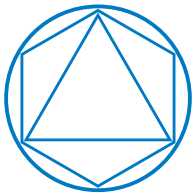
HelmholtzZentrum münchen
German Research Center for Environmental Health

Technische Universität München

Data-driven statistical learning to model cellular heterogeneity

Dissertationsschrift
an der Fakultät für Mathematik
der Technischen Universität
München



vorgelegt von

Thomas Blasi

München, den 22.03.2016

TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl M12: Mathematical Modeling of Biological Systems

"Data-driven statistical learning to model cellular heterogeneity"

Thomas Blasi

Vollständiger Abdruck der von der Fakultät für Mathematik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Univ.-Prof. Dr. Massimo Fornasier

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Dr. Fabian J. Theis
2. Univ.-Prof. Dr. Julien Gagneur
3. Prof. John Marioni, Ph.D. (nur schriftliche Beurteilung)

Die Dissertation wurde am 22.03.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Mathematik am 18.07.2016 angenommen.

Für meine Eltern.

Acknowledgments

First of all I would like to thank my supervisors: Carsten thanks a lot for all the great efforts you made in guiding me through my thesis, for your patience with me and for introducing me into all these fascinating topics in single cell biology. Flo thanks to you for introducing me to the world of machine learning and for always sharing your wisdom and intuition on what new methods to try. Fabian thank you for giving me the opportunity to conduct my thesis at your wonderful institute; thanks for always sharing your fruitful insights and views on my research topics. It has been a great pleasure for me to work at the edge of mathematics, biology and computer science during the last three years – thanks to the three of you for making this possible.

Next I would like to thank the members of my thesis examination committee: Massimo Fornasier for agreeing to chair the examination as well as Julien Gagneur, John Marioni and Fabian for agreeing to review my thesis and to give the exam.

Then I would like to thank Michi St., Justin, Sabine and Sabrina who always had an open ear (and just as importantly the right answers) for the myriads of questions that I had, especially in the beginning of my thesis. Moreover, I would like to thank Jan H. for sharing a lot of knowledgeable advices with me and also for providing a great lecture on parameter estimation for biological systems.

My next big thank you goes to Elisabeth and Marianne – without your support in all the administrative stuff I wouldn't have had the opportunity to focus so much on research.

Furthermore my thanks go to all members of the QSCD and Machine Learning groups for being a constant source of support and a pool of fresh ideas. The same holds for the whole ICB. Thanks to all of you for also creating this great working atmosphere that I have always felt very enjoyable.

Another thank you goes to the Studienstiftung des deutschen Volkes who not only founded me for most of my dissertation here in Munich, but who also enabled me to spend two very fascinating research stays abroad.

This is also where I would like to direct my next thanks to:
Thanks to Anne and to the whole CellProfiler group for hosting me; I found my stay with you very inspiring and I am very happy about the fruitful collaboration we have been having since then. Thanks also to John and the rest of the Marioni group for having me over; I have highly enjoyed our discussions about single cell topics and I have benefited a lot from your feedback on my projects.

Thank you also John, for being on my thesis advisory committee; I have always been very delighted by your ideas and opinions.

Moreover, I would like to thank all my external collaboration partners: Thank you Paul for being the constantly available hub during the writing and revision phase of our 'label-free analysis paper'; thank you also for hosting me over in Swansea, I had a wonderful time there.
Thanks to Christian, Peter and Axel for your input and feedback during the shaping and writing of our 'histone paper'; I think it was a very exciting process.

I would also like to thank the organizers of the HELENA and TUM graduate schools. I always enjoyed their seminars and did profit a lot from them. They also provided a good opportunity to get in touch with PhD students from other disciplines, which I found always very interesting.

Yet another thank you to my three office mates, Laleh, Atefeh and Michi L. I have highly enjoyed spending so much time together with you.

Last but not least I would like to thank my family and friends; thanks for being there for me. Stefanie, thank you!

Abstract

In the last decades the advent of new experimental techniques has lead to a drastic increase of available data in biology. As a consequence the importance of mathematical methods to deduct scientifically relevant hypothesis from this big amount of data is steadily growing. A major challenge for bio-mathematics and bio-statistics therefore lies in both the adaption of existing methods to the, often very specific, properties of the measured data, and in the development of new methods to model these data.

In this thesis we present methods from statistics and machine learning that are suitable to perform this task. The quest for new mathematical methods, thereby, is always pursued in conjunction with the goal to find new scientific insights into the investigated biological system. The biological focus of this work is the analysis of heterogeneity among cells: almost all cells of a living organism share the same DNA, yet there is a multitude of different cell types that may all perform different tasks within the organism. The aim of this thesis is to explore both the biological principles that lead to cellular heterogeneity, and to improve the identifiability of different cellular phenotypes with mathematical methods.

