few hundred molecules, Schwarzfischer et al., submitted). In Fig. 5.2D, cells 2 and 3 independently differentiate at times t_2 and t_3 and the expression of the marker protein starts. Before the protein amount in any of the two cells reaches the detection limit, both cells divide and they daughter cells inherit the state of protein expression from their mother. Note that without loss of generality, we neglect partitioning of molecules at cell division, therefore daughters start exactly at the same state. As daughter cells inherit the state of their mother, they will look correlated with respect to marker onset: If one daughter reaches the detection limit, the other daughter will likely do the same (their initial distance to the threshold is very similar), but because gene expression is intrinsically stochastic, the behavior of both cells will not be exactly identical. However, this statistical correlation is only mediated via the state of the mother: If we condition on the state of the mother cell, the behavior of both daughter will be independent.

5.1.3 Differentiation process

Let us define the proposed model of differentiation events more formally. As in chapter 4, we define a point process with rate $\lambda(t)$ so that $\lambda(t)dt$ is the probability that an event occurs in the interval $[t, t + dt]$, given that the event has not occurred in the interval $[0, t)$. Furthermore, we can define the overall distribution of event times $\Phi(t)$, that is, the probability to observe an event at time t (analogous to $P(g)$ in chapter 3 but for continuous time). Both concepts are related via:

$$\Phi(t) = \lambda(t)e^{-\int_0^t d\tau\lambda(\tau)}$$

For example, if $\lambda(t) = \lambda$ is constant, the above equation yields $\Phi(t) = \lambda e^{-\lambda t}$ which is the probability density of an exponential distribution. Without loss of generality, but motivated by experimental observation (Marr et al., 2012), from now on, we will assume that the rate of differentiation is a linear function of time so that

$$\lambda(t) = a_0 + a_1 \cdot t \tag{5.1}$$

$$\Phi(t) = (a_0 + a_1 \cdot t)e^{-\int_0^t d\tau(a_0+a_1\cdot\tau)} \tag{5.2}$$

$$= (a_0 + a_1 \cdot t)e^{-(a_0\cdot t+\frac{a_1}{2}t^2)} \; . \tag{5.3}$$

This represents a first order approximation to a potentially complex but unknown rate of differentiation. However it allows more flexibility than a zeroth-order approximation

where $a_1 = 0$ and is sufficient to encompass mechanistic models of cell fate choice (see section 5.3.3). From now on, we denote the parameters of the differentiation process as $\theta = (a_0, a_1)$ and write $\Phi(t|\theta)$ to make the dependence on the parameters explicit.

Note that this differentiation model can be extended to depend on other observable features than time, e.g. local cell density (see chapter 4).

5.1.4 Delay process

We model the marker delay as a stochastic gene expression process. Lumping together transcription and translation for simplicity, we obtain a birth-death process with two reactions, one producing a protein with rate α and the other removing a protein with rate γ:

$$\emptyset \xrightarrow{\alpha} X$$
$$X \xrightarrow{\gamma} \emptyset \, .$$

The Chemical Master Equation (see chapter 2) describing how the distribution of protein numbers x evolves over time is

$$\frac{\partial \mathcal{P}(x,t)}{\partial t} = -(\alpha + \gamma x)\mathcal{P}(x,t) + \alpha \mathcal{P}(x-1,t) + \gamma(x+1)\mathcal{P}(x+1,t). \tag{5.4}$$

We are only interested in the dynamics of the system until the protein numbers exceed the detection threshold x^*, where we assume that the marker can be observed, Therefore, we apply the finite state projection (Munsky and Khammash, 2006) to the master equation, truncating the statespace at state $x^* - 1$ and introducing the absorbing state x^*. This is readily achieved by extending Eq. (5.4) with separate equations for the states x^* and $x^* - 1$:

$$\frac{\partial \mathcal{P}(x^*-1,t)}{\partial t} = \alpha \cdot \mathcal{P}(x^*-2,t) - [\alpha + \gamma(x^*-1)] \cdot \mathcal{P}(x^*,t)$$
$$\frac{\partial \mathcal{P}(x^*,t)}{\partial t} = \alpha \cdot \mathcal{P}(x^*-1,t) \, .$$

Hence, all probability that leaves the truncated statespace will be collected in the absorbing state x^*. We are interested in the first passage time distribution $\Psi_{x_0}(t)$, that is, the probability that the protein number crosses the threshold x^* for the first time at time t starting in state x_0:

$$\Psi_{x_0}(t) = P(x^*, t, x(s) < x^* | x_0) \quad \forall s < t \, .$$

In the truncated statespace, this is precisely the probability flow from state $x^* - 1$ to x^* at time t (Van Kampen, 1992):

$$\Psi_{x_0}(t) = \frac{\partial \mathcal{P}(x^*,t)}{\partial t} = \alpha \cdot \mathcal{P}(x^*-1, t|x_0). \tag{5.5}$$

$\Psi_{x_0}(t)$ depends of course on the parameters $\eta = (\alpha, \gamma, x^*)$ of the underlying model, but we have dropped this dependence for readability. To obtain $\Psi_{x_0}(t)$ we have to solve the Chemical Master Equation (Eq. 5.4) up to time t with initial condition $\mathcal{P}(x,0) = \delta_{x,x_0}$

to calculate $\mathcal{P}(x^* - 1, t|x_0)$, which can be done using standard ODE solvers. However, the system of differential equations grows linearly with the size of the statespace, which can quickly become infeasible. For this simple stochastic model, an expression for first passage time distribution can be derived in terms of a renewal equation (see supplement of Shahrezaei and Swain, 2008). However, when dealing with tree structures in section 5.2, we will see that we have to solve the master equation numerically to also obtain the propagator $P_{x\to x'}(t)$, the probability to start a state x and after time t arrive at state x'.

Note that, as in section 5.1.3, the model described above is only an approximation to the underlying complex delay process. For example, the underlying process might involve the upregulation of the marker protein via a cascade of several genes, which is triggered upon differentiation. However, we will see that this model is sufficient to describe the dynamics of such complex delay processes in section 5.3.2.

5.2 Statistical inference

Our goal is to estimate the parameters of the model shown in Fig. 5.2 from observed lineage trees in order to predict the differentiation events in a given tree (Fig. 5.2C). Therefore, we now derive the likelihood of the observed data given the parameters, which is then optimized to find the maximum likelihood estimates.

5.2.1 Derivation of the likelihood

We notice that the entire process of differentiation and marker delay on trees has the Markov property: Given the state of some cell i at time t, the subtree induced by this cell and time is independent of the remaining tree. This allows us to divide the problem into smaller subproblems, where we enumerate on a per cell basis all possibilities of differentiation points in an observed tree.

Consider the observed tree T in Fig. 5.3A. It has three marker onsets (in cells 3, 4 and 5) and with respect to cells, there are three possible differentiation scenarios leading to this tree: Either the root of the tree differentiated, or both its children differentiated or all three leave cells differentiated. Note that in each scenario, the point of differentiation within the cell is not fixed, just the cells that differentiated (this will be accounted for in section 5.2.2). The likelihood of the observed tree given parameters θ and η is the sum of likelihoods of the hidden trees, because these are competing alternatives:

$$\mathcal{L}(T|\theta,\eta) = \sum_{H\in\mathcal{H}(T)} \mathcal{L}(H|\theta,\eta)\ . \tag{5.6}$$

Note that the number of hidden trees grows quickly with the number of generations in the tree (it is doubly exponential in the number of generations, Aho and Sloane, 1973). However, as we will see later in section 5.2.3, we don't have to calculate the likelihood for each hidden tree separately but the calculations overlap substantially, rendering the calculations tractable even for large trees.

Now we derive the likelihood $\mathcal{L}(H|\theta,\eta)$ of a single hidden tree $H \in \mathcal{H}(T)$. Let us partition the hidden tree into the various subtrees D_i induced by the differentiating cells and a single tree U that only contains undifferentiated cells (see Fig. 5.3A). Then, due to

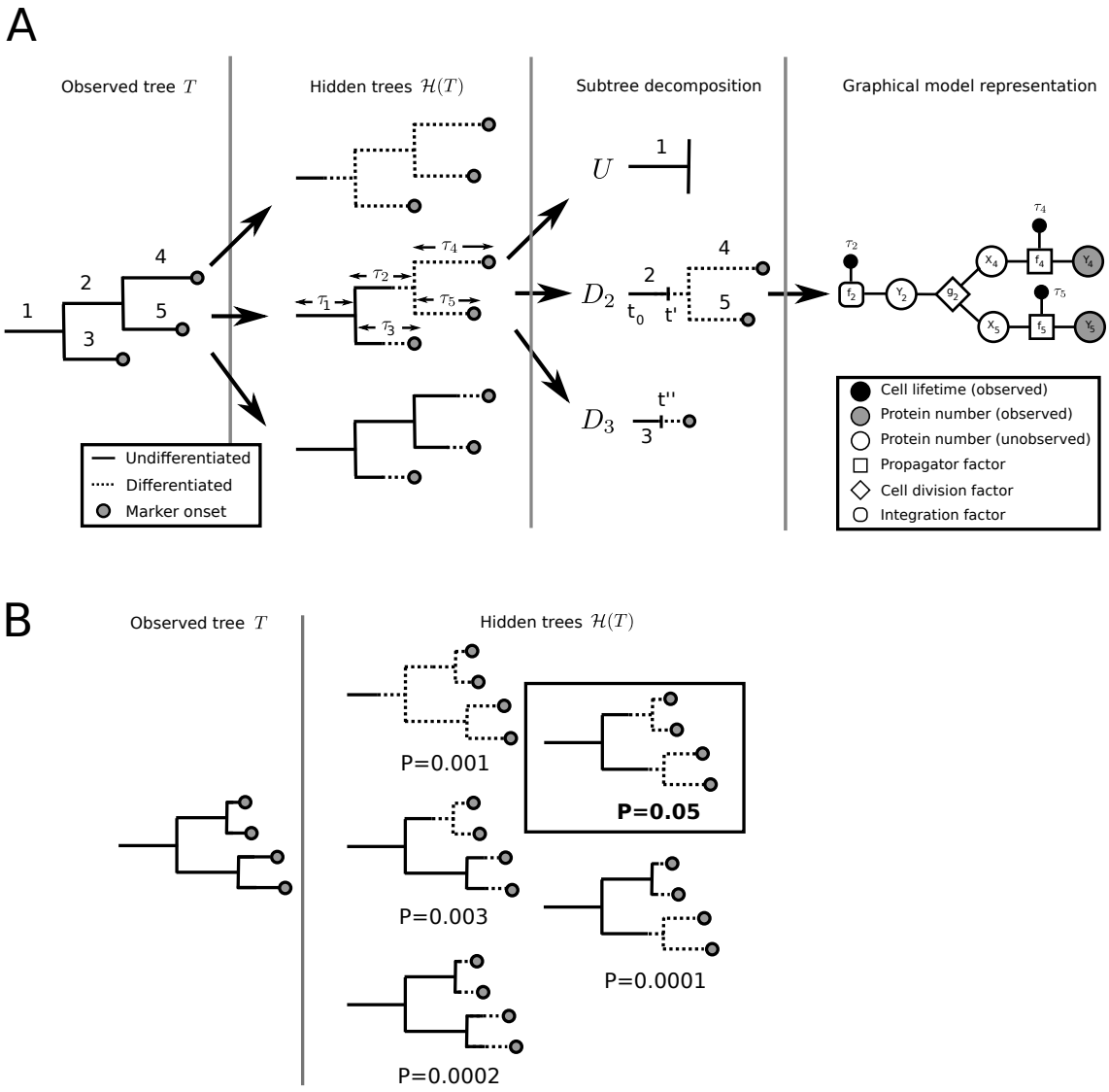


Figure 5.3: **Statistical inference on lineage trees.** A) For one observed tree T, several hidden trees $H \in \mathcal{H}(T)$ can be constructed. A particular hidden tree can be decomposed into a single tree U that contains only undifferentiated cells, and the set of subtrees D_i whose roots are differentiating at unknown timepoints (t', t''). To obtain its likelihood, each subtree D_i is represented as a graphical model and message passing is performed. For details, see main text. B) After estimating the parameters from data, we can predict the most likely hidden tree for an observed tree, by assigning probabilities to the hidden trees $\mathcal{H}(T)$.

the Markov property and our assumption of a point process of differentiation (see above), the likelihood of a hidden tree H factorizes into the likelihood of the tree U generated by the differentiation process and the product of likelihoods of the subtrees D_i generated by the delay:

$$\mathcal{L}(H|\theta,\eta) = \mathcal{L}(U|\theta) \cdot \prod_i \mathcal{L}(D_i|\theta,\eta) \ . \tag{5.7}$$

Note that the parameters θ also appear in the likelihoods for D_i as the root of these subtrees is still undifferentiated for some unknown time (see Fig. 5.3A). The first term is easy to compute because the process generating it has no memory as we assumed a decision to be a point process (see section 5.1.3). Therefore, we can treat the cells within the tree U independently. Consider a single cell c in this tree that is born at time ς_c (since movie start) and divides at time $\varsigma_c + \tau_c$. The probability that this cell does not differentiate in its lifetime $[\varsigma_c, \varsigma_c + \tau_c]$ is given by:

$$\mathcal{L}(c|\theta) = e^{-\int_{\varsigma_c}^{\varsigma_c+\tau_c} \lambda(t|\theta)dt} = e^{-a_0\cdot\tau_c - \frac{a_1}{2}\cdot(\tau_c^2 + 2\varsigma_c\tau_c)} \ , \tag{5.8}$$

where we have used the assumption that the hazard rate λ is a linear function of t (see Eq. 5.3). As all cells in U are undifferentiated and independent (point process), we get

$$\mathcal{L}(U|\theta) = \prod_{c\in U} \mathcal{L}(c|\theta) \ , \tag{5.9}$$

which is straightforward to calculate for any given U.

The factors $\mathcal{L}(D_i|\theta,\eta)$ in the second term of Eq. (5.7) are more difficult to obtain, as the delay process has memory and hence the individual cells of the subtree cannot be treated independently. Also, one has to account for the unknown time interval where the root of the subtree is still undifferentiated (see Fig. 5.3A).

5.2.2 Factor graph representation

We represent each tree D_i as a factor graph (see section 2.4) and obtain the likelihood $\mathcal{L}(D_i|\theta,\eta)$ by performing inference via message passing on the factor graph. For each non-root cell c in D_i, we create two variable nodes, one representing the state X_c of the cell c at its first timepoint, the other representing its state Y_c before division (see Fig. 5.3A). Furthermore, we have to introduce one additional variable node τ_c per cell representing the lifetime of that cell (shown as black circles in Fig. 5.3A). These three nodes associated with cell c are linked via the factor f_c (squares in Fig. 5.3A), that expresses the probability to reach state Y_c in time τ_c starting from state X_c. The factor f_c is the transition matrix or propagator of the associated Markov process: $f_c(X_c, Y_c, \tau_c) = P_{X_c\to Y_c}(\tau_c)$. It is obtained by solving the Master Equation (Eq. 5.4) numerically. Individual cells are linked via cell division factors g_c (shown as diamonds in Fig. 5.3A) that couple the last state of the mother (Y_c) to the first states of the daughters (X_{2c}, X_{2c+1}). For simplicity, we neglect partitioning of molecules at cell division and assume that this factor is the identity: $g_c(Y_c, X_{2c}, X_{2c+1}) = \delta_{Y_c,X_{2c}} \cdot \delta_{Y_c,X_{2c+1}}$. This ensures that both subtrees inherit the same

state from the mother cell. Note that this can easily be extended for example by binomial partitioning of molecules[1].

Finally, for the root cell r we have only a node representing the state before division and couple that to a factor f_r that implements the integration over the unknown timepoint of differentiation:

$$f_r(Y_r) = \int_{t_0}^{t_0+\tau_r} dt' \; \Phi(t_0 + t'|\theta) P_{0\to Y_r}(t_0 + \tau_r - t') \; .$$

The first term in the integral gives the probability that the differentiation occurred at time $t_0 + t'$ and the second term is the probability to reach protein number Y_r in the remaining cell cycle starting from zero proteins.

In this factor graph, which compactly describes the joint distribution over all variables $P(X_1, \ldots, X_n, Y_1, \ldots, Y_n, \tau_1, \ldots, \tau_n)$, certain nodes are observed (filled nodes in Fig. 5.3A): Let's assume we know the cell cycle length of all cells and we also know the state of the leave cells, because we observe the marker in these cells: $Y_l = x^*$, where l denotes any leaf cell.

To obtain the likelihood $\mathcal{L}(D_i|\theta, \eta)$, we apply the sum-product algorithm on this factor graph as introduced in section 2.4.3. Here, the sum-product algorithm is used to obtain the evidence of the data, i.e. $p(Y_{l_1} = x^*, \ldots, Y_{l_L} = x^*)$ for all L leaves of the hidden tree, which is equal to the desired likelihood

$$\mathcal{L}(D_i|\theta, \eta) = p(Y_{l_1} = x^*, \ldots, Y_{l_L} = x^*) \; .$$

For an example of the message passing procedure on a tree, see section 2.4.3.

5.2.3 Resolving the combinatorial complexity

Putting together Eqs. (5.6)–(5.9), we find the likelihood of an observed tree T as:

$$\mathcal{L}(T|\theta, \eta) = \sum_{H \in \mathcal{H}(T)} \left[\left(\prod_{c \in U_H} \mathcal{L}(c|\theta) \right) \cdot \prod_{d \in D_H} \mathcal{L}(d|\theta, \eta) \right] \; . \tag{5.10}$$

The sum over H in Eq. (5.10) consists of a large number of terms (it is double exponential in the number of cells (Aho and Sloane, 1973)), but the computationally expensive calculations take place in the second product term of Eq. (5.10): If a tree T has n_T cells, there are exactly n_T different delay trees D_i to consider, hence the number of expensive evaluation of $\mathcal{L}(D_i|\theta, \eta)$ scales linearly with the number of cells. For each tree T one can precompute the terms corresponding to all possible $D_i, i = 1 \ldots n_T$, which then only have to be added up in many different combinations via the sum over H in Eq. (5.10).

For example, consider a full binary tree with five generations, which has 458330 hidden trees[2]. Instead of evaluating $\mathcal{L}(D_i|\theta, \eta)$ in each of the 458330 hidden trees, we only need to evaluate it for each of the $2^5 = 32$ cells once.

[1] For example, by choosing $g_c(Y_c, X_{2c}, X_{2c+1}) = \binom{Y_c}{X_{c2}} p^{X_{c2}} (1-p)^{Y_c - X_{c2}} \cdot \delta_{X_{c2}+X_{c2+1}, Y_c}$

[2] For a full binary tree with n generations, the number of hidden trees is equivalent to the number y_n of strongly binary trees with generations $\leq n$. This is defined by the quadratic map $y_n = y_{n-1} + 1$ with $y_0 = 1$ and evaluates to 458330 for $n = 5$

Algorithm 4: Recursive algorithm to calculate $LA(i)$

Input: Cell i, tree T
Output: Set of ancestor cells $LA(i)$
Algorithm `LA(`i`,`T`)`

```
if i ∉ T then
    return ∅ ;                                  // if cell does not exist in T
else if mod(i,2) == 1 and (i − 1) ∈ T then
    return {c} ;                                // if c has a left sister cell
else
    m = ⌊c/2⌋ ;                                 // its mother cell
    return {c, LA(m, T)} ;                      // continue recursion in mother cell
end
```

However, even the enumeration of all $H \in \mathcal{H}(T)$ can be prohibitive and we now present a dynamic programming approach which avoids the explicit enumeration of hidden trees H. First we introduce two convenient abbreviations:

- $S(i) = \mathcal{L}(D_i|\theta, \eta)$.
  This is the probability of the subtree rooted in cell i, which we obtain by inference on the graphical model.
- $\tilde{P}(i) = \prod_{c \in LA(i)} \mathcal{L}(c|\theta)$.
  Here, $LA(i)$ is defined via Algorithm 4 and denotes a particular set of undifferentiated ancestor cells, such that every undifferentiated cell in the tree is contained in only one set $LA(i)$. $\mathcal{L}(c|\theta)$ denote the probability of cell c being undifferentiated (see Eq. 5.8).

Now we define a quantify $\kappa(i)$, which aids us in calculating the value of the overall sum in Eq. (5.10):

$$\kappa(i) = \begin{cases} S(i) \cdot \tilde{P}(i) & \text{if } i \in L \\ S(i) \cdot \tilde{P}(i) + \kappa(v) \cdot \kappa(w) & \text{if } i \notin L \text{ and } (i,v) \in E, (i,w) \in E \ , \end{cases} \tag{5.11}$$

and L denotes the set of leaves of T and E is the set of edges. If cell i is a leave, the probability is just the product of the subtree probability and the probability to be still undifferentiated (corrected for the multiple counting, hence $\tilde{P}$). If cell i is not a leave, its contribution to the combinatorics is the following: Either it differentiates and generates the subtree below (first term) or it just leads to the combinations of the two subtrees below (resulting in all possible combinations of the events in the subtree). Finally, we're only interested in

$$\mathcal{L}(T|\theta, \eta) = \kappa(1) \tag{5.12}$$

and we can use the recursive rule Eq. (5.11) to calculate the likelihood of an observed tree (Eq. 5.10).

We can now perform maximum likelihood estimation of the underlying model parameters (θ, η) given a set of observed trees $T_1, \ldots, T_n$:

$$\begin{aligned}(\hat{\theta}, \hat{\eta}) &= \underset{\theta,\eta}{\text{argmax}} \ \log \left[\prod_i \mathcal{L}(T_i|\theta, \eta) \right] \\ &= \underset{\theta,\eta}{\text{argmax}} \ \sum_i \log \left[\mathcal{L}(T_i|\theta, \eta) \right] \ . \end{aligned} \tag{5.13}$$

To solve the above optimization problem, we apply a standard multiple-restart (Latin Hypercube (McKay et al., 1979)) optimization routine.

5.2.4 Predicting the timepoint of differentiation

The final goal of our approach once we have learned the parameters $(\hat{\theta}, \hat{\eta})$ via Eq. (5.13) is to predict differentiation times and cells. For an observed tree T, we select the most likely hidden tree $\hat{H}$ from the set of all possible hidden trees $\mathcal{H}(T)$ according to

$$\hat{H} = \underset{H \in \mathcal{H}(T)}{\text{argmax}} \, \mathcal{L}(H|\hat{\theta}, \hat{\eta}) \ . \tag{5.14}$$

This immediately provides us with the information which cells most likely have differentiated (see Fig 5.3B). Note that the probabilities over hidden trees do not sum up to one. However, another hidden tree H' might have an only slightly smaller likelihood $\left(\frac{\mathcal{L}(H'|\hat{\theta},\hat{\eta})}{\mathcal{L}(\hat{H}|\hat{\theta},\hat{\eta})} \approx 1\right)$, making it difficult to choose what the "best" hidden tree is. This can be dealt with by making predictions only if the maximum in Eq. (5.14) is very distinct, e.g. if there is a five-fold difference in the likelihood of the best and second best hidden tree. Of course, one could also invoke a fully Bayesian treatment, not deciding on one "true" hidden tree but taking into account all hidden trees, weighted by their posterior probabilities.

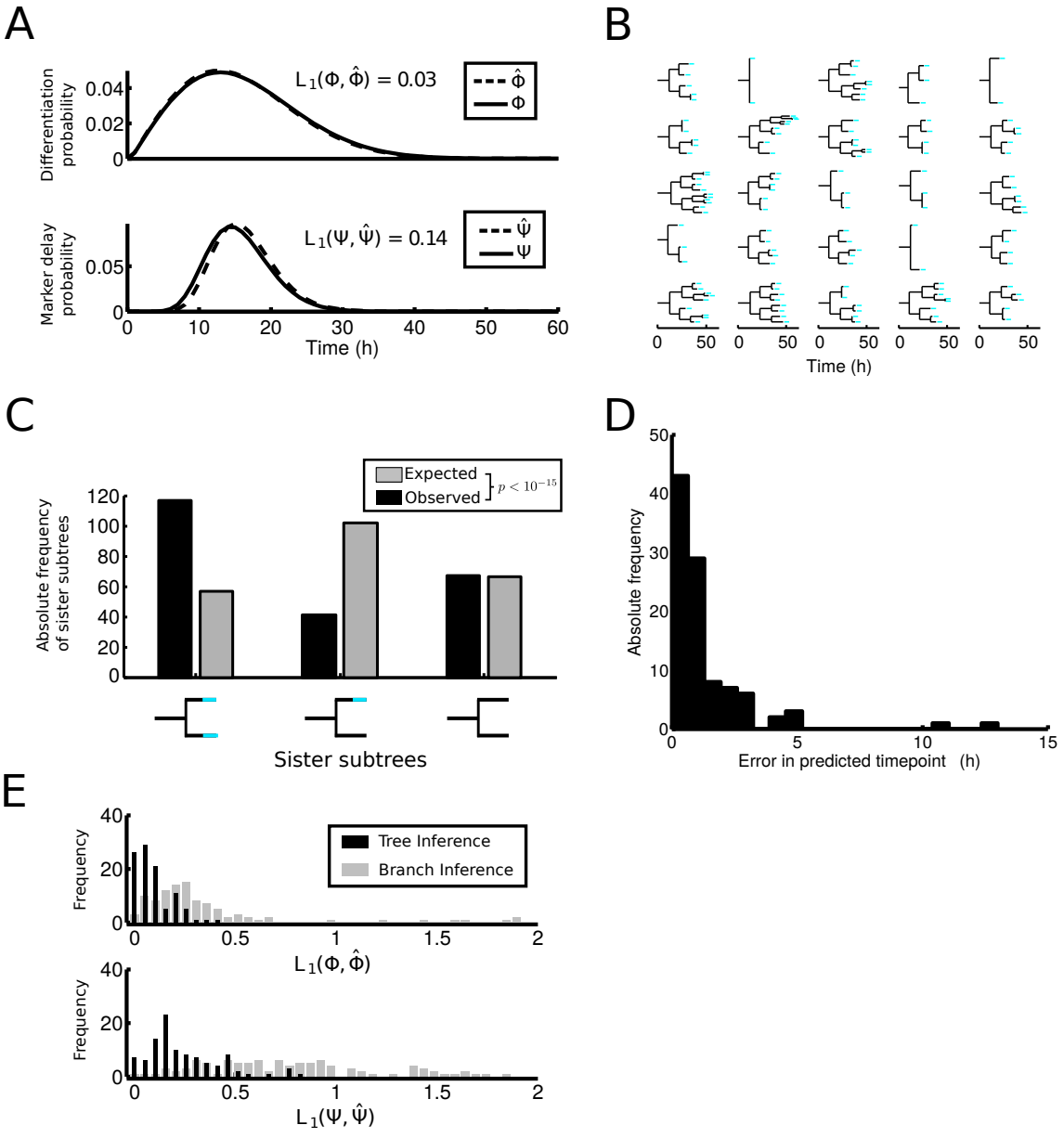
5.3 Application

In the following, we apply our proposed method to two synthetic datasets testing the validity of our approach. Finally, we use the method on data from blood stem cell differentiation to predict differentiating cells and whether these cells show differential PU.1 transcription factor dynamics.

5.3.1 Proof of principle

We now test the tree inference method on synthetic data. We first choose parameters θ and η, giving rise to the a particular differentiation and delay distribution via Eq. (5.3) and Eq. (5.5) (see Fig. 5.4A, solid lines). We generate 50 trees from those distributions as our observations (Fig. 5.4B shows a subset of 25 trees), so that we do not observe the two underlying processes directly, but only the marker onset.

To quantify the amount of correlations induced in sister cells due to the delay process, we apply the statistical test proposed in section 4.2.5, i.e. we fit a generalized linear


A
Differentiation probability
0.04
0.02
0
$L_1(\Phi, \hat{\Phi}) = 0.03$
$\hat{\Phi}$
Φ
Marker delay probability
0.05
0
$L_1(\Psi, \hat{\Psi}) = 0.14$
$\hat{\Psi}$
Ψ
0 10 20 30 40 50 60
Time (h)
B
0 50 0 50 0 50 0 50 0 50
Time (h)
C
Absolute frequency of sister subtrees
120 100 80 60 40 20 0
Expected
Observed
$p < 10^{-15}$
Sister subtrees
D
Absolute frequency
50 40 30 20 10 0
0 5 10 15
Error in predicted timepoint (h)
E
Frequency
40 20 0
Tree Inference
Branch Inference
0 0.5 1 1.5 2
$L_1(\Phi, \hat{\Phi})$
Frequency
40 20 0
0 0.5 1 1.5 2
$L_1(\Psi, \hat{\Psi})$

		Predicted	
		differentiating	not differentiating
Ground truth	differentiating	TP=187	FN=42
	not differentiating	FP=27	TN=792

Table 5.1: Confusion matrix for the predictions of differentiating cells in 100 genealogies in scenario of linear time-dependent differentiation and a single-gene delay. Training was performed on an independent set of 50 genealogies. (TP: true positive, FN: false negative, FP: false positive, TN: true negative)

model that tries to explain observed marker onsets in terms of external influences (in this case the only external influence is time since movie start). The frequencies of sister subtrees expected from this model are statistically different ($p < 10^{-15}$) from the observed frequencies (Fig. 5.4C). This shows that the correlations observed in the data extend beyond what can be explained by external variables and indicates the need for the model proposed in this chapter which can handle these cell intrinsic processes (the marker delay).

We fit our model by solving the optimization problem in Eq. (5.13) numerically (a single optimization run takes approximately 8 minutes), obtain a maximum likelihood estimate $(\hat{\theta}, \hat{\eta})$ and compute the corresponding differentiation $(\hat{\Phi})$, delay $(\hat{\Psi})$ distributions (see Fig. 5.4A, dashed lines). The estimated differentiation and delay distributions are very close but not identical to the true ones due to the finite sample size of $n = 50$ trees. As a simple measure to quantify this difference, we calculate the L_1-distance between true

Figure 5.4 *(facing page)*: **Inference of the differentiation decision from lineage trees can accurately reconstruct the underlying differentiation and marker delay dynamics.** A,B) For a given set of parameters θ, η, the differentiation probability distribution Φ (solid line, upper panel) and the marker delay probability distribution Ψ (solid line, lower panel) are shown. We simulate 50 trees (25 shown in B) from these parameters and apply the tree inference algorithm to obtain estimates $\hat{\theta}, \hat{\eta}$. The corresponding estimates of the differentiation and delay distributions, $\hat{\Phi}, \hat{\Psi}$ (dashed lines in both panel) agree well with the true distributions (solid lines in A), as quantified by their L_1 distance (0.03 and 0.14, respectively). C) Comparison of the observed and expected frequencies of sister pairs (both, one, or none differentiating) of the dataset in A-B). The differences are statistically significant ($p < 10^{-15}$, χ^2-test, see section 4.2.5). D) Histogram of the difference in time between predicted and true differentiation timepoints for an independent test set of 100 trees simulated from the distributions in A). E) Histograms of L_1 distances between true and estimated distributions for 100 randomly chosen parameter sets. While the tree inference is capable of reconstructing the underlying distributions accurately, resulting in small L_1 distances (black bars) using an algorithm based only on branches of lineage trees often fails to reconstruct the corresponding distributions (gray bars).

distribution (p) and estimated distribution ($\hat{p}$) as:

$$L_1(p, \hat{p}) = \sum_t |p(t) - \hat{p}(t)| \, .$$

In our toy example, we find a small distance between the estimated and true differentiation(0.03) distributions and a slightly larger distance for the delay (0.14) distributions.

We now test the predictions of our model trained with the 50 genealogies on a independent test set of 100 genealogies. For each of the 100 trees in the test set, we obtain its most likely hidden tree via Eq. (5.14), thereby predicting the differentiating cells. Comparing these predictions to the ground truth[3], we find that for 77 of 100 observed trees, we predict the correct hidden tree. In terms of single cells, we correctly recall 184 of 198 differentiating cells, while seven cells are falsely identified as differentiating (for the confusion matrix, see Table 5.1). Finally, we calculated the distance in time, between true and predicted timepoints of differentiation to see how accurately we can recover not only the differentiating cell but also the timepoint of this event. We found that the predicted timepoint is typically within 5 hours of the true timepoint of differentiation, and only in two cases was larger than ten hours (Fig. 5.4D). Hence, even if the predicted cell is wrong, the true differentiation event is within the timescale of a single cell cycle, which is 12 hours on average.

To systematically validate the algorithm's performance, we repeat the above analysis for a set of 100 parameters $(\theta, \eta)_i, (i = 1, \ldots, 100)$, randomly sampled from the intervals given in Table 5.2. We simulate 50 trees T_i for each $(\theta, \eta)_i$ and perform the maximum

	a_0	a_1	α	γ	x^*
lower bound	10^{-15}	10^{-15}	10^{-6}	10^{-15}	1
upper bound	10^{-7}	10^{-7}	10^{-1}	10^{-3}	150

Table 5.2: Parameter ranges considered in the optimization.

likelihood estimation using Eq. (5.13). For each set of trees T_i, we obtain estimates of the underlying differentiation and delay distributions ($\hat{\Phi}^i$ and $\hat{\Psi}^i$). A histogram of the L_1-distances between estimated and true distributions is shown in Fig. 5.4B (black bars). The distances in the differentiation distributions are generally small ($\text{Quantile}_{0.95} = 0.25$), Fig. 5.4B, upper panel). For the delay distributions these distances are sometimes larger ($\text{Quantile}_{0.95} = 0.70$, Fig. 5.4B, lower panel), suggesting that it is more difficult to extract these parameters from the data. As a comparison, we developed a similar framework for the estimation of the two processes based only on branches of lineage trees (see Appendix B) and applied it to the same 100 datasets. The resulting L_1 distances (Fig. 5.4B, gray bars) are larger ($\text{Quantile}_{0.95} = 1.50$ and $\text{Quantile}_{0.95} = 1.61$) compared to the L_1 distances achieved using the full tree information, indicating that the correlation structure observed in the trees is critical for inferring the underlying processes. Overall, the analysis suggests that we can reconstruct the underlying parameter from the observation with good accuracy.

[3]The ground truth for the 100 genealogies is available from the data generation process.

5.3.2 A cascade of genes

Until now we have used the very simple model of gene expression combined with a detection limit of marker onset to explain correlations in trees. However, taking into account typical gene expression parameters (Schwanhäusser et al., 2011) and reasonable detection limits (Schwarzfischer et al., submitted), one expects only short delays (in the range of several hours) between the start of expression and the marker detection. Short delays might still cause correlations in sister cells, for example if the expression starts very late in the cell cycle of the mother and is only completed in the two daughters. Correlations across multiple generations cannot be explained by this simple mechanism, but are more likely caused by slow dynamics or long cascades in the underlying gene regulatory network that trigger differentiation. Because the network dynamics are typically unknown, it is impossible to model this process and to learn its parameters.

Here, we now assess if our simple model can cope with a more realistic delay process consisting of a cascade of three genes (Fig. 5.5A):

$$\begin{aligned}
\emptyset &\xrightarrow{\alpha_A} A \xrightarrow{\gamma_A} \emptyset \\
\emptyset &\xrightarrow{\tilde{\alpha}_B(A)} B \xrightarrow{\gamma_B} \emptyset \\
\emptyset &\xrightarrow{\tilde{\alpha}_C(B)} C \xrightarrow{\gamma_C} \emptyset \,.
\end{aligned}$$

Upon differentiation, expression of the first gene in the cascade (gene A in Fig. 5.5A) is triggered, which in turn activates expression of its downstream target (gene B in Fig. 5.5A). Gene B in turn activates gene C, whose expression can be detected once crossing a detection threshold x^* (Fig. 5.5B).

Activation is governed by a Hill function, such that the production rate $\tilde{\alpha}_B(A)$ of gene B is an increasing function of the number of activator molecules[4]. Degradation rates are $\gamma_A = \gamma_B = \gamma_C = 0.1$ h^{-1}, maximal synthesis rates are $\alpha_A = \alpha_C = 100$ h^{-1}, $\alpha_B = 22$ h^{-1}, cooperativity $n = 5$ and dissociation constants $K_A = 800, K_B = 100$. The dynamics of this stochastic process lead to a long and heterogeneous delay ranging from 35 to 60 hours after differentiation.

We now simulate 50 genealogies from a time-dependent differentiation process (parameters as in Fig. 5.4A-D) and the three-gene cascade (a sample of nine genealogies is shown in Fig. 5.5C). As before, we fit the model to the data via Eq. (5.13). Here, a single optimization run takes approximately 60 minutes due to the much larger genealogies (computation time scales linearly with the number of cells, see section 5.2.3). Note that the model still assumes a single gene delay process. Comparing the estimated to the true distributions, we observe good agreement (Fig. 5.5D). We note a slight deviation for the delay distribution, which arises due the difference in delay processes (a single gene as opposed to a cascade).

We evaluate the performance of the fitted model on an independent test set of 100 genealogies. For 91 of 100 genealogies, the mostly likely hidden tree (obtained via Eq. 5.14) indeed corresponds to the true underlying differentiation scenario. Performance in terms

[4] $\tilde{\alpha}_B(A) = \alpha_A \cdot \frac{K_A^n}{K_A^n + A^n}$ and $\tilde{\alpha}_C(B) = \alpha_C \cdot \frac{K_B^n}{K_B^n + B^n}$, where K_x is its dissociation constant of the activator, n the cooperativity, and α_x the maximal synthesis rate.