For this purpose four different mathematical methods are implemented, tested and applied to biological data in order to draw new conclusions about cellular heterogeneity:

i We propose a statistical method to correct for latent confounding effects on single cell transcriptomics data that are due to differences in cell size, which we show to have an impact on the inference of the underlying gene expression mechanism.

ii By applying ordinary-differential-equation-based models on chromatin data we can show that histone acetylation (a certain class of chromatin modifications with known impact on transcriptional regulation) depends specifically on the chromatin status before these modifications occur.

iii We apply transfer entropy to protein time-series data from hematopoietic stem and progenitor cells and find that the information transfer between two key transcription factors differs depending on the final cellular phenotype of the progenitor cells.

iv By the help of machine learning methods we show that cellular phenotypes can be identified without the need for chemical fluorescent stains relying entirely on bright field and dark field images of the cells.

To conclude, we anticipate the contributions of bio-mathematics and biostatistics for the quest of deciphering and understanding the myriad biochemical processes (and the molecular species involved in them) that eventually lead to the emergence of cellular heterogeneity to become more and more important.

Zusammenfassung

Das Auftreten neuer experimenteller Methoden führte in den letzten Jahren zu einer drastischen Zunahme verfügbarer, biologischer Daten geführt. Um aus dieser Datenmenge wissenschaftlich relevante Ergebnisse abzuleiten, wird es daher immer wichtiger geeignete mathematische und rechnergestützte Methoden zu finden. Eine der größten Herausforderungen im Bereich der Biomathematik und Biostatistik ist es daher, bestehende Methoden an die oft sehr spezifischen Eigenschaften der Daten anzupassen und zusätzlich neue Methoden für die Modellierung dieser Daten zu finden.

In dieser Arbeit werden Methoden aus der Statistik und dem maschinellen Lernen vorgestellt, die dieser Aufgabe gerecht werden. Dabei wird die Entwicklung neuer mathematischer Methoden stets mit dem Versuch verknüpft, wissenschaftliche Einblicke in das untersuchte biologische System zu gewinnen. Der biologische Fokus dieser Arbeit liegt auf der Analyse von Heterogenität zwischen Zellen: die Zellen eines lebenden Organismus besitzen fast alle die gleiche DNA, dennoch findet man eine Vielfalt an verschiedenen Zelltypen vor, die jeweils andere Aufgaben im Organismus erfüllen. Ziel dieser Arbeit ist es, mit mathematischen Methoden sowohl die Ursachen dieser Heterogenität zu analysieren, als auch die Identifizierung von verschiedenen Zellzustände zu verbessern.

Dazu werden vier verschiedene mathematische Methoden implementiert, getestet und auf biologische Daten angewandt, um so neue Rückschlüsse über das Auftreten von Heterogenität von Zellen ziehen zu können:

i Es wird eine statistische Methode vorgestellt mit der bei Einzelzell-Transcriptomics Daten auftretende latente Störfaktoren korrigiert werden können, welche durch Unterschiede in der Zellgröße zustande kom-

men und welche die Inferenz des zugrunde liegenden Genexpressionsmechanismus beeinflussen.

ii Durch das Anwenden von Modellen basierend auf gewöhnlichen Differentialgleichungen auf Chromatindaten wird gezeigt dass Histonacetylierungen (eine spezielle Klasse von Chromatinmodifikationen, die für die Regulation der Genexpression von Bedeutung sind) spezifisch vom Zustand des Chromatins vor Eintreten der Modifizierung abhängen.

iii Mit Hilfe der Transferentropie werden zeitaufgelöste Proteindaten von hämatopoetischen Stamm- und Vorläuferzellen untersucht und wird festgestellt, dass sich der Informationsfluss zwischen zwei wichtigen Transkriptionsfaktoren abhängig vom finalen Phänotypen der Zellen unterscheidet.

iv Durch die Anwendung von Methoden aus dem Bereich des maschinellen Lernens wird gezeigt, dass verschiedene Zellphänotypen ohne zusätzliche chemische Farbstoffe klassifiziert werden können.

Mit Blick auf die Zukunft kann davon ausgegangen werden, dass die Rolle der Biomathematik für die Erforschung der zahlreichen biologischen Prozesse, die letztlich zu Heterogenität zwischen Zellen führen, immer wichtiger werden wird.

Contents

List of Figures

Chapter 1

Introduction

Eine Reise, tausend Meilen lang,
mit einem ersten Schritt fing sie an!

Laozi [I]

1.1 Scientific question of this thesis

Cells are the fundamental building units of all life on earth. The human organism for instance is formed by $\sim 4 \times 10^{13}$ single cells (Bianconi et al., 2013). Even though the DNA is the same in almost all cells of an organism (exceptions are, e..g. red blood cells and cells from the immune system), cells occur in many different cellular phenotypes leading to a vast heterogeneity. Important manifestations of different cellular phenotypes are, e.g., the $\sim$ 2,000 different cell types that can be found in the human organism (Hatano et al., 2011; SHOGoiN, 2016) and the different phases of the cell cycle. Such different cellular phenotypes are often but not always reflected in a distinct morphological appearance.

A fundamental question in biology is how does this cellular heterogeneity arise? The answer to this question seems simple: the emergence of cellular heterogeneity is due to differences in gene expression. On a coarse level gene expression consists of two subsequent processes that link three different scales in the cell: first DNA that contains the genetic information is transcribed into mRNA, which is then translated to functional

proteins. Gene expression can be regulated such that the present number of protein species can be very different from cell to cell depending on the details of the regulation. These details, however, are still not fully understood and also the means to tell different phenotypes apart are limited (Pennacchio et al., 2013).

The reason for this is that even an isolated single cell is a highly complex system. In its small volume of $\sim 1,000 - 10,000 \mu m^3$ it contains $\sim 10^{10}$ molecules of proteins and $\sim 10^5 - 10^6$ molecules of mRNAs (in the case of mammalian HeLa cells; Milo and Phillips (2015)), which both can be present in plenty of different species. At first glance one might assume that the number of different mRNA and protein species in the cell is limited by the number of genes in the DNA given by $\sim 22,000$ in case of the human organism (Pertea and Salzberg, 2010). A single gene, however may give rise to multiple mRNA species via a process known as alternative splicing (Matlin et al., 2005). More importantly after translation, proteins can be modified in myriad ways: so far $\sim 150,000$ different species of these so called post-translational modifications (PTMs) have been reported (The-UniProt-Consortium, 2015, 2016) and estimates anticipate $> 1,000,000$ different PTMs (Jensen, 2004).

From a computational point of view it is very hard (if not to say infeasible) to keep track of the positions and velocities of all these individual molecules and their biochemical interactions in the cell. However, by making certain assumptions on the system, it is possible to find mathematical formulations, which allow to describe a biological process of interest (e.g. the regulation of gene expression). A very successful approach that fulfills this purpose is given by the chemical master equation (CME; Gillespie (1992)), where the positions and the velocities of the molecules are marginalized out and the dynamics of the system can be described in terms of the present molecule numbers. The CME provides a fully parametrized mathematical formulation of biochemical processes that can be easily interpreted. Even though the CME cannot be solved analytically in most situations there is an efficient way to simulate individual realizations from it using the stochastic simulation algorithm (SSA; Gillespie (1976)).

To investigate cellular processes, however, we have to go one step further and link its mathematical formulation with experimental data. Biological data is – from a mathematical point of view – nothing else than a tensor, the dimensions of which equal the number of measured features (e.g. molecular species, or morphological properties)

times the number of samples times the number of measurement time points. Although great experimental progresses have been made in the last decades making more and more data available even on a single cell level (Hoppe et al., 2014; Rubakhin et al., 2011; Tang et al., 2011), the data is still often limited due to the experimental procedures that are performed to measure it. While measurements of the morphology of a cell can conveniently be performed using microscopes and do not constitute an invasive procedure to the cells, a measurement of the molecular species present within the cell is often very complicated and restrictive. In many cases it is only possible to measure a small fraction of the molecular species that are involved in a biological process rendering the data incomplete. Moreover, in most cases the experiments disrupt the cell such that only a snapshot of the species at a single point in time can be measured. Lastly, the data quality is often limited due to technical noise that comes from the experiments or even from biological sources of additional variability that may not be of interest for the specific question at hand.

This leaves us with the following problem statement: given the data of only a fraction of the features from a biological process of interest, can we still infer something about it? In this thesis we use methods from statistics and machine learning to pursue this goal. We work with parameterized models (as derived from the CME) and link them to biological data by formulating and maximizing the likelihood for the model parameters given the data and a particular model topology. This approach allows to estimate probability distributions for the considered parameters and to analyze the uncertainty in the estimate enabling to assess if the information content of the data is sufficient to draw conclusions. Moreover, we use methods from machine learning that do not rely on a particular model definition, but look for significant differences within the data itself. These methods can be used to dissect cellular heterogeneity and to identify different cellular phenotypes. Often these methods have to be adapted to both, the specific biological question that we aim to answer and the experimental procedure. Further challenges lie in finding mathematical problem formulations that can be answered with some given data, but also in the development of new experimental and computational workflows that lead to more informative data.

The specific scientific questions that we aim to address with this thesis are:

i Can we improve on the existing methods to infer the regulatory status of a gene and to dissect cellular heterogeneity from single cell transcriptomics data by taking latent differences in cell size into account that are caused by the cell cycle?

ii Can we find a signature of specific and targeted activity in the acetylation of histone protein tails, which is known to affect the accessibility of genes in its proximity and therefore influences their expression?

iii Can we find distinguishing properties in the interplay of two key transcription factors in hematopoietic stem and progenitor cells that differentiate into two distinct lineages?

iv Can we discriminate the different stages of the cell cycle based entirely on morphological properties of the cells?