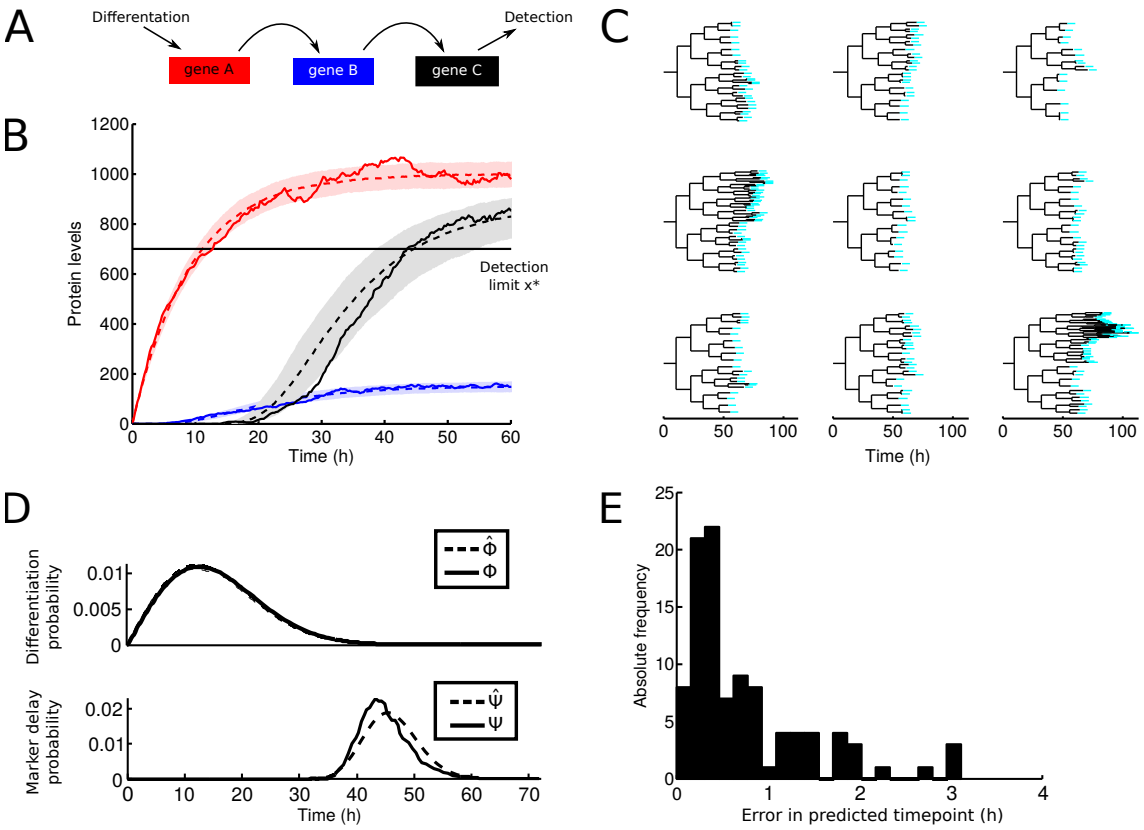


Figure 5.5: **The model can accurately fit delays arising from cascades of genes.** A) In a three-gene cascade, the upstream gene A (red) is activated by the differentiation event and activates its downstream target gene B (blue), which then activates the marker gene C (black) which is detectable. B) Dynamics of the process depicted in A). A single realization is shown as solid lines and shaded areas represent 5-95% regions across 1000 realizations. Time is relative to the differentiation event at $t = 0$. The detection threshold of gene C is indicated a horizontal black line. C) Nine genealogies simulated from a linear time-dependent differentiation process (parameters as in Fig 5.4) but with a delay arising from a three-gene cascade. D) Estimated differentiation and marker delay probability distributions (dashed lines) from 100 simulated observed trees agree well with the true distributions (solid lines). The true delay distribution is calculated from 1000 stochastic simulations. E) Histogram of the error in the predicted timepoint of differentiation. All predicted timepoints in the 100 genealogies from D) are within three hours of the true differentiation timepoint.

of single cell prediction is summarized in Table 5.3. Note that due to the much longer delay compared to Fig. 5.4, many more non-differentiating cells are present, which are mostly classified correctly. In terms of time difference between predicted and actual timepoint of differentiation, we find that the predicted timepoint is always within 3 hours of the true timepoint (Fig. 5.5E) and the misclassifications in Table 5.3 happen close to cell division: For example, the mother cell might differentiate at the end of its cell cycle, but the methods predicts that its daughter cells difference at the beginning of their cell cycles.

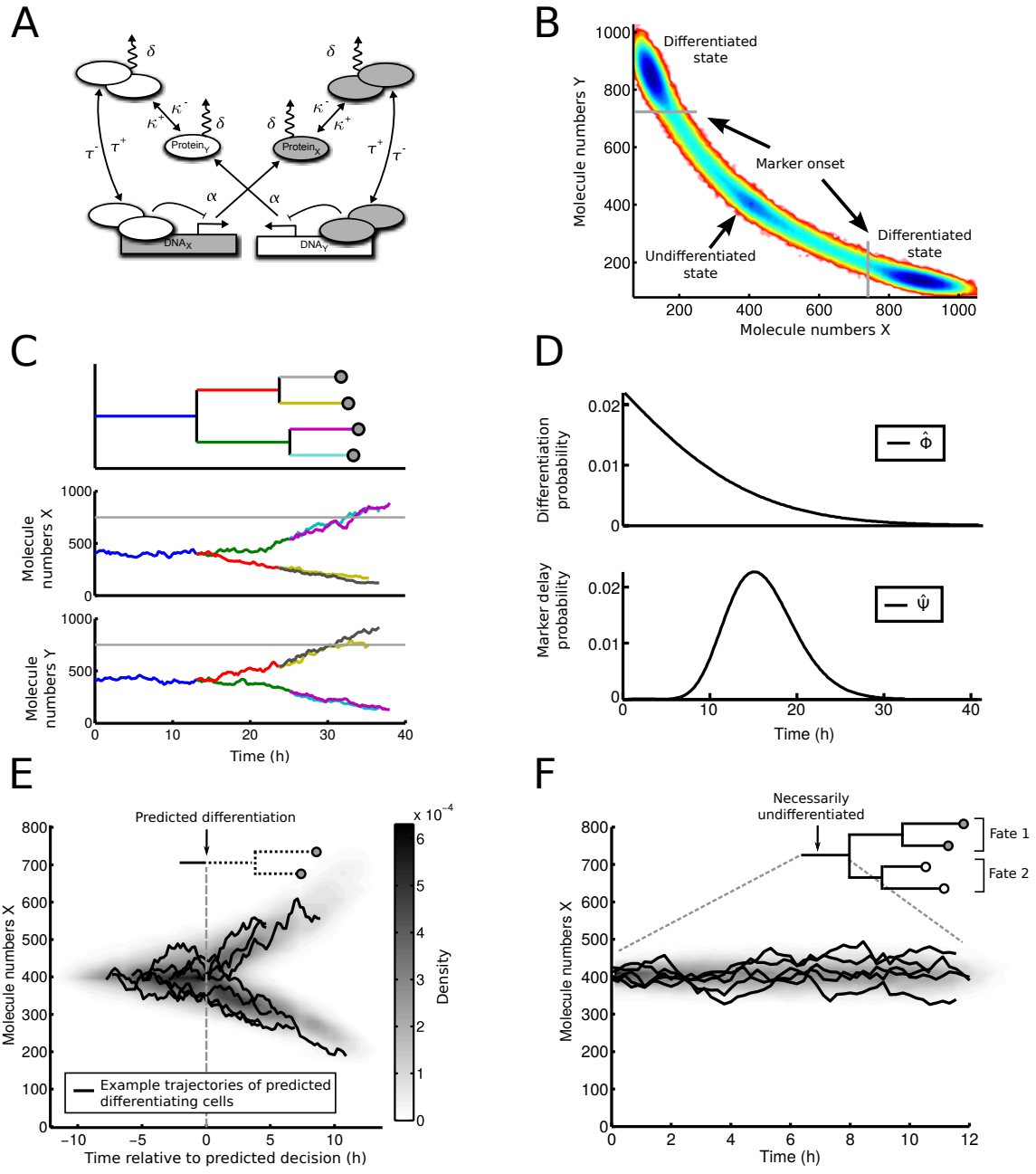
		Predicted	
		differentiating	not differentiating
Ground truth	differentiating	TP=180	FN=18
	not differentiating	FP=9	TN=5401

Table 5.3: Confusion matrix for the predictions of differentiating cells in 100 genealogies in scenario of linear time-dependent differentiation and a three-gene cascade delay. Training was performed on an independent set of 50 genealogies. (TP: true positive, FN: false negative, FP: false positive, TN: true negative)

5.3.3 Toggle switch model

Now we ask whether our method can be used to infer the differentiation decision in a more realistic model of cell fate decisions, deliberately ignoring its mechanistic details. To address this, we implement a toggle switch composed of two mutually repressing transcription factors as an underlying cell fate decision mechanism (see Fig. 5.6A). This model was already discussed in chapter 3 (see Eqs. 3.36–3.39).

The model exhibits three stable states that can be seen as wells in the quasi-potential of the system (see Fig. 5.6B): One state, where both proteins are expressed at similar levels is associated with an undifferentiated cell. In the two other states, either one or the other protein is strongly upregulated, thereby repressing the other. These two states corresponds to mutually exclusive differentiated lineages. Differentiation occurs via noise driven transitions from the undifferentiated to one of the differentiated states. Using Gillespie's algorithm (see chapter 2 and Gillespie, 1976) to obtain sample trajectories from the associated Chemical Master Equation, we simulate trees from this toggle switch model (see Fig. 5.6C): The root cell of each tree starts in the undifferentiated state, from where it evolves over time according to the laws of the underlying toggle switch model until it divides. The cell cycle time is drawn from a log-normal distribution with a mean cell cycle time of 12h and a standard deviation of 1h. Upon division, two identical daughter cells are created that inherit the state of the mother cell. We ignore asymmetric partitioning of molecules at cell division for simplicity. The daughter cells evolve independently of each other according to the toggle switch dynamics. Once a cell arrives at one of the differentiated states (empirically defined as crossing 750 molecules of either protein, gray lines in Fig. 5.6B,C), we stop its simulation and annotate this cell as being differentiated (gray circles in Fig. 5.6C). We do not distinguish between the two differentiated states when annotating the marker, emulating a "loss of bipotency" marker, such as LysM::GFP


A
B
Differentiated state
Marker onset
Undifferentiated state
Differentiated state
Molecule numbers Y
Molecule numbers X
C
Molecule numbers X
Molecule numbers Y
Time (h)
D
Differentiation probability
Marker delay probability
Time (h)
E
Predicted differentiation
Example trajectories of predicted differentiating cells
Molecule numbers X
Time relative to predicted decision (h)
Density
F
Necessarily undifferentiated
Fate 1
Fate 2
Molecule numbers X
Time (h)

in hematopoiesis (Rieger et al., 2009). It is apparent from Fig. 5.6C that the decision of differentiation, that is the time when the system leaves the undifferentiated state, occurs much earlier than the observation of our differentiation marker (gray circles in Fig. 5.6B). This delay arises because the transition from the undifferentiated to the differentiated states requires some non-negligible time, in which a cell might divide several times, causing correlated behavior in terms of marker onset between related cells.

In order to test our method, we assume that we cannot directly observe the process that drives the differentiation (Fig. 5.6C upper and middle panel), but only observe whether a cell has finally arrived at one of the differentiated states (Fig. 5.6C bottom panel). We apply the tree inference method and learn the parameters θ, η of our model from 100 simulated trees. Note that these parameters do not correspond to the parameters of the toggle switch model itself, but to an abstract description of the differentiation and delay process. In Fig. 5.6D, we show the distributions Φ and Ψ learned from the data. The differentiation distribution is almost exponential, indicating that the process of leaving the undifferentiated state is well described by a point process with constant rate (in agreement with findings from chapter 3). Intuitively, this rate reflects the frequency of a large fluctuation that pushes the cell out of the undifferentiated state. Inspecting the delay distribution Ψ we find an average delay of 15 ± 4 hours which visually coincides with the typical transition times we observe in Fig. 5.6C.

Now, we look at the underlying timecourses of the toggle switch in context of the

---

Figure 5.6 *(facing page)*: **Inferring the differentiation decision in lineage trees with an underlying toggle switch model recovers the change in the unobserved underlying dynamics.** A) A symmetric model of two mutually inhibiting transcription factors X and Y. Proteins X and Y are created with rate $f(Y)$ and $f(X)$, respectively. Transcription and translation are lumped together and mutual inhibition is incorporated in the Hill-type synthesis rates f. Proteins decay with rate γ. B) The model in A) gives rise to one undifferentiated (central) and two differentiated states (upper left, lower right) that can be identified as wells in the quasi-potential $(-\log(P))$ of the system. We define a cell to be differentiated and hence marker positive once it enters the basin of attraction of a differentiated state (gray lines). C) Genealogies are generated from the toggle switch model. Shown are the resulting genealogy (top panel) as well as the timecourses of both factors X and Y (middle panel and lower panel). Individual cells are color coded across panels. D) Estimated differentiation and marker delay probability distributions from 100 simulated observed trees. The differentiation is close to an exponential and the mean marker delay is 15h. E) Trajectories of predicted differentiating cells are centered on the timepoint of differentiation ($t = 0$). Single trajectories (black lines) as well as the density across all predicted cells (color map) are shown. The predicted differentiation event coincides with the branching of the toggle switch dynamics. F) Trajectories of cells whose progeny differentiates into both fates (see inset). Different fates are indicated by black and white circles. By definition, these cells are undifferentiated. Parameters used for simulation are $\gamma_X = \gamma_Y = 0.7\ \mathrm{h}^{-1}$ (degradation rate), $\alpha_X = \alpha_Y = 700\ \mathrm{h}^{-1}$ (maximal synthesis rate), $n = 2$ (cooperativity), and $K_X = K_Y = 330$ (dissociation constants). See also Eqs. (3.36)–(3.39).

predicted differentiation events. For a set of 500 observed trees (independent of the 100 trees used for inferring the parameters), we predict the differentiating cells as well as the timepoint of differentiation within those cells based on Eq. (5.14).

The timecourses of the two proteins X and Y in these predicted cells are then aligned at the predicted differentiation point ($t = 0$ on the x-axis of Fig. 5.6E). In Fig. 5.6E, we show some example cells and their corresponding aligned trajectories (black lines) as well as the density over all trajectories of predicted cells. These trajectories show the distinct pattern of state transitioning. In the beginning ($t < 0$), cells reside in the undifferentiated central state. Later, some cells exit this state by increasing the abundance of protein X, differentiating into one lineage, whereas other cells leave the central state by decreasing the abundance of X and thereby differentiate into the other lineage. Notice that the predicted timepoint of differentiation coincides with the point where trajectories split up in the one or the other direction.

For cells that give rise to both differentiated cell types in their progeny, we definitively know that they are still undifferentiated, assuming differentiation to be irreversible (see Fig. 5.6F inset). We use these cells as a further validation of our prediction. In Fig. 5.6F, we plot the trajectories of those cells. We see that an undifferentiated cell samples all states $X \in [350, 450]$, similar to the predicted cells at $t < 0$ in Fig. 5.6D. Therefore, we conclude that the cells in Fig. 5.6E before the predicted differentiation ($t < 0$) are indistinguishable from undifferentiated cells.

We performed similar analyses for a different parameter set (see Appendix Fig. C.1) and for a toggle switch coupled to a three-gene marker cascade (see Appendix Fig. C.2) and obtained the same results: The predicted cells are the ones where the balance of the two toggle switch proteins is broken and the system tilts towards one or the other differentiated state.

Without knowledge of the underlying process, but just from the correlations of onsets in the trees and assuming a linearly time dependent differentiation hazard, our method has identified cells where the dynamics of the toggle switch undergo large changes that lead to a state transition. Therefore, it seems plausible that, even if the underlying dynamics are much more complicated, our simple model of independent differentiation and a delay due to marker expression can be applied and useful predictions about the underlying dynamics can be made.

5.3.4 Blood stem cell differentiation

One decision within the hematopoietic differentiation tree is the choice of hematopoietic stem and progenitor cells (HSPCs) between the megakaryocytic-erythroid (MegE) and the granulocyte-macrophage (GM) lineage (see Fig. 3.8A). Based by diverse experimental indications (see e.g. (Krumsiek et al., 2011) for an overview), the mutual binding of lineage-specific transcriptions factors PU.1 and Gata1 inspired toggle switch models predicting TF dynamics during this decision (Roeder and Glauche, 2006; Huang et al., 2007; Bokes et al., 2009; Strasser et al., 2012).

Here, we test whether PU.1 levels change during the differentiation decision as determined by our tree inference algorithm. To that end, we use sorted HSPCs (see Hoppe et al., in revisionfor experimental details) from genetically engineered mice, where PU.1 has

been tagged with enhanced yellow fluorescent protein (eYFP). PU.1 levels are determined by tracking single cells, where the definite GM lineage choice is read out via manual annotation of CD16/32 expression via antibody staining (Fig. 5.7A), quantifying PU1.eYFP intensities and mapping intensities to absolute molecule numbers using quantification of Western blot analysis (for details, see Schwarzfischer et al., submitted). In a typical branch of a differentiating HSPC tree, both the number of PU.1eYFP proteins, as well as the cellular PU.1eYFP concentration rise before CD16/32 onset (Fig. 5.7B). We use our tree inference method to predict the most likely differentiation timepoint in each tree, sort cells accordingly into generations before, at, and after the predicted differentiation and fit the slope of the PU.1eYFP concentration in all cells to quantify protein production during differentiation. We find no significant difference in PU.1eYFP production between cells at predicted differentiation and one generation before ($p = 0.87$, ranksum test), or one generation afterwards ($p = 0.72$, ranksum test). In contrast, PU.1eYFP production is significantly higher in cells with CD16/32 onset as compared to cells one generation before, at, and and after the predicted differentiation ($p = 0.011$, $p = 0.009$, $p = 0.018$, respectively). This is in sharp contrast to the PU.1 dynamics in the toggle switch model (see Fig. 5.6): Here, PU.1 production clearly changes in cells that are predicted to differentiate ($p = 8 \cdot 10^{-11}$, ranksum test). This discrepancy rejects the involvement of PU.1 in the MegE vs GM lineage decision via the proposed toggle switch mechanism.