To conclude, the scientific aim of this thesis is the application of existing and the development of new mathematical methods to (i) gain new insights into the underlying biological mechanisms that lead to cellular heterogeneity and (ii) to improve the ability to dissect different cellular phenotypes in a data-driven way. Here, the aim to find new mathematical methods is always linked with the aim to find new insights about the underlying biological processes. This makes the scientific topic of this thesis an interdisciplinary field of research, where mathematics, computer science and biology come together.

1.2 Overview over this thesis

In Chapters 2-4 we summarize the background of the biology and the existing mathematical methods that we extend and/or apply in the original contributions of this thesis in Chapters 5-8.

In Chapter 2 we introduce the biological background relevant for this thesis. We discuss the biological process of gene expression and two major ways for its regulation, namely DNA accessibility and gene regulatory networks. We then explain how regulation of gene expression eventually leads to the emergence of cellular heterogeneity. Moreover, we distinguish heterogeneity from variability, which is caused by stochastic gene expression. We then elaborate on the properties of biological data and explain to what extent

current biological data is limited for the inference of cellular heterogeneity. To this end we elucidate the differences between single cell measurements and measurements based on cell populations and discuss what features can be extracted from the cell. We conclude Chapter 2 by naming two important examples of cellular heterogeneity: hematopoiesis (the formation of adult blood cells) and cell cycle.

Chapter 3 provides the mathematical framework to describe biochemical processes, the chemical master equation (CME), and discusses how realizations of biochemical processes can be generated with the stochastic simulation algorithm (SSA). Moreover, we show how in certain limits the CME, which governs stochastic and discrete dynamics, can be converted to the reaction rate equation (RME), which is deterministic and continuous. Finally, we discuss the implications of stochasticity for biological processes and simulate an example for regulated gene expression.

Chapter 4 provides the statistical learning methods that we use to bring together biological data and the mathematical framework that describes biological processes. These methods range from parameter estimation and model selection, over time-series analysis, clustering and dimension reduction to classification and regression.

In Chapter 5 we present cgCorrect (cell growth correction) a novel statistical method to correct snapshot data sets for differences in cell size, which we point out to constitute a confounding source of variability for current analysis methods. We use this method to infer the regulatory status of several genes that are important during hematopoiesis

In Chapter 6 we developed a tailored mathematical framework to study general properties of histone modifications, which influence DNA accessibility, a key ingredient for regulation of gene expression. With our framework we find that combinatorial acetylation patterns on the histone H4 tail are specifically acetylated.

In Chapter 7 we use transfer entropy, a recently published mathematical method, to analyze information transfer among two protein species in single cells during hematopoiesis. We show that the information transferred between the two protein species is distinct for two different lineages the cells differentiate to.

In Chapter 8 we use probabilistic learning methods within a novel computational workflow that facilitates the dissection of cellular heterogeneity during the cell cycle. We show that the identification of cell cycle phase, which was traditionally performed by

labeling the cells with fluorescent stains, can also be achieved in a purely non-invasive way based on morphological properties.

Chapter 9 contains a conclusion on the mathematical methods presented in this thesis and gives an outlook on what further steps can be undertaken to infer cellular heterogeneity.

1.3 Main scientific contributions

The main scientific contributions of this thesis are the development of new mathematical methods and the adaption of existing mathematical methods to gain new insights into cellular heterogeneity.

Specifically, we developed novel methods to:

- correct confounding cell size effects that are due to cell cycle.
- predict reaction pathways for histone tail modifications.
- analyze protein time-series with respect to their information transfer.
- predict the cellular phenotype of cells based on their morphological properties.

This lead to the following new biological insights:

- Cell growth effects may obscure the underlying biological processes.
- Histone modification patterns are set in a highly specific manner.
- Information transfer from PU.1 to Gata-1 during hematopoiesis is distinct between cells of the megakaryocyte-erythrocyte and the granulocyte-macrophage lineage.
- Information about molecular species inside the cell can be inferred from the morphology alone.

These contributions are in part already contained in peer-reviewed publications and in a provisional patent application. Parts of this thesis therefore correspond to or are in parts identical to the publications listed here:

Blasi, T., Buettner, F., Strasser, M.K., Marr, C. and Theis, F.J. cgCorrect: A method to correct for confounding cell-cell variation due to cell growth in single-cell transcriptomics. *In preparation.*

Blasi, T., Feller, C., Feigelman, J., Hasenauer, J., Imhof, A., Theis, F.J., Becker, P.B. and Marr, C. (2016). Combinatorial histone acetylation patterns are generated by motif-specific reactions. *Cell Systems* 2:49–58.

Gumpinger, A.*, **Blasi, T.***, Hennig, H., Theis, F.J. and Marr, C. Transfer entropy: PU.1 regulates Gata1 in MEPs but not in GMPs. *In preparation.*

Blasi, T., Hennig, H., Summers, H.D., Theis, F.J., Cerveira, J., Patterson, J.O., Davies, D., Filby, A., Carpenter, A. E. and Rees, P. (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nature Communications* 7:10256.

Hennig, H.*, **Blasi, T.***, Rees, P.* and Carpenter, A.E.* (2014). Method for Label-Free Image Cytometry. US 61/985,236.

(* equal contributions)

The content of the third of the listed manuscripts, which is the basis for Chapter 7 of this thesis is also partially contained in a master thesis (Gumpinger, 2015) that the author of this thesis co-supervised. We explicitly state the individual contributions in the beginning of Chapter 7.

Chapter 2

Biological background: cellular heterogeneity

> Doch Forschung strebt und ringt, ermüdend nie,
> Nach dem Gesetz, dem Grund, Warum und Wie.
>
> *Johann Wolfgang von Goethe* [II]

Even though almost all cells in an organism have the same DNA they can be found in many different phenotypes. So far great progress has been made to understand this remarkable concept of nature: we know that the underlying reason for cellular heterogeneity is that different cellular phenotypes express different sets of protein species. For this, cells have the ability to highly regulate the expression of genes into proteins. We start this Chapter by giving a brief (but by no means fully comprehensive) summary of the current understanding of these regulatory processes on a molecular level. In Section 2.1 we present the central dogma of molecular biology that explains how DNA is transcribed into mRNA, which in turn is translated into proteins and discuss how differences in gene expression give rise to cellular heterogeneity and explain how gene expression can be regulated. There, we focus on two ways of transcriptional regulation, namely DNA accessibility and transcription initiation.

In contrast to cellular heterogeneity, which is caused by the regulation of gene expression there is a fundamental limit to what extend cellular processes can be regulated given by the laws of thermodynamics that cause gene expression to be a stochastic process.

In Section 2.2 we discuss of stochastic gene expression leads to variability in gene expression among cells that we distinguish from cellular heterogeneity in this thesis.

In order to obtain insights about cellular processes and to learn about cellular heterogeneity we always need experimental evidence. The data that can be measured from the cell, however, is limited and only information on certain scales (i.e. chromatin, mRNA, proteins, morphology, etc.) of the cell can be extracted. If we had a way to watch biological processes of interest in live cells we would be in great position to answer most of the questions that are still open. This, however, is not always possible and we have to rely on information extracted from the cell via complex biochemical procedures, which are often incomplete (i.e. only one or a few scales of gene expression can be measured at a time), invasive (often the cells are even killed) and subjected to high levels of measurement noise. In Section 2.3 we describe the difference between experiments that are performed on single cells as compared to experiments that are performed on cellular material from a population of cells. Then, in Section 2.4 we discuss the scales within a cell that can currently be experimentally assessed and the properties of the measured features (e.g. on the mRNA scale the features correspond to the abundance of the measured mRNA species). Finally, in Section 2.5, we give two examples of biological systems that are of particular interest for this thesis, namely hematopoiesis and the cell cycle, which both exhibit cellular heterogeneity.

It is important to note that despite the understanding that we currently have about cellular processes (as described in Section 2.1) there are still a lot of open questions that are partly due to a lack of available data but also due to a lack of suitable mathematical models to analyze existing data sets (a point we will take on in Chapter 4). The two questions that we aim to address in this thesis are (i) how can we gain knowledge about the mechanism that leads to cellular heterogeneity and (ii) how can we improve the classification of cell types given information about a mixture of cells.

2.1 Regulation of gene expression leads to cellular heterogeneity

In this Section we discuss how cellular heterogeneity arises even though the DNA of almost all cells in an organism is the same. The key to this is that gene expression is highly regulated in order to generate differences in the abundance of protein species from cell to cell. We start by introducing the central dogma of molecular biology, which describes how genetic information is transferred from DNA to functional proteins within a cell (Crick, 1958, 1970). It involves three scales, DNA and its chromatin environment, mRNAs and proteins, which are linked via transcription and translation, respectively.

In fact every single step involved in this process can be (and in fact is) subjected to regulation (see also Alberts et al. (2006)). Recent findings include the regulation of RNA processing (Licatalosi and Darnell, 2010), RNA transport and localization (Wickramasinghe and Laskey, 2015), translation (Jackson et al., 2010), mRNA degradation (Schoenberg and Maquat, 2012) and protein activity (Petsko and Ringe, 2004). Here, however, we focus only on transcriptional regulation, which regulates the initiation of transcription. This process is of particular importance for the cell since it is the first step of gene expression and its control ensures that the cell produces no superfluous products. Transcriptional regulation cannot only switch the expression of specific genes entirely on or off but can also lead to a subsequent transition between these two states.