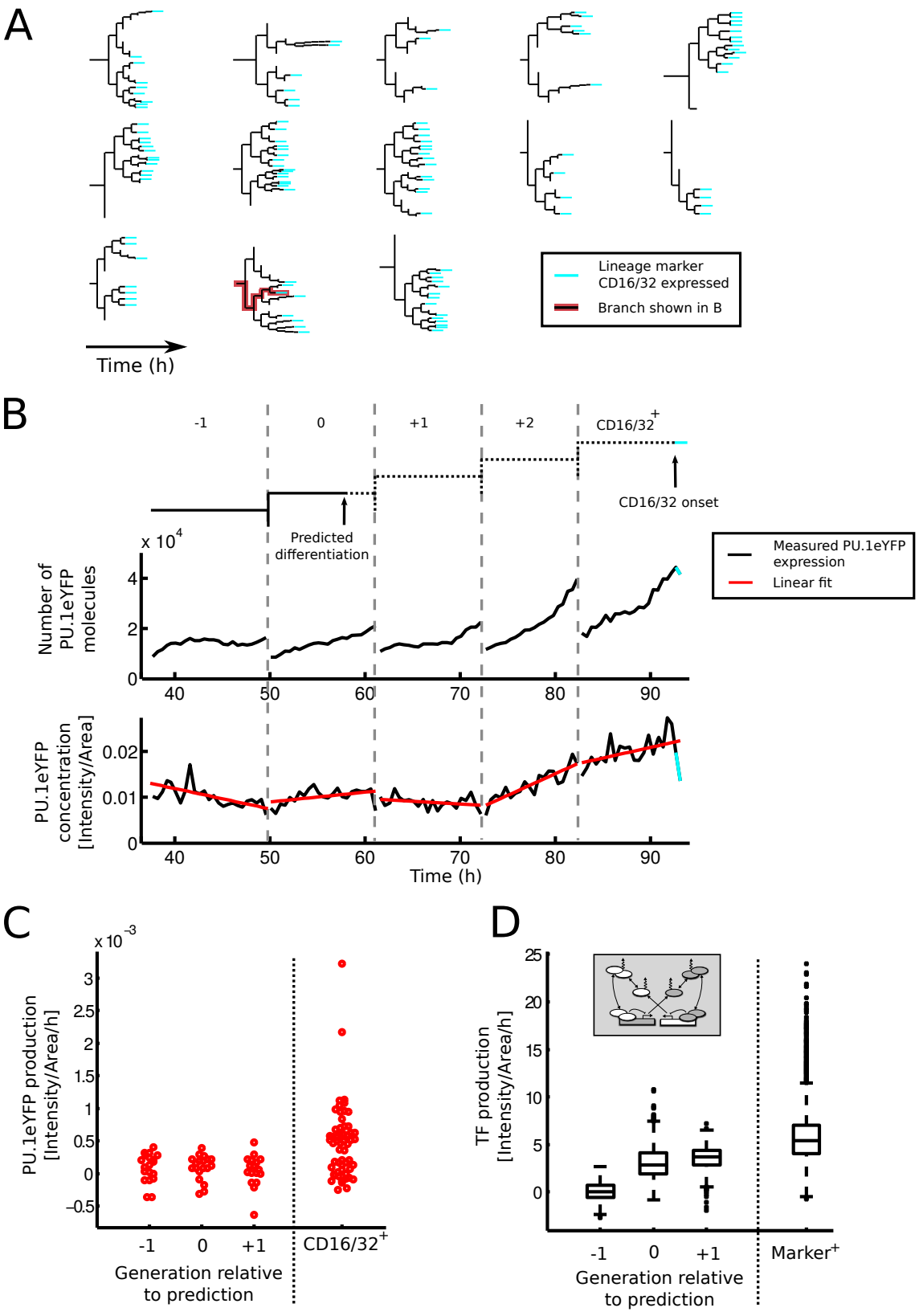
5.4 Discussion

In this chapter, have developed a computational method to infer cell fate decisions in genealogies from correlated marker onsets. We showed for a simple toy example, that the tree information is essential to make reliable predictions about the timepoint of decisions and also demonstrated the this simple model of independent differentiation and marker delay is capable to explain data generated from a more complicated differentiation model, a toggle switch of two mutually inhibiting genes.

Having established that the model can capture the cell fate dynamics of an underlying toggle switch system, we applied it to genealogies of differentiating hematopoietic stem and progenitor cells to infer the timepoints of cell fate choice from annotated onsets of

---

Figure 5.7 *(facing page)*: **In contrast to toggle switch model predictions, PU.1 expression does not change in hematopoietic stem cells at the inferred differentiation timepoint.** (A) Graphical representation of the 13 trees used to infer differentiation timepoints. Manually annotated CD16/32 onset is labeled in cyan. (B) We linearly fit the PU.1eYFP expression in each cell before, at, and after the inferred differentiation timepoint (red lines). (C) PU.1eYFP production is comparable (p-value $\gg 0.05$, rank-sum test) in cells one generation before, at , and one generation after the predicted differentiation decision. In cells with annotated CD16/32 onset, a sub-population markedly increases PU.1eYFP levels. (D) In contrast to the experimental data, the toggle switch model predicts a clear change of PU.1 production at the differentiation decision. Here, the fitted slope of PU.1 concentrations in differentiating cells is significantly higher as compared to undecided cells one generation before (($p = 8 \cdot 10^{-11}$),ranksum test).


A
Lineage marker
CD16/32 expressed
Branch shown in B
Time (h)
B
-1
0
+1
+2
CD16/32+
CD16/32 onset
Predicted differentiation
x 10^4
Number of PU.1eYFP molecules
Measured PU.1eYFP expression
Linear fit
PU.1eYFP concentration [Intensity/Area]
Time (h)
C
x 10^-3
PU.1eYFP production [Intensity/Area/h]
Generation relative to prediction
CD16/32+
D
TF production [Intensity/Area/h]
Generation relative to prediction
Marker+

CD16/32. Contrary to the prediction from a toggle switch model including PU.1, the inferred timepoints of differentiation precede the change in PU.1 expression by several generations. This suggests that PU.1 is not actively involved in the GM/MegE lineage decision, but acts after the decision has been made, as suggested also by Hoppe et al., in revision. Considering the rare occurrence of genealogies where both cell fates (GM and MegE) are observed (10%, see Hoppe et al., in revision) and the late onset of markers (approximately 5 generations after moviestart, see Fig. 5.7A), this also argues for an early lineage choice and a delay observation via markers: Let us assume the each cell individually decides its fate at the timepoint of marker onset (with probability p it differentiates into a GMP, with probability $1-p$ it differentiates into a MegE-progenitor). The probability of observing both fates with these 32 cells[5] is close to 1; even in the extreme case when $p = 0.9$ and almost all cells become GMPs, the probability to observe both fates within the same tree is still 0.97 in stark contrast to the observed 10%. This discrepancy also argues for a cell fate decision several generations before the marker onset, otherwise the lack of genealogies containing both fates remains unexplainable.

We focused on the observation of only one cell fate marker, but often two or more markers can be quantified Hoppe et al., in revision, that can have different delay dynamics. One can split up the observed trees into the ones that only contain one marker and into the ones that contain only the other marker and discard the trees that have both markers. Then, two separate models can be trained for both groups. However, as we showed in Fig. 5.6F, the trees containing both marker onsets are particularity informative: For some cells, one can a priori tell that they must be undifferentiated. To include this information, one has to extend the hidden tree enumeration depicted in Fig. 5.2C to account for both possible fates and apply the graphical model to a subtree D_i (Fig. 5.3) with one of two sets of parameters, depending on whether D_i has the one or other type of marker onsets in the leaves.

For the differentiation process we assumed that is is a linear function of time. In general, differentiation can depend on other external factors, e.g. spatial interactions between cells. For example, it is well known that in vivo, the blood stem and progenitor cells interact with niche cells, e.g. with osteoblasts, and that this interaction influences the fate of the progenitor cells (Wang and Wagers, 2011). Therefore, it is likely that also in vitro experiments, spatial dependencies play a role for cell fate decision. As a next step, one has to extend the proposed method to account for these spatial interactions, such as cell-density dependent differentiation. This is straight forward to incorporate, because spatial location and cellular density can easily be quantified from time lapse microscopy data. However, this introduces more unknown parameters and the benefit of these more complex models has to be rigorously evaluated using model comparison techniques to avoid overfitting.

We modeled the marker delay as a simple stochastic gene expression. This is certainly a simplification but due to a lack of knowledge about the internal processes more complex models are only speculative. In fact, this simple gene expression model captures the most relevant features of such a process: an average delay time, an overall variance between cells and a mechanism resembling (epigenetic) inheritance at cell division. Additionally, we neglected partitioning of molecules at cell division. However, we can simply exchange

[5] which is $\sum_{i=1}^{31} \binom{32}{i} p^i (1-p)^{32-i}$; the first summand is the probability of one GM and 31 MegE cells, the second term is the probability for 2 GM and 30 MegE cells, etc.

the cell division nodes g_c in Fig. 5.3A, that are implemented as identities by more complex functions, modeling either deterministic halving or even stochastic partitioning of proteins at cell division (Huh and Paulsson, 2011).

Throughout the chapter, we assumed that the observed marker that reports cell fate is a fluorescence signal, either intracellular (e.g. GFP-fusion) or extracellular (live antibody staining), that is triggered downstream of the decision. However, the proposed method is not limited to fluorescence markers, but can in principle be applied to any observable feature of a cell that is thought to be relevant for a particular cell fate. For example, one can quantify and use cell morphology as a marker for cell fate (Cohen et al., 2010; Held et al., 2010) or use cell survival and cell death as readout to learn if apoptosis is initiated only in the dying cell or was already initiated in one of its ancestor cells.

Provided its extendability and generality, we are confident that the proposed method can be applied to a wide range of cellular decision problems and that it will support the analysis and understanding of lineage tree data that is becoming more and more popular in single cell biology.

Chapter 6

Summary and outlook

In this thesis, I presented and analyzed models for cell fate choice in differentiating stem and progenitor cells and developed methods to link those models to experimental data. Of particular biological interest was the role of the two transcription factors PU.1 and Gata1 in hematopoietic cell fate choice. Those two factors form a genetic toggle switch and serve as the current (but unproven) paradigm of cell fate choice on a molecular level. Therefore, the toggle switch is a recurring theme in this thesis and is extensively analyzed. Motivated and driven by single cell time-lapse data from the Schroeder lab (Department of Biosystems Science and Engineering, ETH Zürich), the presented work put emphasis on single cell models, their benefits and challenges. Single cell data and hence single cell models are in fact necessary to study the dynamics of these cell fate decisions: The underlying dynamics are inherently multimodal, i.e. progenitor cells choose either one fate or the other and with potentially different dynamics. For example, GM-fated blood progenitors upregulate PU.1 whereas erythroid-fated cells downregulate PU.1, and averaging over the entire population will yield non-representative intermediate PU.1 dynamics.

In chapter 3, we introduced a stochastic model of the toggle switch without autoactivation and cooperative binding, but including an mRNA stage in the expression process. While a deterministic model is monostable in the absence of autoactivation and cooperativity of inhibition, stochastic models of the toggle switch were shown to be multimodal (Lipshtat et al., 2006). In these models, transcription and translation are lumped into a single synthesis reactions (one-stage gene expression), which is justified by a much faster timescale of mRNA turnover e.g. in bacteria. Due to the mounting evidence that in mammalian cells, mRNA and protein turnover of transcription factors happen on a similar timescale (Schwanhäusser et al., 2011) and the fact that small mRNA numbers introduce strong stochastic fluctuations that propagate even in the presence of high protein numbers, we developed a toggle switch model explicitly accounting for the mRNA stage of gene expression. We showed that inclusion of the mRNA stage indeed changes the dynamics compared to one-stage toggle switches. While the one-stage model is trimodal and most probability mass is located in a deadlock state where both factors are expressed in low levels, in our two-stage model, this state splits into two intermediate states and most of the probability mass shifts to the two states where either one or the other factor is strongly expressed. Interestingly, in the context of cell fate choice, the intermediate states can be associated with undecided states, which are however already biased towards one or the

other fate. This phenomenon was for example observed by Chang et al. (2008) in blood progenitor cells: While the progenitor population was provably undifferentiated, within the population two subtypes existed that were shown to be biased towards either myeloid or erythroid fate. Furthermore, we studied the process of state transitions within this toggle switch model, and showed that the transitions are initiated by the occurrence of a few elementary reactions, e.g. repressor unbinding followed by a single mRNA synthesis. Hence, we could approximate the transition as a point process and provided analytical expressions for the transition rates in terms of elementary reaction rates. Next, we showed how the dynamics of a toggle switch can be linked to existing time-lapse microscopy data of differentiating granulocyte-monocyte progenitor cells in permissive conditions. Using the annotated expression of the "loss of bipotency" marker LysM as a proxy for the timepoint of differentiation, we extracted the timing of the cell fate decision from the genealogies and used Approximate Bayesian Computation to fit a stochastic toggle switch model to that data. The model could indeed fit the observed timing of differentiation and predicts that the protein degradation rate of one factor as to be roughly one order of magnitude smaller than the other, meaning that cells commit to the one fate much faster than to the other. Unfortunately, with the existing data it was not possible to verify this hypothesis, because no cell fate specific markers are contained in the dataset, and LysM expression only indicates that a cell is no longer bipotent but does not provide information about the chosen fate itself.

For chapter 4, we digressed from the toggle switch as a molecular mechanism of cell fate choice and investigated how the influence of cell extrinsic variables, such as local cell density or nutrient concentration on cell fate choice can be detected. For simplicity, we considered only one cell fate, where for example cells are either undifferentiated or differentiated. We assumed the cell fate choice to be a point process (as motivated by the stochastic toggle switch), whose rate is now a function of external features. We showed how this rate can be reconstructed from genealogies using a non-parametric estimator, but also pointed out its drawbacks: (i) estimating the rate as function of multiple variables is infeasible with limited sample size, (ii) indirect effects due to correlations of variables cannot be identified. In turn, we estimated the rate via generalized linear models equipped with regularization to remove indirect effects due to correlated features. Furthermore, we proposed to use the correlations in terms of cell fate choice between genealogically related cells in order to validate of the model, i.e. whether the model of the transition rate can or cannot reproduce the observed correlation patterns. We demonstrated the success of our approach on different synthetic datasets and illustrated exemplarily how the interaction kernel of local cell density could be reconstructed from that data. Next, we tested how our method performs for varying numbers of samples and amount of cell tracking errors to give guidelines for future experiments. Although the analysis was performed for specific sets of parameters, one obtains already a rough estimate of requirements: a few thousand observed transitions (in our case corresponding to less then 100 genealogies) and a tracking error of less than 5%. Note that the chosen linear dependence of the rate on features is already a challenging scenario: Due to the gradual linear change of the rate with respect to the feature, the transition events are only weakly correlated to the underlying features. Non-linear relationships, where the rate abruptly changes by a large amount with respect to an underlying feature (e.g. a step function) are easier to detect, as there is clear link

between the transition event and the feature.

In chapter 5 we merged the previous ideas on mechanistic, cell-intrinsic and cell-extrinsic mechanisms of cell fate choice. We assumed that differentiation itself is a point process with a rate depending on external factors, but relax the assumption of the cell fate marker immediately reporting the cell fate choice. Therefore we accounted for the fact that the marker gene is upregulated most likely only as a delayed consequence of the newly established fate, but not immediately when the cell fate choice is made. This delay, which is modeled by a simple gene expression process, can cause correlated marker onsets beyond what is expected from external influences (as indicated by the test proposed in chapter 4). We developed a likelihood-based inference method that learns the model parameters from genealogies and show that using the tree structure is crucial to identify the correct parameters. Furthermore, we used the fitted model to predict the timepoints of differentiation within the observed genealogies. Testing the method on various synthetic datasets, we showed that the same simple model is also capable of predicting the correct timepoints of differentiation in datasets generated from more complex models of differentiation, e.g. a toggle switch, or a toggle switch coupled to a cascade of genes.

Before proceeding to the application of the method to the PU.1/Gata1 data, it is insightful to discuss how the our model relates to previously published models on cell fate choice via a toggle switch. In the work of Huang et al. (2007), the cell fate decision is modeled via a toggle switch assuming deterministic dynamics. In a deterministic system, a state transition cannot occur spontaneously as the cell will always converge to the stable state, in whose basin of attraction the cell is located. Therefore, Huang et al. (2007) assumed a gradual change of a system parameter (the strength of autoactivation) over time, eventually leading to a bifurcation. This destabilizes the progenitor state and leads to a transition into one of the two differentiated states which the cell reaches after a given period of time (approximately 72h in their experiments). Our model of a point process decision and a delay can be viewed as a generalization of the model by Huang et al. (2007): We would consider the change of the activation strength as an external time-dependent influence, which modulates the differentiation rate λ. For example, to recapitulate the scenario by Huang et al. (2007), one would choose $\lambda(t) = \Theta(t - t^*)$, where Θ is the Heaviside step function and t^* is the timepoint at which the bifurcation occurs in their model. Hence, cells differentiated with certainty after timepoint t^*. Similarly, the time it takes for the deterministic system to settle into one of the differentiated stable states would be accounted for by or model as a sharply peaked gene expression delay. Hence, our model would allow to reconstruct the timing of the underlying bifurcation event, which in this scenario correspond to the timepoint of differentiation.

Having established that our model is capable of accurately predicting timepoints of differentiation when the driving mechanism is a toggle switch, we finally analyzed the role of PU.1 in the GM/MegE cell fate decision in hematopoiesis. Here we used a recent dataset by Hoppe et al., in revision, where for the first time expression levels of PU.1 and Gata1 were observed in single blood stem and progenitor cells continuously over time across several generations. We focused on GM-fated cells, whose commitment is indicated by expression of CD16/32, a GM-lineage specific surface marker. For this fate choice, it is expected that PU.1 is upregulated from intermediate expression in CMPs to high levels in GMPs. However, whether this upregulating is the origin or a consequence of the cell

fate choice is yet unknown. We fitted the proposed model to the CD16/32 onsets in GM-fated genealogies and predicted the timepoints of differentiation within those genealogies. Investigating the PU.1 expression levels in the predicted cells, we found that PU.1 is upregulated several generations later, whereas from a toggle switch driven cell fate decision one would expect PU.1 upregulating within the predicted cells. This led to the conclusion that PU.1 is not actively involved in the cell fate decision, but is merely regulated in response to the the chosen fate where it then coordinates the fate-specific gene expression in its role as a myeloid master regulator.

While the long delay (> 4 generations) between the cell fate decision and the observation of the marker onset was unexpected, it fits well to another independent observation: Genealogies where both myeloid- and erythroid-fated cells are present are very rare ($< 10\%$, Hoppe et al., in revision). This can be explained if the respective cell fate decisions happen early on in the genealogy. Otherwise, one would in fact expect the majority of genealogies to contain both cell fates just by chance (see Marr et al., 2012 for mathematical details). Another possible explanation for the lack of these mixed genealogies is the presence of a strong lineage bias in the starting population, i.e. single cells are undifferentiated, but latently already biased towards either the myeloid or erythroid lineage. Such lineage bias was readily observed in a progenitor cell line (Chang et al., 2008) and also reported for hematopoietic stem cells, albeit for different lineages (Dykstra et al., 2007; Müller-Sieburg et al., 2002) and might be implemented molecularly by the toggle switch discussed in chapter 3. However, such lineage bias would not produce the observed correlated marker onsets. Cells within a genealogy would only choose the same fate, but the timing of this decision would not be correlated. Hence, we argue that the lack of mixed genealogies originates from an early cell fate decision rather than from a lineage bias.

To verify the predictions of our model, additional experimental proof is required. To that end, we propose to measure the rate of the point process governing differentiation via cytokine instruction: Assuming that cytokines can instruct undecided cells but exert no effect on already decided cells, one can determine the timepoints of differentiation via colony assays[1]. Applying a myeloid-promoting cytokine at the start of the experiment should result exclusively in GM-colonies, as all cell were undecided and can be instructed via the cytokine. Applying the cytokine at a later timepoint will result in some MegE-colonies, as some cells already differentiated into the MegE-lineage and cannot be instructed by the cytokine. Hence, one can derive the fraction of differentiated cells and therefore the rate of differentiation at each time from the number of non-instructable colonies.