For the purpose of this thesis we subdivide transcriptional regulation into two parts: (i) regulation of DNA accessibility and (ii) regulation of transcription initiation. In general, both processes are intimately related and both of them are performed by transcription factors that may even be present in a single protein complex. Here, however, we discuss them as a two step process such that the densely packaged DNA has to be made accessible for the transcription machinery first and then the initiation of transcription can be regulated in the next step. The first of the following two subsections is in parts adopted from Phillips and Shaw (2008) and the second subsection is in parts adopted from chapter 7 of Alberts et al. (2006).

Gene expression: the central dogma of molecular biology

Nowadays the fact that the genetic information of all cells is decoded in the deoxyribonucleic acid (DNA) is taken for granted. The first evidence, however, that DNA is the likely carrier of genetic information goes only back to the 1940s (Avery et al., 1944) and it became only clear in 1953 how DNA can both decode the building instructions for proteins and be faithfully replicated within the cell cycle when Watson and Crick correctly predicted the double-helical structure of DNA. The DNA molecule consists of two complementary DNA strands that are composed of four different nucleotides, which only differ with respect to their nucleobases, adenine (A), cytosine (C), guanine (G) and thymine (T). It is the sequence of many of these 4 base pairs that build up the genetic information of a cell: $\sim 2.9 \times 10^9$ in the case of the human genome (International-Human-Genome-Sequencing-Consortium, 2004).

In eukaryotes the DNA is placed in the nucleus, which forms a subunit of the cell with a diameter of $\sim 6\mu$m (in case of the human genome). The length of the human genome, however, if it was rolled out to a line, would have a length of $\sim$ 2m. From this fact it becomes clear that DNA has to be efficiently packaged to fit into the nucleus. This task is performed by the formation of chromatin. The basic unit of chromatin consists of nucleosomes (Kornberg, 1974), a complex of DNA and histone proteins that serve as a structuring unit, around which $\sim$ 200 nucleotide pairs of DNA are wrapped. The core of the nucleosome is formed by a histone octamer built by two proteins of each histone H2A, H2B, H3 and H4. Nucleosomes are further packed into 30 nm wide chromatin fibers, which then are coiled on several levels to form $\sim$ 700 nm wide domains. These chromatin domains can then form highly condensed chromosomes with a width of $\sim$ 1,400 nm (see Figure 2.1 A). Even though DNA is highly packaged in the chromatin, specific regions of it must be made accessible in order to be read out. Chromatin therefore has to be a dynamic structure that allows fast, localized and on-demand DNA accessibility. This requirement makes chromatin a highly conserved concept in evolution that is very similar for all eukaryotic cells.

Before the genetic information on DNA can be transformed into functional proteins an intermediate step, known as transcription, takes place. A gene on the template strand of DNA containing the information to build its corresponding protein species is

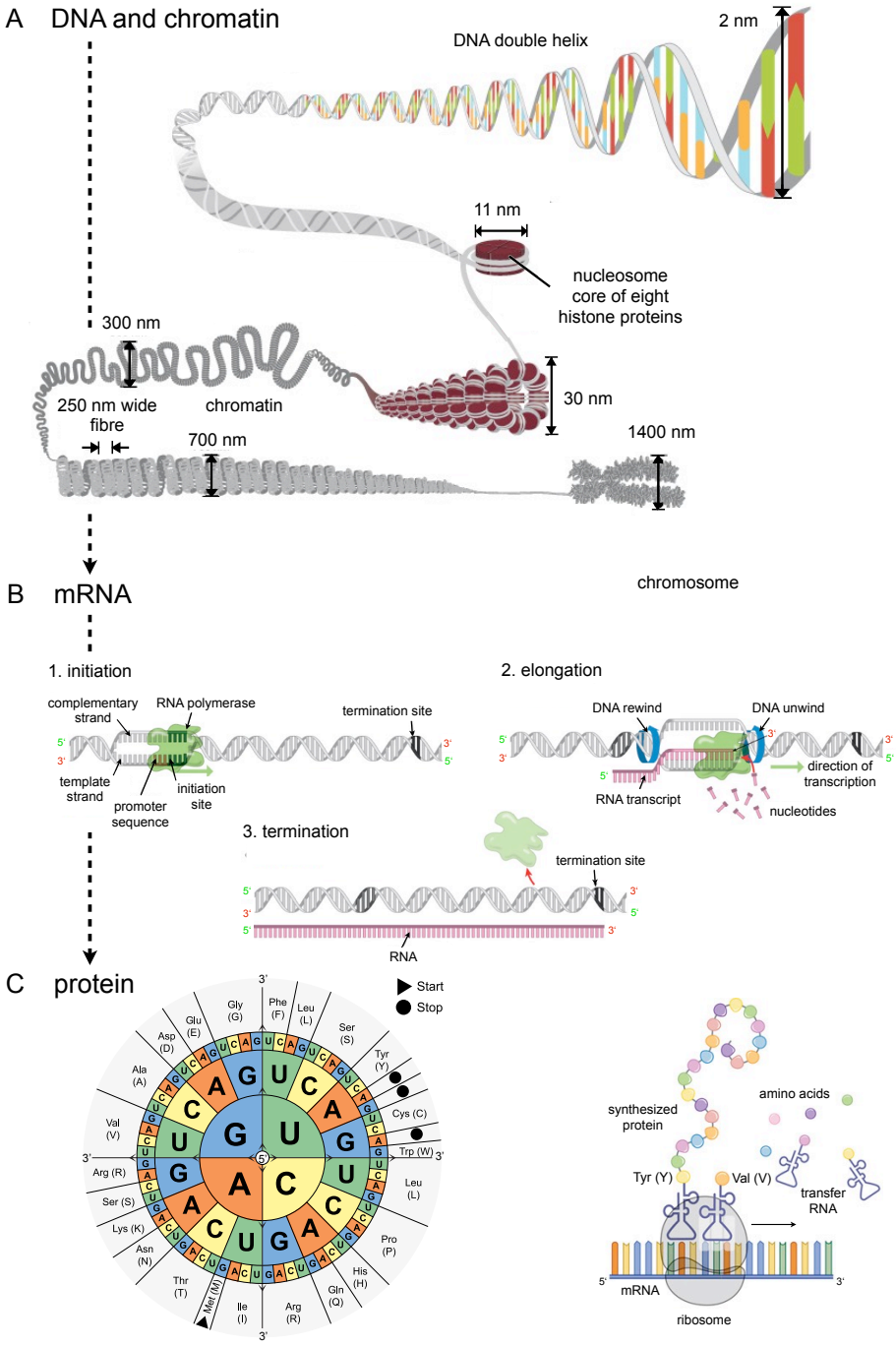


Figure 2.1: The central dogma of molecular biology: DNA, transcription and translation (Figure legend on next page).

Figure 2.1: (From previous page). The central dogma of molecular biology. (A) For DNA to fit into the nucleus it has to be densely packed. Its two complementary DNA strands have a double helix structure with a diameter of ~ 2 nm. The DNA double helix is then wrapped around a nucleosome consisting of an histone octamer. Nucleosomes are further organized into a 30 nm broad chromatin fibre, which is again wrapped on several scales until a 250 nm wide and 700 nm broad chromatin coil. This chromatin coil is further packed into 1,400 nm wide rod-shaped chromosomes that can be seen during cell division (Figure adopted from Annunziato (2008)). (B) Transcription, the process in which DNA is copied into RNA. In the transcription initiation step, RNA polymerase binds to the promoter sequence on the template strand of the DNA and transcription starts. In the elongation step, the RNA transcript becomes assembled by the RNA polymerase adding nucleotides to the transcript. This process does not necessarily proceed smoothly, since the DNA has to be unwound before it is accessible for transcription. In the termination step, the polymerase unbinds from the DNA, once it reached the termination site on the template strand (Figures adopted from Clancy (2008)). (C) Left: Representation of the genetic code. The sequence of three nucleotides – a codon (to be read from the centre outward) – becomes translated into an amino acid (on the outer circle). Since there are 64 different sequences that can be built with three nucleotides most of the amino acids are represented by multiple sequences (Figure adopted from Wikipedia (2016)). Right: Translation occurs in the ribosome. Specialized transfer RNA molecules can recognize the codons and have the appropriate amino acid attached that they add to the newly synthesized protein (Figure adopted from O'Connor and Adams (2010)).

transcribed into a complementary single strand of ribonucleic acid (RNA) containing exactly the same information as the template DNA strand. This process consists of three steps: initiation, elongation and termination (see Figure 2.1 B). In general, there are different sorts of RNA, only one of them, called messenger RNA (mRNA), contains building instructions for proteins. The primary RNA template becomes - still during its assembly - subsequently modified: first, a methyl cap is added to the protruding end of the mRNA, which makes it distinct from other RNA; then the transcribed intron sequences are removed (a process called splicing); finally a poly-A tail is added to the other end of the mRNA. Once these processes are finished, the mRNA is selectively exported from the nucleus to the cytosol of the cell.