While it was already postulated that the PU.1/Gata1 toggle switch as a whole cannot be the determinant of myeloid/erythroid fate (Hoppe et al., in revision; Schwarzfischer, 2013), e.g. due to absence of Gata1 in CMPs, here our model showed more generally that also a toggle switch consisting of PU.1 and another yet unknown transcription factor does not implement the cell fate decision molecularly. Note that we did not analyze the timecourses of Gata1 during the cell fate decision as Gata1 is not detectable in GM-fated cells, as shown by Schwarzfischer (2013). Our findings do not in general rule out the paradigm of toggle switches as molecular implementations of binary cell fate choice, but simply show that the most promising candidate pair of PU.1 and Gata1 is not an instance

[1] Cells are plated in clonal density and after a given period of time, e.g. one week, the composition of the formed colonies in terms of cell types is assessed.

of this paradigm. For the future, an important question arises: How can one find the true transcription factors that might mediate the myeloid/erythroid cell fate choice?

From the experimental side, one could rely on increasing automation of time-lapse microscopy, cell tracking and quantification, e.g. using microfluidics, and perform the analysis on a whole library of mouse lines, where different transcription factors have been tagged by fluorescence (analogous to the Yeast GFP library, Huh et al., 2003) to find transcription factors that are differentially regulated during the cell fate decision.

On the computational side, we envision a combination of time-lapse experiments, yielding timecourses of a few proteins, and static large scale methods such as single-cell qPCR or RNAseq. Based on the idea that the gene regulatory network should restrict the transcription factor dynamics to a low dimensional manifold (Huang, 2012), one might be able to infer from timecourses of a few factors the current location of a cell on manifold and hence extrapolate timecourses of the other factors that have only been observed statically. This would allow to assess the dynamics of the entire gene regulatory network at the timepoint of a cell fate decision and to find those factors that drive it. In a first attempt, it would be interesting to compare our results about the timing of the myeloid/erythroid decision from time-lapse data with static single cell qPCR data. Ideally, the qPCR dataset should contain the same cell types observed also in the course of the time-lapse experiment, i.e. hematopoietic stem cells, CMPs, GMPs and MEPs and has to contain PU.1 and Gata1 in its gene-set. Performing dimension reduction using e.g. diffusion maps (Coifman et al., 2005) creates a low dimensional embedding of the high dimensional qPCR data and can also recapitulate the branching structure of the underlying cell type hierarchies (Haghverdi et al., in revision) and the time ordering of cells. Mapping the time ordering of the diffusion map into real time by e.g. comparing the rates of change in PU.1 and Gata1 between datasets, one could assess where our predicted delay of 4-5 generations between cell fate decision and CD16/32 lies with respect to the branching of the two lineages in the qPCR dataset and which factors are differentially regulated at that point.

Although during my thesis, the methods and models were developed in the context of stem cell differentiation and hematopoiesis in particular, the ubiquity of stochastic cell state transitions opens up many other applications: Reprogramming somatic cells into iPS cells is believed to be a stochastic process (Hanna et al., 2009; Buganim et al., 2012) and e.g. analyzing the timing of reprogramming (Morris et al., 2014) might give important insight into this complex procedure. Similarly, it is thought that tumorogenesis and tumor heterogeneity is the result of stochastic state transitions between cancer stem cells and non-tumorogenic cells (Gupta et al., 2011). Furthermore, metastases are generated when cells randomly undergo a epithelial-mesenchymal transition, detach from the tumor and spread the cancer into other body parts (Magee et al., 2012). Here, our methods could be used e.g. to understand how external environment influences transition probabilities within the tumor or to predict at which cells in the tumor initiated an epithelial-mesenchymal transition, thus promoting the development of more effective therapies.

Appendix A

The master equations of an interacting population of dividing cells

We consider a population of cells, each defined via its position x, its state s and its age τ (see section 4.1.1). Cells change state s with rate $\lambda(F)$, where F are the cell's features (cell density, age, etc.), and divide with age dependent rate $\gamma(\tau)$. First, let us describe how the distribution of a single cell evolves:

$$\begin{aligned}\dot{\mathcal{P}}(x,s,\tau,t) = \nabla^2\mathcal{P}(x,s,\tau,t) + \frac{\partial}{\partial\tau}\mathcal{P}(x,s,\tau,t)\\ - \delta_{s,0}\cdot\lambda(x,t,\tau)\cdot\mathcal{P}(x,s,\tau,t)\\ + \delta_{s,1}\cdot\lambda(x,t,\tau)\cdot\mathcal{P}(x,s-1,\tau,t)\\ - \gamma(\tau)\cdot\mathcal{P}(x,s,\tau,t)\end{aligned}$$

This is analogous to the equation given in section 4.1.1, but contains an extra term that accounts for the loss of the single cell due to cell division giving rise to a pair of cells. When considering pairs of cells, we must describe the evolution of their joint distribution $\mathcal{P}(x_1,s_1,\tau_1,x_2,s_2,\tau_2,t)$:

$$\begin{aligned}\dot{\mathcal{P}}(x_1,&s_1,\tau_1,x_2,s_2,\tau_2,t) = \left(\nabla^2_{x_1,x_2} + \frac{\partial}{\partial\tau_2} + \frac{\partial}{\partial\tau_2}\right)\cdot\mathcal{P}(x_1,s_1,\tau_1,x_2,s_2,\tau_2,t)\\ &- [\delta(s_1)\cdot\lambda(F_1(x_1,\tau_1,x_2,\tau_2)) + \delta(s_0)\cdot\lambda(F_2(x_1,\tau_1,x_2,\tau_2))]\cdot\mathcal{P}(x_1,s_1,\tau_1,x_2,s_2,\tau_2,t)\\ &+ \delta(s_1-1)\cdot\lambda(F_1(x_1,\tau_1,x_2,\tau_2))\cdot\mathcal{P}(x_1,s_1-1,\tau_1,x_2,s_2,\tau_2,t)\\ &+ \delta(s_2-1)\cdot\lambda(F_2(x_1,\tau_1,x_2,\tau_2))\cdot\mathcal{P}(x_1,s_1,\tau_1,x_2,s_2-1,\tau_2,t)\\ &+ \delta(\tau_1)\cdot\delta(\tau_2)\cdot\delta(x_1-x_2)\cdot\int_0^t d\tau'\gamma(\tau')\mathcal{P}(x_1,s_1,\tau',t)\\ &- [\gamma(\tau_1)+\gamma(\tau_2)]\cdot\mathcal{P}(x_1,s_1,\tau_1,x_2,s_2,\tau_2,t)\end{aligned}$$

The first line represent diffusion in space and drift in time. The second line corresponds to loss due to a state change out of state A, where the rate λ depends on the features F_i of

cell i, which is function depending on potentially all system variables (e.g. cell locations when representing cell density). The third line accounts for a cell in state A ($s_1 = 0$) at position x_1 that transitions into state B (the forth line is analogous for the second cell). The fifth line includes the gain of probability due to a division event in the single cell equation creating two cells of age 0. This term gives rise to the coupling of the equations. For simplicity we assumed that a division event at location x gives rise to two cells both located at x as well. Finally, the last term models loss due to either cell 1 or cell 2 dividing.

The equation for higher numbers of cells (triplets, etc.) become more and more complex as one has to deal with arising symmetries (see e.g. Dodd and Ferguson (2009)).

Appendix B

Inferring lineage decisions from branches

B.1 Theory

Here, we briefly develop an alternative approach to infer the differentiation events, based only on branches of the lineage trees. One can simply reduce the intricate tree-structured data to single branches and estimate the parameters from these branches. Having observed a set of trees $T_i, (i = 1, \ldots, N)$, for each tree T_i one extracts all possible branches B_i^j, one branch per leave cell j (Fig. B.1A). In order to obtain independent data points required for parameter estimation, we have to randomly sample a single branch $B_i^{\tilde{j}}$ from each tree T_i (see Fig. B.1A). For example, if we observe 100 trees we can only obtain 100 independent branches, thereby discarding a lot of data. Having sampled N independent branches $B_i, (i = 1, \ldots, N)$ we extract the times $t_i, (i = 1, \ldots, N)$ between the root of the branch and the onset of the marker. We can now derive the likelihood of these observations t_i given the parameters (θ, η) of our model. Recall that θ and η represent the parameters associated with the differentiation process and delay process, respectively.

Both the differentiation process and the delay process contribute to the length of the observed branch t_i. The branch length t_i is the sum of two random variables, one representing the time until differentiation, the other representing the time from differentiation to marker onset. Hence, we observe the convolution (Fig. B.1B bottom panel) of the hidden event densities of the differentiation $\Phi(t)$ (Fig. B.1B upper panel) and the delay $\Psi(t)$ (Fig. B.1B middle panel). For simplicity of notation we defined $\Psi(t) = \Psi_0(t)$, the first passage time distribution starting from state $x_0 = 0$. The likelihood for a single marker onset observation at time t is therefore:

$$\mathcal{L}(t|\theta, \eta) = (\Phi * \Psi)(t) = \int_0^t d\tau \; \Phi(\tau|\theta) \cdot \Psi(t - \tau|\eta) \; . \tag{B.1}$$

Here, we explicitly note the dependence of Φ and Ψ on their associated parameters θ and η. The likelihood of multiple observations t_i $(i = 1, ..., N)$ is just the product of the individual likelihoods due to their independence. Note that this is only possible because we enforced independence by sampling just one branch per tree. In order to estimate the

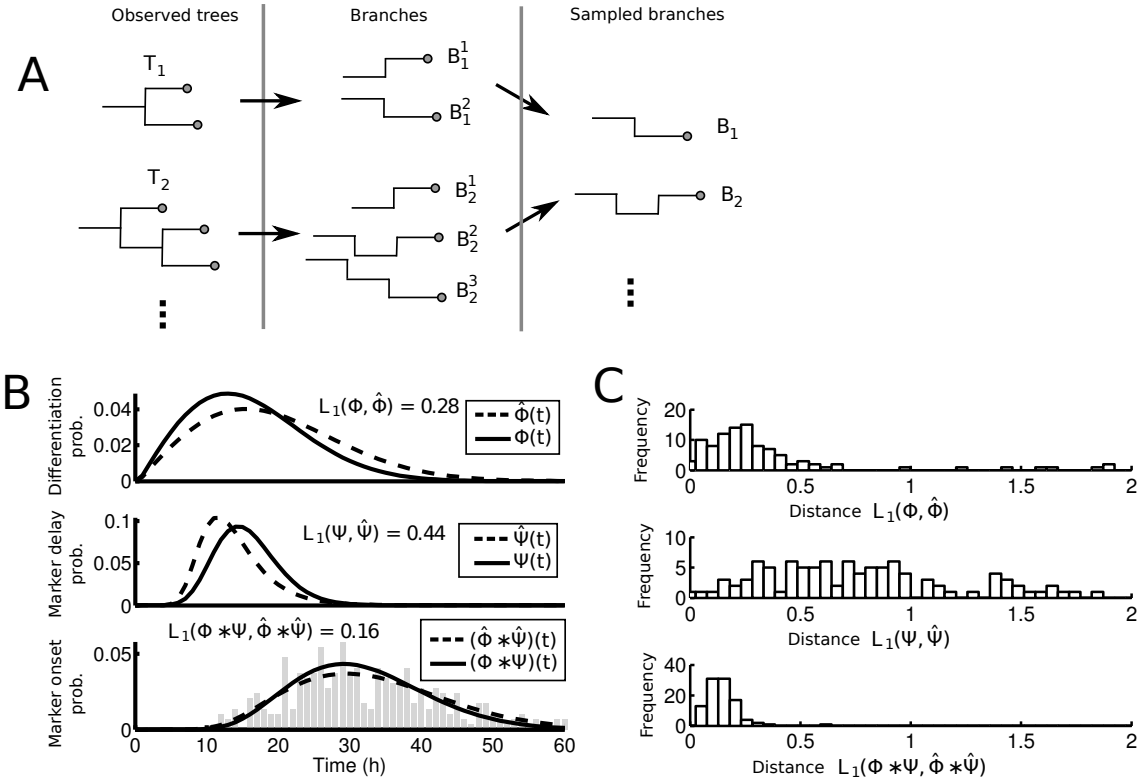


Figure B.1: Inference of the differentiation decision using branches estimates observed data well but fails to infer differentiation and delay distributions properly. A) Having observed a set of trees T_i with marker onsets indicated by gray circles, for each tree, one constructs all possible branches B_i^j and obtains independent samples B_i by randomly choosing one branch per tree. B) The marker onset distribution (bottom panel) is the convolution of the differentiation distribution Φ (upper panel) and the marker delay distribution Ψ (middle panel). Based on 50 observed branches generated from one parameter set, we obtain an estimate of the marker onset distribution (gray bars in bottom panel). Applying the deconvolution based on a linearly increasing differentiation hazard, we find estimates of differentiation and delay distributions, $\hat{\Phi}, \hat{\Psi}$ (dashed lines in all panel). Even though the marker onset distributions and its estimate agree very well (L_1 distance = 0.16), the estimates $\hat{\Phi}$ and $\hat{\Psi}$ are apparently different from the distributions Φ and Ψ (L_1 distance = 0.28 and L_1 distance = 0.44). C) Histograms of L_1-distance between estimated and true differentiation-, marker delay-, and marker onset distributions for 100 randomly chosen parameter sets. Even though the distance with respect to the convolution is small ($\text{Quantile}_{0.95} = 0.25$ in the bottom panel) indicating a good fit, the fits of the underlying unobserved components Φ and Ψ don't agree ($\text{Quantile}_{0.95} = 1.50$ and $\text{Quantile}_{0.95} = 1.61$ in the upper and middle panel).

parameters θ and η of the two underlying processes, we maximize the log-likelihood with respect to the parameters to obtain the maximum likelihood estimate:

$$(\hat{\theta}, \hat{\eta}) = \underset{\theta,\eta}{\operatorname{argmax}} \log \left[\prod_i \mathcal{L}(t_i|\theta, \eta) \right] = \underset{\theta,\eta}{\operatorname{argmax}} \sum_i \log \left[\mathcal{L}(t_i|\theta, \eta) \right] \tag{B.2}$$

To solve the above optimization problem, we apply a standard multiple-restart (Latin Hypercube (McKay et al., 1979)) optimization routine.

B.2 Simulation study

We now test the branch inference method on synthetic data. We first choose parameters θ and η, giving rise to the a particular differentiation and delay distribution (via Eq. 5.3 and Eq. 5.5) as well as their convolution (see Fig. B.1B, solid lines). Next, we draw 300 samples from this convolution as our observation (grey bars in Fig. B.1B bottom panel). By solving the optimization problem in Eq. B.2 numerically, we obtain a maximum likelihood estimate $(\hat{\theta}, \hat{\eta})$ and compute the corresponding differentiation $(\hat{\Phi})$, delay $(\hat{\Psi})$ and convolved distributions $(\hat{\Phi} * \hat{\Psi})$ (see Fig. B.1B, dashed lines). One observes that the true and estimated convolved distributions (Fig. B.1B, solid and dashed lines, bottom panel) are very similar. This is expected, because we are directly fitting this distribution via the optimization. Since we provided only a finite sample of size $N = 300$ from this distribution (Fig. B.1B, bars), the fitted distribution is slightly different from the true one. Examining the differentiation and delay distribution (Fig. B.1B, upper and middle panel), we notice a considerable discrepancy between the true and estimated distributions. Even though we are able to find parameters that closely reproduce the convolution, the underlying two distributions $\hat{\Phi}$ and $\hat{\Psi}$ are different from the true distributions Φ and Ψ.

As a simple measure to quantify this difference, we calculate the L_1-distance between true distribution (p) and estimated distributions $(\hat{p})$ as:

$$L_1(p, \hat{p}) = \sum_t |p(t) - \hat{p}(t)| \, .$$

In our toy example (Fig. B.1B), we find a small distance between the estimated and true convolution (0.1), but larger distance for the differentiation (0.28) and delay (0.44) distributions. This toy example suggests that the deconvolution is not unique within the error introduced by finite sample size. This is not surprising, as noisy deconvolution problems often are ill-posed if not sufficiently constrained. In our example, we constrain the distributions Φ and Ψ to be of a specific form, but these constraints are still not sufficient to make the problem solvable. For example, a trivial indeterminacy occurs if both distributions can closely resemble each other (within the error introduced by finite sample size) for certain choices of parameters. Since the convolution is symmetric $(f * g = g * f)$, in this case one cannot tell what is differentiation and what is delay just from observing the sum of both processes, e.g swapping Φ and Ψ in Fig. B.1B will yield exactly the same convolution and hence the same likelihood.

In order to systematically assess this non-identifiability using branches, we repeat the above analysis for a set of 100 parameters $(\theta, \eta)_i, (i = 1, \ldots, 100)$, randomly sampled

from the intervals given in Table 5.2. We simulate 50 trees T_i for each $(\theta, \eta)_i$, sample a set of branches B_i (as discussed in Fig. B.1A) and perform the maximum likelihood estimation using Eq. B.2. For each set of branches B_i, we obtain estimates of the underlying differentiation and delay distributions ($\hat{\Phi}^i$ and $\hat{\Psi}^i$). In Fig. B.1C, a histogram of the L_1-distances between estimated and true distributions is shown. We observe that the distance with respect to the convolution distribution is rather small ($\text{Quantile}_{0.95} = 0.25$), indicating that the the estimated parameters can fit the true marker distribution very well. Again, one cannot expect this distance to be zero because we provided only a finite sample (as discussed for Fig. B.1B). The distances with respect to differentiation and delay distribution are more wide spread, especially the delay distributions often fit very badly ($\text{Quantile}_{0.95} = 1.61$). This confirms our previous statement about the ill-posed nature of this problem: Observing only branches is insufficient to reconstruct the two underlying processes of differentiation and marker delay.

Note that the prediction of differentiating cells within a tree is not possible using the branch inference. One can of course estimate Φ and Ψ by maximizing Eq. B.2 for a set of observed branches and use Eq. B.1 to find the cell within a branch that has most likely differentiated. However, there is no way to predict the differentiating cells within a whole tree with that approach: The branch inference method cannot per se handle the tree structure, because it treats cells independently. If we predict the branches within a tree independently, we can easily run into inconsistencies. For example consider the tree T_1 in Fig. B.1A. If for some reason in branch $B_1^{(1)}$ the leave is predicted to differentiate, but in branch $B_1^{(2)}$ the root is predicted to differentiate we cannot put those prediction back together into a consistent tree.

Appendix C

Tree inference for the toggle switch

C.1 A different parameter regime

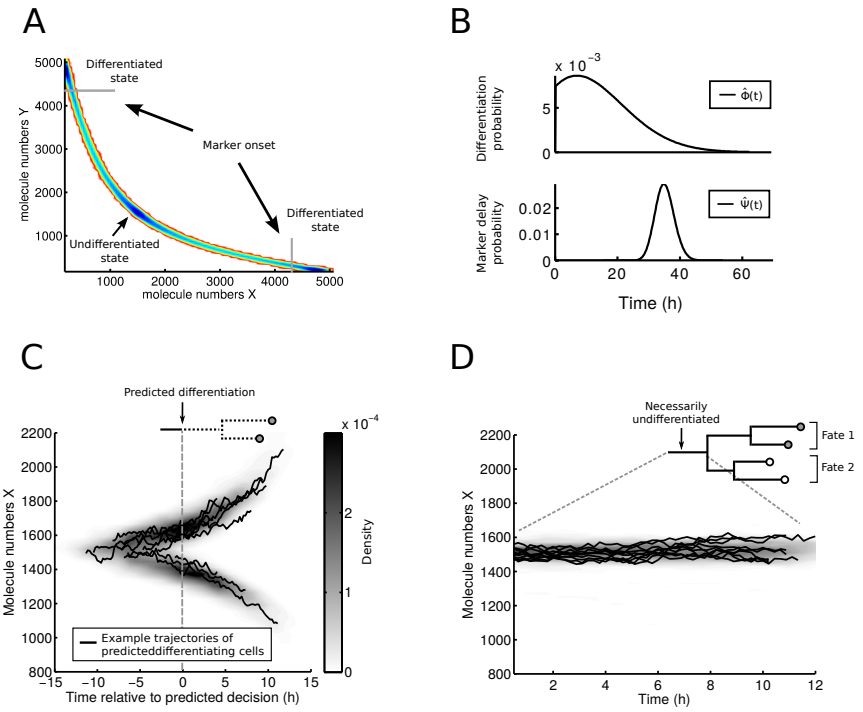


Figure C.1: **Inferring the differentiation timepoint in genealogies simulated with a toggle switch model (see Fig. 5.6) and a different set of parameters.** A) The model gives rise to a similar attractor landscape as before (compare to Fig. 5.6B). However, overall protein numbers are increased. B) Estimated differentiation and marker delay probability distributions from 100 simulated observed trees. C) Trajectories of predicted differentiating cells are centered on the timepoint of differentiation ($t = 0$). D) Trajectories of cells whose progeny differentiates into both fates (see inset). Parameters used for simulation are $\gamma_X = \gamma_Y = 0.29\ \mathrm{h}^{-1}$ (degradation rate), $\alpha_X = \alpha_Y = 1440\ \mathrm{h}^{-1}$ (maximal synthesis rate), $n = 2$ (cooperativity), and $K_X = K_Y = 1000$ (dissociation constants), $x^* = 4500$ (detection threshold).

C.2 A toggle switch coupled to a three-gene cascade

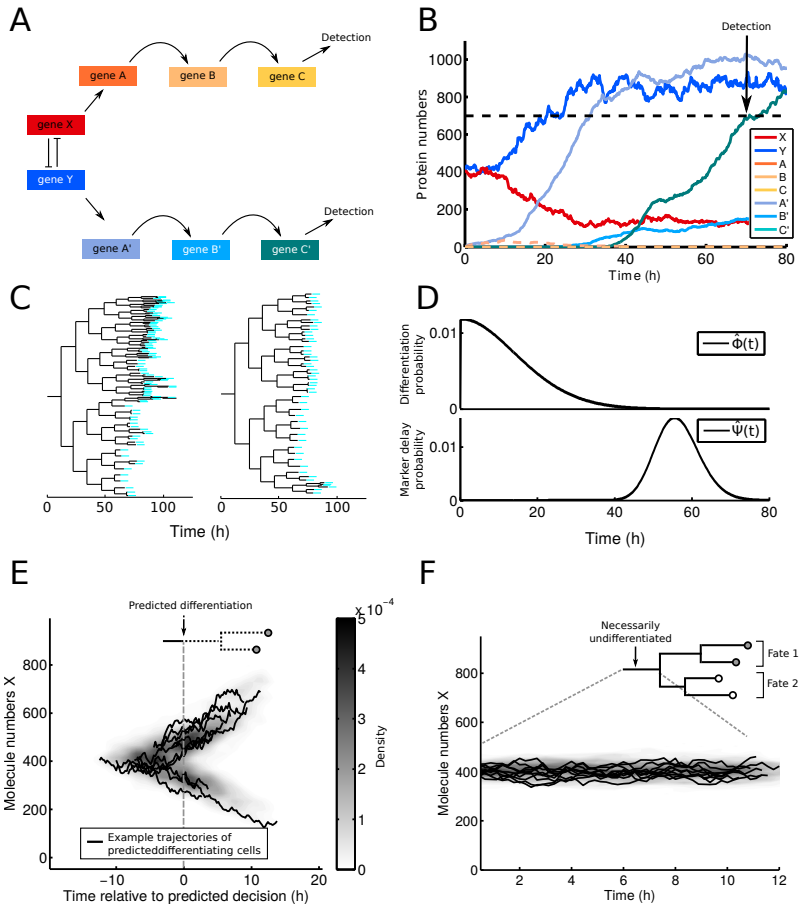


Figure C.2: A) Schematic of the model: A toggle switch (genes X and Y, see Fig. 5.6) implements the cell fate decision while the marker proteins that are detectable (genes C, C') are coupled to the toggle switch via a cascade of intermediate genes. For simplicity, the system is considered to be symmetric, i.e. both cascades share the same parameters. Parameters of the toggle switch (cascade) are the same as for Fig. 5.6 (Fig. 5.5). B) An exemplary trajectory of the system in A (for simplicity no cell division/branching is considered). The switch tilts at $t = 10$ hours in favor of protein Y, but the outcome of the decision is visible only at $t = 70$ when marker protein C' crosses the detection threshold (dashed black line). C) Two exemplary genealogies simulated with the model in A,B. Detection of the marker proteins C,C' is indicated in cyan. D) Estimated differentiation and marker delay probability distributions from 100 simulated observed trees. E) Trajectories of predicted differentiating cells are centered on the timepoint of differentiation ($t = 0$). F) Trajectories of cells whose progeny differentiates into both fates (see inset).

Bibliography

J. Abkowitz, S. Catlin, P. Guttorp, and Others. Evidence that hematopoiesis may be a stochastic process in vivo. *Nature medicine*, 2(2):190–197, 1996.

A. Aho and N. Sloane. Some doubly exponential sequences. *Fibonacci Quarterly*, 1973.

K. Akashi, X. He, J. Chen, H. Iwasaki, C. Niu, B. Steenhard, J. Zhang, J. Haug, and L. Li. Transcriptional accessibility for genes of multiple tissues and hematopoietic lineages is hierarchically controlled during early hematopoiesis. *Blood*, 101(2):383–390, 2003.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular biology of the cell*. Garland Science, 2002.

U. Alon. *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall/CRC Mathematical and Computational Biology. Taylor & Francis, 2006.

L. Alonso and E. Fuchs. Stem cells of the skin epithelium. *Proceedings of the National Academy of Sciences*, 100 Suppl: 11830–11835, 2003.

N. Amariglio, A. Hirshberg, B. W. Scheithauer, Y. Cohen, R. Loewenthal, L. Trakhtenbrot, N. Paz, M. Koren-Michowitz, D. Waldman, L. Leider-Trejo, A. Toren, S. Constantini, and G. Rechavi. Donor-derived brain tumor following neural stem cell transplantation in an ataxia telangiectasia patient. *PLoS Medicine*, 6(2):0221–0231, 2009.

F. Amat, W. Lemon, D. Mossing, and K. McDole. Fast, accurate reconstruction of cell lineages from large-scale fluorescence microscopy data. *Nature Methods*, (July), 2014.

M. Ashburner, C. a. Ball, J. a. Blake, D. Botstein, H. Butler, J. M. Cherry, a. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. a. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, 25:25–29, 2000.

A. Auger, P. Chatelain, and P. Koumoutsakos. R-leaping: accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of chemical physics*, 125(8):084103, August 2006.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine Learning*, 2008.

S. S. Bajikar, C. Fuchs, A. Roller, F. J. Theis, and K. A. Janes. Parameterizing cell-to-cell regulatory heterogeneities via stochastic transcriptional profiles. *Proceedings of the National Academy of Sciences*, 111:E626–35, 2014.

N. Barker, J. H. van Es, J. Kuipers, P. Kujala, M. van den Born, M. Cozijnsen, A. Haegebarth, J. Korving, H. Begthel, P. J. Peters, and H. Clevers. Identification of stem cells in small intestine and colon by marker gene Lgr5. *Nature*, 449 (October):1003–1007, 2007.

B. Barzel and O. Biham. Calculation of switching times in the genetic toggle switch and other bistable systems. *Physical Review E*, 78(4):41919, October 2008.

D. Basiji, W. Ortyn, and L. Liang. Cellular image analysis and imaging by flow cytometry. *Clinics in laboratory medicine*, 27:653–670, 2007.

M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, and C. P. Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, October 2009.

A. J. Becker, E. A. McCulloch, and J. E. Till. Cytological demonstration of the clonal nature of spleen colonies derived from transplanted mouse marrow cells. *Nature*, 197(4866):452–454, 1963.

G. Bel, B. Munsky, and I. Nemenman. The simplicity of completion time distributions for common complex biochemical processes. *Physical biology*, 7(1):016003, March 2010.

A. Beuter. *Nonlinear Dynamics in Physiology and Medicine*. Interdisciplinary Applied Mathematics. Springer, 2003.

S. Bhattacharya, Q. Zhang, and M. E. Andersen. A deterministic map of Waddington's epigenetic landscape for cell fate specification. *BMC systems biology*, 5(1):85, 2011.

W. Bialek. Stability and noise in biochemical switches. In *Advances in neural information processing systems 13: proceedings of the 2000 conference*, volume 13, page 103. The MIT Press, 2001.

D. A. Birch and W. R. Young. A master equation for a spatial population model with pair interactions. *Theoretical Population Biology*, 70:26–42, 2006.

C. M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag New York, Inc., 2006.

K. Blagovic, L. Y. Kim, and J. Voldman. Microfluidic perfusion for regulating diffusible signaling in stem cells. *PLoS ONE*, 6(8), 2011.

A.-M. Boehm, K. Khalturin, F. Anton-Erxleben, G. Hemmrich, U. C. Klostermeier, J. A. Lopez-Quintero, H.-H. Oberg, M. Puchert, P. Rosenstiel, J. Wittlieb, and T. C. G. Bosch. FoxO is a critical regulator of stem cell maintenance in immortal Hydra. *Proceedings of the National Academy of Sciences*, 109:19697–702, 2012.

P. Bokes, J. R. King, and M. Loose. A bistable genetic switch which does not require high co-operativity at the promoter: a two-timescale model for the PU.1-GATA-1 interaction. *Mathematical medicine and biology : a journal of the IMA*, 26(2):117–32, June 2009.

P. Bokes, J. R. King, A. T. Wood, and M. Loose. Exact and approximate distributions of protein and mRNA levels in the low-copy regime of gene expression. *Journal of mathematical biology*, June 2011.

P. Brennecke, S. Anders, J. K. Kim, A. A. Kołodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler. Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods*, 10(11):1093–5, 2013.

F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle. Computational dissection of cell-to-cell heterogeneity in single-cell RNA-Seq data reveals novel structure between cells. *Nature biotechnology*, in press, 2014.

Y. Buganim, D. A. Faddah, A. W. Cheng, E. Itskovich, S. Markoulaki, K. Ganz, S. L. Klemm, A. van Oudenaarden, and R. Jaenisch. Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell*, 150(6):1209–1222, September 2012.

F. Buggenthin, C. Marr, M. Schwarzfischer, P. S. Hoppe, O. Hilsenbeck, T. Schroeder, and F. J. Theis. An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy. *BMC Bioinformatics*, 14(1): 297, 2013.

L. Cai, C. K. Dalal, and M. B. Elowitz. Frequency-modulated nuclear localization bursts coordinate gene regulation. *Nature*, 455(September):485–490, 2008.

X. Cai and Z. Xu. K-leap method for accelerating stochastic simulation of coupled chemical reactions. *The Journal of chemical physics*, 126(7):074102, February 2007.

Y. Cao, H. Li, and L. R. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The Journal of chemical physics*, 121(9):4059–67, September 2004.

Y. Cao, D. T. Gillespie, and L. R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of chemical physics*, 122(1):14116, January 2005.

A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome biology*, 7(10):R100, 2006.

I. Chambers, J. Silva, D. Colby, J. Nichols, B. Nijmeijer, M. Robertson, J. Vrana, K. Jones, L. Grotewold, and A. Smith. Nanog safeguards pluripotency and mediates germline development. *Nature*, 450(December):1230–1234, 2007.

H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–7, May 2008.

P. K. Chattopadhyay, T. M. Gierahn, M. Roederer, and J. C. Love. Single-cell technologies for monitoring immune systems. *Nature immunology*, 15(2):128–35, 2014.

N. Chenouard, I. Smal, F. de Chaumont, M. Maška, I. F. Sbalzarini, and E. Meijering. Objective comparison of particle tracking methods. *Nature methods*, 11(3):281–9, March 2014.

J. L. Cherry and F. R. Adler. How to make a biological switch. *Journal of theoretical biology*, 203(2):117–33, March 2000.

V. Chickarmane and C. Peterson. A computational model for understanding stem cell, trophectoderm and endoderm lineage determination. *PLoS ONE*, 3(10):1–8, 2008.

V. Chickarmane, C. Troein, U. A. Nuber, H. M. Sauro, and C. Peterson. Transcriptional dynamics of the embryonic stem cell switch. *PLoS Computational Biology*, 2(9):1080–1092, 2006.

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

V. Chickarmane, V. Olariu, and C. Peterson. Probing the role of stochasticity in a model of the embryonic stem cell – heterogeneous gene expression and reprogramming efficiency. *BMC Systems Biology*, 6:98, 2012.

M. F. Ciaccio, J. P. Wagner, C.-P. Chuu, D. A. Lauffenburger, and R. B. Jones. Systems analysis of EGF receptor signaling dynamics with microwestern arrays. *Nature methods*, 7(2):148–155, 2010.

C. Clopper and E. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26: 404–413, 1934.

A. R. Cohen, F. L. a. F. Gomes, B. Roysam, and M. Cayouette. Computational prediction of neural progenitor cell fates. *Nature methods*, 7(3):213–8, March 2010.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–31, May 2005.

P. Coolen-Schrijner and E. A. Van Doorn. On the convergence to stationarity of birth-death processes. *Journal of applied probability*, 38(3):696–706, 2001.

M. R. Costa, F. Ortega, M. S. Brill, R. Beckervordersandforth, C. Petrone, T. Schroeder, M. Götz, and B. Berninger. Continuous live imaging of adult neural stem cell division and lineage progression in vitro. *Development*, 138(6): 1057–68, March 2011.

D. L. Coutu and T. Schroeder. Probing cellular processes by long-term live imaging–historic problems and current solutions. *Journal of cell science*, 126:3805–15, 2013.

D. Cox. Regression models and life tables. *JR stat soc B*, 34(2):187–220, 1972.

R. Dahl. Development of Macrophages and Granulocytes. In *Molecular Basis of Hematopoiesis*, pages 127–149. Springer New York, 2009.

R. Dahl, J. C. Walsh, D. Lancki, P. Laslo, S. R. Iyer, H. Singh, and M. C. Simon. Regulation of macrophage and neutrophil cell fates by the PU.1:C/EBPalpha ratio and granulocyte colony-stimulating factor. *Nature immunology*, 4(10):1029–1036, 2003.

M. E. Dahlberg and S. J. Benkovic. Kinetic mechanism of DNA polymerase I (Klenow fragment): identification of a second conformational change and evaluation of the internal equilibrium constant. *Biochemistry*, 30(20):4835–4843, May 1991.

A. Dakic, D. Metcalf, L. Di Rago, S. Mifsud, L. Wu, and S. L. Nutt. PU.1 regulates the commitment of adult hematopoietic progenitors and restricts granulopoiesis. *The Journal of experimental medicine*, 201(9):1487–1502, 2005.

P. Dalerba, T. Kalisky, D. Sahoo, P. S. Rajendran, M. E. Rothenberg, A. A. Leyrat, S. Sim, J. Okamoto, D. M. Johnston, D. Qian, M. Zabala, J. Bueno, N. F. Neff, J. Wang, A. A. Shelton, B. Visser, S. Hisamori, Y. Shimono, M. van de Wetering, H. Clevers, M. F. Clarke, and S. R. Quake. Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12):1120–1127, 2011.

G. Q. Daley. Cellular Alchemy and the Golden Age of Reprogramming. *Cell*, 151(6):1151–1154, December 2012a.

G. Q. Daley. The promise and perils of stem cell therapeutics. *Cell Stem Cell*, 10(6):740–749, 2012b.

E. Davidson and D. Erwin. Gene regulatory networks and the evolution of animal body plans. *Science*, 311(2006):796–800, 2006.

L. A. Diaz Jr, R. T. Williams, J. Wu, I. Kinde, J. R. Hecht, J. Berlin, B. Allen, I. Bozic, J. G. Reiter, M. A. Nowak, K. W. Kinzler, K. S. Oliner, and B. Vogelstein. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature*, 486(7404):537–540, 2012.

J.-E. Dietrich and T. Hiiragi. Stochastic patterning in the mouse pre-implantation embryo. *Development*, 134:4219–4231, 2007.

P. J. Dodd and N. M. Ferguson. A many-body field theory approach to stochastic models in population biology. *PloS one*, 4(9):e6855, January 2009.

C. Duff, K. Smith-Miles, L. Lopes, and T. Tian. Mathematical modelling of stem cell differentiation: the PU.1-GATA-1 interaction. *Journal of mathematical biology*, April 2011.

M. J. Dunlop, R. S. Cox, J. H. Levine, R. M. Murray, and M. B. Elowitz. Regulatory activity revealed by dynamic correlations in gene expression noise. *Nature genetics*, 40(12):1493–8, December 2008.

B. Dykstra, D. Kent, M. Bowie, L. McCaffrey, M. Hamilton, K. Lyons, S. J. Lee, R. Brinkman, and C. Eaves. Long-Term Propagation of Distinct Hematopoietic Differentiation Programs In Vivo. *Cell Stem Cell*, 1:218–229, 2007.

H. M. Eilken, S.-I. Nishikawa, and T. Schroeder. Continuous single-cell imaging of blood generation from haemogenic endothelium. *Nature*, 457(7231):896–900, February 2009.

A. Eldar and M. B. Elowitz. Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–73, September 2010.

J. Elf and M. Ehrenberg. Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Res*, 13(11):2475–2484, November 2003.

V. Elgart, T. Jia, A. T. Fenley, and R. Kulkarni. Connecting protein and mRNA burst distributions for stochastic models of gene expression. *Physical Biology*, 8(4):046001, August 2011.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science*, 297(5584): 1183–6, August 2002.

S. Engblom. Computing the moments of high dimensional solutions of the master equation. *Applied Mathematics and Computation*, 180(2):498–515, September 2006.

M. Evans and M. Kaufman. Establishment in culture of pluripotential cells from mouse embryos. *Nature*, 292:154–156, 1981.

N. Faust, F. Varas, L. M. Kelly, S. Heck, and T. Graf. Insertion of enhanced green fluorescent protein into the lysozyme gene creates mice with green fluorescent granulocytes and macrophages. *Blood*, 96(2):719–26, July 2000.

A. Fisher and M. Merkenschlager. Fresh powder on Waddington's slopes. *EMBO reports*, 2010.

D. V. Foster, J. G. Foster, S. Huang, and S. A. Kauffman. A model of sequential branching in hierarchical cell fate determination. *Journal of theoretical biology*, 260(4):589–97, October 2009.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010.

N. Friedman, L. Cai, and X. Xie. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16):1–4, October 2006.

G. Fritz, N. E. Buchler, T. Hwa, and U. Gerland. Designing sequential transcription logic: a simple genetic circuit for conditional memory. *Systems and synthetic biology*, 1(2):89–98, April 2007.

G. Fritz, J. A. Megerle, S. a. Westermayer, D. Brick, R. Heermann, K. Jung, J. O. Rädler, and U. Gerland. Single cell kinetics of phenotypic switching in the arabinose utilization system of E. coli. *PLoS ONE*, 9(2), 2014.

C. Fuchs. *Inference for Diffusion Processes: With Applications in Life Sciences*. Springer, 2013.

C. Furusawa and K. Kaneko. A Dynamical-Systems View of Stem Cell Biology. *Proceedings of the National Academy of Sciences*, (October):215–217, 2011.

C. Gardiner. *Handbook of stochastic methods for physics, chemistry and the natural sciences*. Springer, 3rd edition, 2004.

T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in Escherichia coli. *Nature*, 403 (6767):339–42, January 2000.

D. Geiger, T. Verma, and J. Pearl. Identifying independence in Bayesian networks. *Networks*, 1990.

U. Gerland, J. D. Moroz, and T. Hwa. Physical constraints and functional characteristics of transcription factor-DNA interaction. *Proceedings of the National Academy of Sciences*, 99(19):12015–20, September 2002.

M. A. Gibson and J. Bruck. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, March 2000.

D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976.

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

D. T. Gillespie. The chemical Langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716, 2001.

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

I. Glauche, M. Herberg, and I. Roeder. Nanog variability and pluripotency regulation of embryonic stem cells–insights from a mathematical model analysis. *PloS one*, 5(6):e11238, 2010.

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(September):781–788, 2005.

F. L. a. F. Gomes, G. Zhang, F. Carbonell, J. a. Correa, W. a. Harris, B. D. Simons, and M. Cayouette. Reconstruction of rat retinal progenitor cell lineages in vitro reveals a surprising degree of stochasticity in cell fate decisions. *Development*, 138(2):227–35, January 2011.

T. Graf and T. Enver. Forcing cells to change lineages. *Nature*, 462(7273):587–594, December 2009.

T. Graf and M. Stadtfeld. Heterogeneity of embryonic and adult stem cells. *Cell stem cell*, 3(5):480–3, November 2008.

R. Grima. Investigating the robustness of the classical enzyme kinetic equations in small intracellular compartments. *BMC systems biology*, 3:101, 2009a.

R. Grima. Noise-induced breakdown of the Michaelis-Menten equation in steady-state conditions. *Physical Review Letters*, 102(May):1–4, 2009b.

R. Grima. An effective rate equation approach to reaction kinetics in small volumes: theory and application to biochemical reactions in nonequilibrium steady-state conditions. *The Journal of chemical physics*, 133(3):035101, July 2010a.

R. Grima. An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *Journal of Chemical Physics*, 133:0–15, 2010b.

R. Grima, P. Thomas, and A. V. Straube. How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations? *The Journal of chemical physics*, 135(8):084103, August 2011.

R. Grima, D. R. Schmidt, and T. J. Newman. Steady-state fluctuations of a genetic feedback loop: An exact solution. *Journal of Chemical Physics*, 137(9):0–13, 2012.

M. Grskovic, A. Javaherian, B. Strulovici, and G. Q. Daley. Induced pluripotent stem cells–opportunities for disease modelling and drug discovery. *Nature Reviews. Drug discovery*, 10(December):915–29, 2011.

C. M. Guldberg and P. Waage. Concerning chemical affinity. *Erdmann's Journal f{ü}r practische Chemie*, 127:69–114, 1879.

M. Guns, V. Vanacker, and T. Glade. Logistic regression applied to natural hazards: rare event logistic regression with replications. *Nat. Hazards Earth Syst. Sci*, 12:1937–1947, 2012.

G. Guo, M. Huss, G. Q. Tong, C. Wang, L. Li Sun, N. D. Clarke, and P. Robson. Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell*, 18(4):675–685, 2010.

P. B. Gupta, C. M. Fillmore, G. Jiang, S. D. Shapira, K. Tao, C. Kuperwasser, and E. S. Lander. Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, 146(4):633–644, 2011.

J. H. Hanna, K. Saha, B. Pando, J. van Zon, C. J. Lengner, M. P. Creyghton, A. van Oudenaarden, and R. Jaenisch. Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, 462(7273):595–601, December 2009.

C. V. Harper, B. Finkenstaedt, D. J. Woodcock, S. Friedrichsen, S. Semprini, L. Ashall, D. G. Spiller, J. J. Mullins, D. Rand, J. R. E. Davis, and M. R. H. White. Dynamic analysis of stochastic transcription cycles. *PLoS biology*, 9(4): e1000607, April 2011.

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

J. Hasenauer, V. Wolf, A. Kazeroonian, and F. J. Theis. Method of conditional moments (MCM) for the Chemical Master Equation : A unified framework for the method of moments and hybrid stochastic-deterministic models. *Journal of mathematical biology*, August 2013.

T. Hashimshony, F. Wagner, N. Sher, and I. Yanai. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, 2(3):666–673, 2012.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer New York Inc., 2009.

H. He and E. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9): 1263–1284, September 2009.

M. Heinäniemi and M. Nykter. Gene-pair expression signatures reveal lineage control. *Nature Methods*, 10(6):7–9, 2013.

M. Held, M. H. A. Schmitz, B. Fischer, T. Walter, B. Neumann, M. H. Olma, M. Peter, J. Ellenberg, and D. W. Gerlich. CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nature methods*, 7(9):747–54, September 2010.

M. Herberg, T. Kalkan, I. Glauche, A. Smith, and I. Roeder. A model-based analysis of culture-dependent phenotypes of mESCs. *PloS one*, 9(3):e92496, 2014.

C. Heyworth, S. Pearson, G. May, and T. Enver. Transcription factor-mediated lineage switching reveals plasticity in primary committed progenitor cells. *The EMBO journal*, 21(14):3770–81, July 2002.

P. Hillenbrand, G. Fritz, and U. Gerland. Biological signal processing with a genetic toggle switch. *PloS one*, 8(7):e68345, January 2013.

J. E. M. Hornos, D. Schultz, G. Innocentini, J. Wang, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes. Self-regulating gene: An exact solution. *Physical Review E*, 72(5):1–5, November 2005.

S. Huang. The molecular and mathematical basis of Waddington's epigenetic landscape: a framework for post-Darwinian biology? *BioEssays*, 34(2):149–57, February 2012.

S. Huang. Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. *Cancer metastasis reviews*, 32(3-4):423–48, December 2013.

S. Huang, G. Eichler, Y. Bar-Yam, and D. E. Ingber. Cell Fates as High-Dimensional Attractor States of a Complex Gene Regulatory Network. *Physical Review Letters*, 94(12):128701, April 2005.

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

D. Huh and J. Paulsson. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, August 2011.

W.-K. Huh, J. Falvo, W.-K. Huh, L. Gerke, A. S. Carroll, L. C. Gerke, R. Howson, J. S. Weissman, R. W. Howson, and E. K. O'Shea. Global analysis of protein localization in budding yeast. *Nature*, 425:686–91, 2003.

S. Islam, U. Kjällquist, A. Moliner, P. Zajac, J.-B. Fan, P. Lönnerberg, and S. Linnarsson. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome research*, 21:1160–1167, 2011.

S. Islam, A. Zeisel, S. Joost, G. La Manno, P. Zajac, M. Kasper, P. Lönnerberg, and S. Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature methods*, 11(1):163–6, 2014.

H. Iwasaki, C. Somoza, H. Shigematsu, E. A. Duprez, J. Iwasaki-Arai, S.-I. Mizuno, Y. Arinobu, K. Geary, P. Zhang, T. Dayaram, M. L. Fenyus, S. Elf, S. Chan, P. Kastner, C. S. Huettner, R. M. Murray, D. G. Tenen, and K. Akashi. Distinctive and indispensable roles of PU. 1 in maintenance of hematopoietic stem cells and their differentiation. *Blood*, 106(5):1590–1600, 2005.

T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of mathematical biology*, 54(1):1–26, January 2007.

H. Jeffreys. *The Theory of Probability*. Oxford University Press, 1939.

R. R. Jenq and M. R. M. van den Brink. Allogeneic haematopoietic stem cell transplantation: individualized stem cell and immune therapy of cancer. *Nature Reviews. Cancer*, 10(February):213–221, 2010.

M. Kaern, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews. Genetics*, 6(6):451–464, June 2005.

T. Kalmar, C. Lim, P. Hayward, S. Muñoz Descalzo, J. Nichols, J. Garcia-Ojalvo, and A. M. Arias. Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, 7(7):33–36, 2009.

A. Kashiwagi, I. Urabe, K. Kaneko, and T. Yomo. Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. *PloS one*, 1(1):e49, January 2006.

S. A. Kauffman. Metabolic stability and epigenesis in randomly constructed genetic nets. *Journal of theoretical biology*, pages 437–467, 1969.

B. B. Kaufmann, Q. Yang, J. T. Mettetal, and A. van Oudenaarden. Heritable stochastic switching revealed by single-cell genealogy. *PLoS biology*, 5(9):e239, September 2007.

K. Kaushansky, M. A. Lichtman, E. Beutler, T. Kipps, U. Seligsohn, and J. Prchal. *Williams Hematology*. Williams Hematology. McGraw-Hill Professional, 8th edition, 2010.

D. G. Kendall. An artificial realization of a simple" birth-and-death" process. *Journal of the Royal Statistical Society. Series B*, 12(1):116–119, 1950.

M. G. Kendall and A. Stuart. *The advanced theory of statistics*. Number Bd. 1 in The Advanced Theory of Statistics. Griffin, 1967.

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

T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal*, 81(6):3116–36, December 2001.

B. Knapp, I. Rebhan, A. Kumar, P. Matula, N. A. Kiani, M. Binder, H. Erfle, K. Rohr, R. Eils, R. Bartenschlager, and L. Kaderali. Normalizing for individual cell population context in the analysis of high-content cellular screens. *BMC bioinformatics*, 12(1):485, January 2011.

M. Komorowski, B. Finkenstaedt, C. V. Harper, and D. Rand. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, 10:343, 2009.

J. Krumsiek, C. Marr, T. Schroeder, and F. J. Theis. Hierarchical Differentiation of Myeloid Progenitors Is Encoded in the Transcription Factor Network. *PLoS ONE*, 6(8):e22649, August 2011.

H. Y. Kueh, A. Champhekhar, S. L. Nutt, M. B. Elowitz, and E. V. Rothenberg. Positive Feedback Between PU.1 and the Cell Cycle Controls Myeloid Differentiation. *Science*, 670, July 2013.

H. Kulessa, J. Frampton, and T. Graf. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes & development*, 9:1250–1262, 1995.

B. Lambolez, E. Audinat, P. Bochet, F. Crepel, and J. Rossier. AMPA receptor subunits expressed by single Purkinje cells. *Neuron*, 9:247–258, 1992.

P. Laslo, C. J. Spooner, A. Warmflash, D. W. Lancki, H.-J. Lee, R. Sciammas, B. N. Gantner, A. R. Dinner, and H. Singh. Multilineage transcriptional priming and determination of alternate hematopoietic cell fates. *Cell*, 126(4):755–66, August 2006.

P. Laslo, J. M. R. Pongubala, D. W. Lancki, and H. Singh. Gene regulatory networks directing myeloid and lymphoid cell fates within the immune system. *Seminars in Immunology*, 20(4):228–235, 2008.

C. H. Lee, K.-H. Kim, and P. Kim. A moment closure method for stochastic reaction networks. *The Journal of chemical physics*, 130(13):134107, April 2009.

E. Lee and O. Go. Survival analysis in public health research. *Annual review of public health*, 1997.

K. Lewis. Persister cells, dormancy and infectious disease. *Nature reviews. Microbiology*, 5(January):48–56, 2007.

J. Liepe, C. Barnes, E. Cule, K. Erguler, P. Kirk, T. Toni, and M. P. H. Stumpf. ABC-SysBio–approximate Bayesian computation in Python with GPU support. *Bioinformatics*, 26(14):1797–9, July 2010.

G. Lillacci and M. Khammash. The signal within the noise: efficient inference of stochastic gene regulation models using fluorescence histograms and stochastic simulations. *Bioinformatics*, 29(18):2311–9, September 2013.

A. Lipshtat, A. Loinger, N. Q. Balaban, and O. Biham. Genetic Toggle Switch without Cooperative Binding. *Physical Review Letters*, 96(18):1–4, May 2006.

J. Liu, C. Hansen, and S. R. Quake. Solving the "world-to-chip" interface problem with a microfluidic matrix. *Analytical Chemistry*, 75(18):4718–4723, 2003.

A. Loinger, A. Lipshtat, N. Q. Balaban, and O. Biham. Stochastic simulations of genetic switch systems. *Physical Review E*, 75(2 Pt 1):21904, February 2007.

M. T. Lorincz. Optimized Neuronal Differentiation of Murine Embryonic Stem Cells. In K. Turksen, editor, *Embryonic Stem Cell Protocols*, pages 55–69. Humana Press, 2006.

R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65–8, April 2008.

A. J. Lotka. *Elements of Physical Biology*. Williams and Wilkins, 1925.

S. E. Luria and M. Delbrück. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics*, 28(November): 491–511, 1943.

H. Ma, R. Morey, R. C. O'Neil, Y. He, B. Daughtry, M. D. Schultz, M. Hariharan, J. R. Nery, R. Castanon, K. Sabatini, R. D. Thiagarajan, M. Tachibana, E. Kang, R. Tippner-Hedges, R. Ahmed, N. M. Gutierrez, C. Van Dyken, A. Polat, A. Sugawara, M. Sparman, S. Gokhale, P. Amato, D. P. Wolf, J. R. Ecker, L. C. Laurent, and S. Mitalipov. Abnormalities in human pluripotent cells due to reprogramming mechanisms. *Nature*, 511(7508):177–83, 2014.

P. Macchiarini, P. Jungebluth, T. Go, M. A. Asnaghi, L. E. Rees, T. A. Cogan, A. Dodson, J. Martorell, S. Bellini, P. P. Parnigotto, S. C. Dickinson, A. P. Hollander, S. Mantero, M. T. Conconi, and M. A. Birchall. Clinical transplantation of a tissue-engineered airway. *The Lancet*, 372(9655):2023–2030, 2008.

J. a. Magee, E. Piskounova, and S. J. Morrison. Cancer Stem Cells: Impact, Heterogeneity, and Uncertainty. *Cancer Cell*, 21(3):283–296, 2012.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavare. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–8, December 2003.

C. Marr, M. Strasser, M. Schwarzfischer, T. Schroeder, and F. J. Theis. Multi-scale modeling of GMP differentiation based on single-cell genealogies. *The FEBS journal*, 279(18):3488–500, September 2012.

G. R. Martin. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proceedings of the National Academy of Sciences*, 78(12):7634–7638, 1981.

M. Maska, V. Ulman, D. Svoboda, P. Matula, P. Matula, C. Ederra, A. Urbiola, T. España, S. Venkatesan, D. M. W. Balak, P. Karas, T. Bolcková, M. Streitová, C. Carthel, S. Coraluppi, N. Harder, K. Rohr, K. E. G. Magnusson, J. Jaldén, H. M. Blau, O. Dzyubachyk, P. Krízek, G. M. Hagen, D. Pastor-Escuredo, D. Jimenez-Carretero, M. J. Ledesma-Carbayo, A. Muñoz Barrutia, E. Meijering, M. Kozubek, and C. Ortiz-de Solorzano. A benchmark for comparison of cell tracking algorithms. *Bioinformatics*, 30(11):1–8, 2014.

H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(February):814–819, 1997.

H. McAdams and A. Arkin. It's a noisy business: Genetic regulation at the nanomolecular scale. *Trends in Genetics*, 15 (1998):65–69, 1999.

J. M. McCollum, G. D. Peterson, C. D. Cox, M. L. Simpson, and N. F. Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Computational Biology and Chemistry*, 30(1):39–49, February 2006.

P. McCullagh, J. Nelder. *Generalized linear models*. Chapman and Hall/CRC, 1989.

M. McKay, R. Beckman, and W. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245, 1979.

J. A. Megerle, G. Fritz, U. Gerland, K. Jung, and J. O. Rädler. Timing and dynamics of single cell gene expression in the arabinose utilization system. *Biophysical journal*, 95(4):2103–15, August 2008.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(1953):1087, 1953.

L. Michaelis and M. L. Menten. Die Kinetik der Invertinwirkung. *Biochem z*, pages 333–369, 1913.

E. Mjolsness, D. Orendorff, P. Chatelain, and P. Koumoutsakos. An exact accelerated stochastic simulation algorithm. *The Journal of chemical physics*, 130(14):144110, May 2009.

V. Moignard, I. C. Macaulay, G. Swiers, F. Buettner, J. Schütte, F. J. Calero-Nieto, S. Kinston, A. Joshi, R. Hannah, F. J. Theis, S. E. Jacobsen, M. F. de Bruijn, and B. Göttgens. Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature cell biology*, 15(4):363–72, April 2013.

N. Molina, D. M. Suter, R. Cannavo, B. Zoller, I. Gotic, and F. Naef. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proceedings of the National Academy of Sciences*, 110(51):20563–8, December 2013.

K. A. Moore and I. R. Lemischka. Stem cells and their niches. *Science (New York, N.Y.)*, 311(2006):1880–1885, 2006.

P. Moral, A. Doucet, and A. Jasra. An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, 22(5):1009–1020, August 2011.

M. J. Morelli, S. Tanase-Nicola, R. J. Allen, and P. R. ten Wolde. Reaction coordinates for the flipping of genetic switches. *Biophysical journal*, 94(9):3413–23, May 2008.

R. Morris, I. Sancho-Martinez, T. O. Sharpee, and J. C. Izpisua Belmonte. Mathematical approaches to modeling development and reprogramming. *Proceedings of the National Academy of Sciences*, 111(14), March 2014.

S. A. Morris, R. T. Y. Teo, H. Li, P. Robson, D. M. Glover, and M. Zernicka-Goetz. Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proceedings of the National Academy of Sciences*, 107 (14):6364–6369, 2010.

S. J. Morrison and A. C. Spradling. Stem cells and niches: mechanisms that promote stem cell maintenance throughout life. *Cell*, 132(4):598–611, February 2008.

A. Mugler, A. M. Walczak, and C. H. Wiggins. Spectral solutions to stochastic models of gene expression with bursts and regulation. *Physical Review E*, 80(4):1–19, October 2009.

U. Müller-Herold. General mass-action kinetics. Positiveness of concentrations as structural property of Horn's equation. *Chemical Physics Letters*, (3), 1975.

C. E. Müller-Sieburg, R. H. Cho, M. Thoman, B. Adkins, and H. B. Sieburg. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. *Blood*, 100(4):1302–1309, 2002.

B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of chemical physics*, 124(4):044104, January 2006.

B. Munsky and G. Neuert. Using Gene Expression Noise to Understand Gene Regulation. *Science*, 183, 2012.

B. Munsky, B. Trinh, and M. Khammash. Listening to the noise : random fluctuations reveal gene network parameters. *Molecular Systems Biology*, (318), 2009.

K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, Cambridge, Massachusetts, 2012.

J. D. Murray. Mathematical Biology I: An Introduction, 2002.

J. Narula, A. M. Smith, B. Gottgens, and O. A. Igoshin. Modeling reveals bistability and low-pass filtering in the network module determining blood stem cell fate. *PLoS Computational Biology*, 6(5):1–16, 2010.

C. Nerlov and T. Graf. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes & development*, 12:2403–2412, 1998.

G. Neuert, B. Munsky, R. Z. Tan, L. Teytelman, M. Khammash, and A. van Oudenaarden. Systematic Identification of Signal-Activated Stochastic Gene Regulation. *Science*, 339(6119):584–587, January 2013.

J. R. S. Newman, S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. *Nature*, 441(7095):840–6, June 2006.

C. Nicoleau, P. Viegas, M. Peschanski, and A. L. Perrier. Human pluripotent stem cell therapy for Huntington's disease: technical, immunological, and safety challenges. *Neurotherapeutics : the journal of the American Society for Experimental NeuroTherapeutics*, 8:562–76, 2011.

H. Niwa, Y. Toyooka, D. Shimosato, D. Strumpf, K. Takahashi, R. Yagi, and J. Rossant. Interaction between Oct3/4 and Cdx2 determines trophectoderm differentiation. *Cell*, 123:917–929, 2005.

A. Novick and M. Weiner. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences*, pages 553–566, 1957.

Y. Okuno, G. Huang, F. Rosenbauer, E. K. Evans, H. S. Radomska, H. Iwasaki, K. Akashi, F. Moreau-gachelin, Y. Li, and D. G. Tenen. Potential Autoregulation of Transcription Factor PU . 1 by an Upstream Regulatory Element. 25(7): 2832–2845, 2005.

S. H. Orkin and L. I. Zon. SnapShot: Hematopoiesis. *Cell*, 132(4):712.e1 – 712.e2, 2008.

J. R. Passweg, H. Baldomero, A. Gratwohl, M. Bregni, S. Cesaro, P. Dreger, T. D. Witte, D. Farge-Bancel, B. Gaspar, J. Marsh, M. Mohty, C. Peters, A. Tichelli, A. Velardi, C. R. de Elvira, F. Falkenburg, A. Sureda, and A. Madrigal. The EBMT activity survey: 1990–2010. *Bone Marrow Transplantation*, 47(March):906–923, 2012.

J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, 2005.

J. Paulsson and M. Ehrenberg. Random signal fluctuations can reduce random fluctuations in regulated components of chemical regulatory networks. *Physical Review letters*, 84(23):5447–50, June 2000.

H. Pendar, T. Platini, and R. V. Kulkarni. Exact protein distributions for stochastic models of gene expression using partitioning of Poisson processes. *Physical Review E*, 87(4):042720, April 2013.

X.-j. Peng and Y.-f. Wang. L-leap: accelerating the stochastic simulation of chemically reacting systems. *Applied Mathematics and Mechanics*, 28(10):1361–1371, 2007.

A. O. Pisco, A. Brock, J. X. Zhou, A. Moor, M. Mojtahedi, D. Jackson, and S. Huang. Non-Darwinian dynamics in therapy-induced cancer drug resistance. *Nature communications*, 4:2467, January 2013.

J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–8, December 1999.

X. Qiu, S. Ding, and T. Shi. Understanding the Development Landscape of the Canonical Fate-Switch Pair to Constructing a Dynamic Landscape for Two-Step Neural Differentiation. *PloS one*, 7(12), 2012.

Y. Rais, A. Zviran, S. Geula, O. Gafni, E. Chomsky, S. Viukov, A. A. Mansour, I. Caspi, V. Krupalnik, M. Zerbib, I. Maza, N. Mor, D. Baran, L. Weinberger, D. a. Jaitin, D. Lara-Astiaso, R. Blecher-Gonen, Z. Shipony, Z. Mukamel, T. Hagai, S. Gilad, D. Amann-Zalcenstein, A. Tanay, I. Amit, N. Novershtern, and J. H. Hanna. Deterministic direct reprogramming of somatic cells to pluripotency. *Nature*, 502(7469):65–70, 2013.

A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS biology*, 4(10):e309, October 2006.

R. Ramaswamy, N. González-Segredo, and I. F. Sbalzarini. A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks. *The Journal of chemical physics*, 130(24):244104, June 2009.

R. Ramaswamy, N. González-Segredo, I. F. Sbalzarini, and R. Grima. Discreteness-induced concentration inversion in mesoscopic chemical systems. *Nature Communications*, 3:779, April 2012.

A. F. Ramos, G. Innocentini, and J. E. M. Hornos. Exact time-dependent solutions for a self-regulating gene. *Physical Review E*, 83(6):1–4, June 2011.

D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent, G. P. Schroth, and R. Sandberg. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30(8):777–782, 2012.

J. M. Raser and E. K. O'Shea. Control of stochasticity in eukaryotic gene expression. *Science*, 304(5678):1811–4, June 2004.

M. A. Rieger, P. S. Hoppe, B. Smejkal, A. C. Eitelhuber, and T. Schroeder. Hematopoietic cytokines can instruct lineage choice. *Science*, 325(July), 2009.

I. Roeder and I. Glauche. Towards an understanding of lineage specification in hematopoietic stem cells: a mathematical model for the interaction of transcription factors GATA-1 and PU.1. *Journal of theoretical biology*, 241(4):852–65, August 2006.

I. Roeder and F. Radtke. Stem cell biology meets systems biology. *Development*, 136:3525–3530, 2009.

I. Roeder, L. M. Kamminga, K. Braesel, B. Dontje, G. De Haan, and M. Loeffler. Competitive clonal hematopoiesis in mouse chimeras explained by a stochastic model of stem cell organization. *Blood*, 105:609–616, 2005.

P. Rompolas, K. R. Mesa, and V. Greco. Spatial organization within a niche as a determinant of stem-cell fate. *Nature*, 502(7472):513–518, October 2013.

S. Roy and M. Lévesque. Limb regeneration in axolotl: is it superhealing? *The Scientific World Journal*, 6:12–25, 2006.

T. Sandoval-Guzmán, H. Wang, S. Khattak, M. Schuez, K. Roensch, E. Nacu, A. Tazaki, A. Joven, E. M. Tanaka, and A. Simon. Fundamental differences in dedifferentiation and stem cell recruitment during skeletal muscle regeneration in two salamander species. *Cell Stem Cell*, 14:174–187, 2014.

K. R. Sanft, S. Wu, M. Roh, J. Fu, R. K. Lim, and L. R. Petzold. StochKit2: Software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458, 2011.

Y. Sasagawa, I. Nikaido, T. Hayashi, H. Danno, K. D. Uno, T. Imai, and H. R. Ueda. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome biology*, 14(4):R31, 2013.

N. Scherf, M. Herberg, K. Thierbach, T. Zerjatke, T. Kalkan, P. Humphreys, A. Smith, I. Glauche, and I. Roeder. Imaging, quantification and visualization of spatio-temporal patterning in mESC colonies under different culture conditions. *Bioinformatics*, 28(18):i556–i561, September 2012.

T. Schroeder. Imaging stem-cell-driven regeneration in mammals. *Nature*, 453(7193):345–51, May 2008.

T. Schroeder. Long-term single-cell imaging of mammalian stem cells. *Nature Methods*, 8(4s):S30–S35, March 2011.

D. Schultz, J. N. Onuchic, and P. G. Wolynes. Understanding stochastic simulations of the smallest genetic networks. *The Journal of chemical physics*, 126(24):245102, June 2007.

D. Schultz, A. M. Walczak, J. N. Onuchic, and P. G. Wolynes. Extinction and resurrection in gene networks. *Proceedings of the National Academy of Sciences*, 105(49):19165–70, December 2008.

B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen, and M. Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–42, May 2011.

S. D. Schwartz, J.-P. Hubschman, G. Heilwell, V. Franco-Cardenas, C. K. Pan, R. M. Ostrick, E. Mickunas, R. Gay, I. Klimanskaya, and R. Lanza. Embryonic stem cell trials for macular degeneration: a preliminary report. *Lancet*, 379 (9817):713–20, 2012.

M. Schwarzfischer. *Quantification and analysis of single-cell protein dynamics in stem cells using time-lapse microscopy*. Phd thesis, Technische Universität München, 2013.

M. Schwarzfischer, C. Marr, J. Krumsiek, P. S. Hoppe, T. Schroeder, and F. J. Theis. Efficient fluorescence image normalization for time lapse movies. In *Proceedings of the Microscopic Image Analysis with Applications in Biology*, Heidelberg, 2011. MIAAB.

V. Shahrezaei and P. S. Swain. Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45):17256–61, November 2008.

A. K. Shalek, R. Satija, X. Adiconis, R. S. Gertner, J. T. Gaublomme, R. Raychowdhury, S. Schwartz, N. Yosef, C. Malboeuf, D. Lu, J. J. Trombetta, D. Gennert, A. Gnirke, A. Goren, N. Hacohen, J. Z. Levin, H. Park, and A. Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453):236–40, 2013.

A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, S. Schwartz, B. Fowler, S. Weaver, J. Wang, X. Wang, R. Ding, R. Raychowdhury, N. Friedman, N. Hacohen, H. Park, A. P. May, and A. Regev. Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, 509 (7505):363–9, 2014.

E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9):618–30, 2013.

M. A. Shea and G. K. Ackers. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *Journal of molecular biology*, 181:211–230, 1985.

N. Shevde. Stem Cells: Flexible friends. *Nature*, 483(7387):S22–S26, March 2012.

A. Shivanandan, A. Radenovic, and I. F. Sbalzarini. MosaicIA: an ImageJ/Fiji plugin for spatial pattern and interaction analysis. *BMC bioinformatics*, 14:349, January 2013.

D. Siegal-Gaskins, M. K. Mejia-Guerra, G. D. Smith, and E. Grotewold. Emergence of Switch-Like Behavior in a Large Family of Simple Biochemical Networks. *PLoS Computational Biology*, 7(5):e1002039, May 2011.

D. Silk, S. Filippi, and M. P. H. Stumpf. Optimizing threshold-schedules for approximate Bayesian computation sequential Monte Carlo samplers: applications to molecular systems. *arXiv preprint arXiv:1210.3296*, pages 1–18, 2012.

K. Singh, A. Srivastava, S. S. Patel, and M. J. Modak. Participation of the fingers subdomain of Escherichia coli DNA polymerase I in the strand displacement synthesis of DNA. *J Biol Chem*, 282(14):10594–10604, April 2007.

D. R. Sisan, M. Halter, J. B. Hubbard, and A. L. Plant. Predicting rates of cell state change caused by stochastic fluctuations using a data-driven landscape model. *Proceedings of the National Academy of Sciences*, October 2012.

S. Sisson, Y. Fan, and M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 106(39):16889–16889, September 2007.

D. S. Sivia and J. Skilling. *Data analysis: a Bayesian tutorial*. Oxford science publications. Oxford University Press, 2006.

J. B. Skeath and S. B. Carroll. Regulation of proneural gene expression and cell fate during neuroblast segregation in the Drosophila embryo. *Development*, 114:939–946, 1992.

A. Slepoy, A. P. Thompson, and S. J. Plimpton. A constant-time kinetic Monte Carlo algorithm for simulation of large biochemical reaction networks. *The Journal of chemical physics*, 128(20):205101, May 2008.

B. Snijder and L. Pelkmans. Origins of regulated cell-to-cell variability. *Nature Reviews. Molecular cell biology*, 12(2): 119–25, March 2011.

B. Snijder, R. Sacher, P. Rämö, E.-M. Damm, P. Liberali, and L. Pelkmans. Population context determines cell-to-cell variability in endocytosis and virus infection. *Nature*, 461(7263):520–3, September 2009.

S. L. Spencer, S. Gaudet, J. G. Albeck, J. M. Burke, and P. K. Sorger. Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. *Nature*, 459(May):428–432, 2009.

J. L. Spudich and D. E. Koshland. Non-genetic individuality: chance in the single cell. *Nature*, 262:467–471, 1976.

J. Stingl and C. Caldas. Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis. *Nature Reviews. Cancer*, 7(10):791–799, October 2007.

T. Stopka, D. F. Amanatullah, M. Papetti, and A. I. Skoultchi. PU.1 inhibits the erythroid program by binding to GATA-1 on DNA and creating a repressive chromatin structure. *The EMBO journal*, 24(21):3712–3723, 2005.

M. Strasser, F. J. Theis, and C. Marr. Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Biophysical journal*, 102(1):19–29, January 2012.

R. Strohman. Epigenesis: the missing beat in biotechnology? *Nature Biotechnology*, 1994.

R. C. Strohman. The coming Kuhnian revolution in biology. *Nature biotechnology*, 15:194–200, 1997.

T. Suda, J. Suda, and M. Ogawa. Disparate differentiation in mouse hemopoietic colonies derived from paired progenitors. *Proceedings of the National Academy of Sciences of the United States of America*, 81(April):2520–2524, 1984.

D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science*, 472, March 2011.

A. Taguchi, Y. Kaku, T. Ohmori, S. Sharmin, M. Ogawa, H. Sasaki, and R. Nishinakamura. Redefining the in vivo origin of metanephric nephron progenitors enables generation of complex kidney structures from pluripotent stem cells. *Cell Stem Cell*, 14(1):53–67, 2014.

K. Takahashi and S. Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell*, 126(4):663–76, August 2006.

K. Takahashi, K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka, K. Tomoda, and S. Yamanaka. Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell*, 131:861–872, 2007.

H. Takano, H. Ema, K. Sudo, and H. Nakauchi. Asymmetric division and lineage commitment at the level of hematopoietic stem cells: inference from differentiation in daughter cell and granddaughter cell pairs. *The Journal of experimental medicine*, 199(3):295–302, 2004.

F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

Y. Taniguchi, P. J. Choi, G.-W. Li, H. Chen, M. Babu, J. Hearn, A. Emili, and X. S. Xie. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science*, 329(5991):533–8, July 2010.

S. Temple. Division and differentiation of isolated CNS blast cells in microculture. *Nature*, 340:471–473, 1989.

M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, July 2001.

E. D. Thomas, H. L. Lochte, W. C. Lu, and J. W. Ferrebee. Intravenous infusion of bone marrow in patients receiving radiation and chemotherapy. *The New England journal of medicine*, 257(11):491–496, September 1957.

J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *Science*, 282(1998):1145–1147, 1998.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1): 267–288, 1996.

J. E. Till, E. a. McCulloch, and L. Siminovitch. a Stochastic Model of Stem Cell Proliferation, Based on the Growth of Spleen Colony-Forming Cells*. *Proceedings of the National Academy of Sciences of the United States of America*, 51: 29–36, 1964.

M. Tipping. Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, 2001.

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

N. G. Van Kampen. *Stochastic processes in physics and chemistry*, volume 11. 1992.

J.-W. Veening, W. K. Smits, and O. P. Kuipers. Bistability, epigenetics, and bet-hedging in bacteria. *Annual review of microbiology*, 62:193–210, January 2008.

J. Visvader and A. Elefanty. GATA-1 but not SCL induces megakaryocytic differentiation in an early myeloid line. *The EMBO journal*, 1(12):4557–4564, 1992.

C. H. Waddington. *The Strategy of the Genes*. George Allen & Unwin, 1957.

D. E. Wagner, I. E. Wang, and P. W. Reddien. Clonogenic neoblasts are pluripotent adult stem cells that underlie planarian regeneration. *Science*, 332:811–816, 2011.

A. M. Walczak, J. N. Onuchic, and P. G. Wolynes. Absolute rate theories of epigenetic stability. *Proceedings of the National Academy of Sciences*, 102(52):18926–31, December 2005a.

A. M. Walczak, M. Sasai, and P. G. Wolynes. Self-consistent proteomic field theory of stochastic gene switches. *Biophysical journal*, 88(2):828–50, February 2005b.

A. M. Walczak, A. Mugler, and C. H. Wiggins. A stochastic spectral analysis of transcriptional regulatory cascades. *Proceedings of the National Academy of Sciences*, 106(16):6529–34, April 2009.

A. M. Walczak, A. Mugler, and C. H. Wiggins. Analytic methods for modeling stochastic regulatory networks. *Arxiv preprint arXiv10052648*, pages 1–37, 2010.

J. Wang, L. Xu, and E. Wang. Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation, and coherence of biochemical oscillations. *Proceedings of the National Academy of Sciences*, 105:12271–12276, 2008.

J. Wang, L. Xu, E. Wang, and S. Huang. The potential landscape of genetic circuits imposes the arrow of time in stem cell differentiation. *Biophysical journal*, 99(1):29–39, July 2010.

J. Wang, K. Zhang, L. Xu, and E. Wang. Quantifying the Waddington landscape and biological paths for development and differentiation. *Proceedings of the National Academy of Sciences*, May 2011.

L. D. Wang and A. J. Wagers. Dynamic niches in the origination and differentiation of haematopoietic stem cells. *Nature Reviews. Molecular cell biology*, 12, September 2011.

L. Warren, D. Bryder, I. L. Weissman, and S. R. Quake. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences*, 103(47):17807–12, November 2006.

P. B. Warren and P. R. ten Wolde. Enhancement of the Stability of Genetic Switches by Overlapping Upstream Regulatory Domains. *Physical Review Letters*, 92(12):128101, March 2004.

P. B. Warren and P. R. ten Wolde. Chemical models of genetic toggle switches. *J Phys Chem B*, 109(14):6812–6823, April 2005.

D. J. Wilkinson. *Stochastic modelling for systems biology*. CRC press, 2011.

C. Yu, A. B. Cantor, H. Yang, C. Browne, R. a. Wells, Y. Fujiwara, and S. H. Orkin. Targeted deletion of a high-affinity GATA-binding site in the GATA-1 promoter leads to selective loss of the eosinophil lineage in vivo. *The Journal of experimental medicine*, 195(11):1387–1395, 2002.

C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros, and H. Koeppl. Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21), May 2012.

P. Zhang, G. Behre, J. Pan, A. Iwama, N. Wara-Aswapati, H. S. Radomska, P. E. Auron, D. G. Tenen, and Z. Sun. Negative cross-talk between hematopoietic regulators: GATA proteins repress PU.1. *Proceedings of the National Academy of Sciences*, 96(July):8705–8710, 1999.

J. X. Zhou and S. Huang. Understanding gene circuits at cell-fate branch points for rational cell reprogramming. *Trends in genetics*, 27(2):55–62, February 2011.

J. X. Zhou, M. D. S. Aliyu, E. Aurell, and S. Huang. Quasi-potential landscape in complex multi-stable systems. *Journal of The Royal Society Interface*, 9(August):3539–3553, 2012.

R. Zhu, A. S. Ribeiro, D. Salahub, and S. A. Kauffman. Studying genetic regulatory networks at the molecular level: Delayed reaction stochastic models. *Journal of Theoretical Biology*, 246:725–745, 2007.

G. Zou. A Modified Poisson Regression Approach to Prospective Studies with Binary Data. *American Journal of Epidemiology*, 159(7):702–706, April 2004.