After the mRNA has been transported to the cytosol it is translated into proteins by ribosomes. The information of the mRNA consisting of a sequence of 4 different nucleotides has to be transferred into proteins that are built up of 20 different amino

acids. The rules by which this translation occurs is known as the genetic code and was deciphered in the 1960s (Crick et al., 1961; Nirenberg and Matthaei, 1961). The mRNA is read out in subsequent groups of three nucleotides called codons, which can be present in $4 \times 4 \times 4 = 64$ possible combinations. Some of the 64 codon combinations are therefore redundant and some amino acids are specified by multiple codons (see Figure 2.1 C left). In the ribosome the codons are recognized by transfer RNA (tRNA) that are attached to the specific amino acid the codon stands for. In this way amino acids are subsequently added to the newly synthesized protein (see Figure 2.1 C right) until a stop codon is reached and the protein is released from the ribosome. Note that mRNAs can also be translated by many ribosomes simultaneously such that a single mRNA can give rise to many proteins very quickly. Before the synthesized proteins can perform their functions in the cell they are folded, modified and undergo a complex quality control process that ensures that the process of translation was successful.

Regulation of DNA accessibility

The key for the establishment of cellular heterogeneity is the ability of the cell to regulate the expression of genes. In this subsection we focus on the regulation of DNA accessibility to perform this task before we discuss the regulation of transcription initiation in the next subsection. As discussed in the previous paragraph DNA is highly packaged into chromatin, which in turn consists of many individual nucleosomes around which the DNA is wrapped. Before the transcriptional process can start the DNA sequence containing the gene that is to be transcribed must be made accessible for the transcription machinery (see Figure 2.2 A). This topic is studied by the field of epigenetics, which is the study of functional effects that are not caused directly by the DNA but by the state of its surrounding. Epigenetic modifications can be grouped into the following main categories: DNA methylation, histone modifications and chromatin remodeling (Portela and Esteller, 2010).

DNA methylation can occur at a di-nucleotide sequence consisting of C and G and is generally associated with inhibition of gene expression. The histone proteins that form the core of the nucleosome have protruding tails, which can be modified in many different ways (among others they can be methylated and acetylated). Depending on where and in what way the histone tails of a nucleosome are modified the gene in its

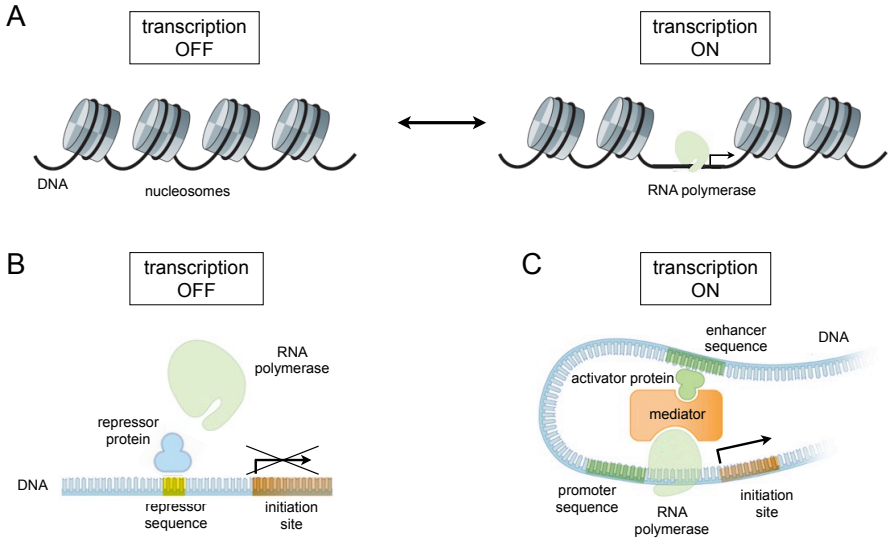


Figure 2.2: Transcriptional regulation of gene expression. (A) DNA accessibility is regulated: in order for RNA polymerase to bind on the DNA, the nucleosome around which the promoter sequence and the initiation site are wound have to be made accessible. If all nucleosomes are in place and the DNA is tightly wound around them the transcription is off (left). If the chromatin structure was locally remodeled, which can e.g. be an effect of histone modifications, RNA polymerase can bind and the transcription process starts (right; Figures modified from Voss and Hager (2014)). (B) Transcription initiation can also be regulated: in presence of repressor proteins that can bind to a repressor sequence on the DNA located close to the initiation site the RNA polymerase cannot bind and transcription is off (Figure in parts adopted from O'Connor and Adams (2010)). (C) Many genes also need the presence of activator proteins and a mediator in order for transcription to start. The activator proteins bind on an enhancer sequence in proximity of the promoter sequence where the RNA polymerase binds (Figure in parts adopted from O'Connor and Adams (2010)).

vicinity is either actively transcribed or not. Chromatin remodeling is necessary to loosen the DNA that is wrapped around the histone proteins and to expose it to the transcription machinery. It is important to note that all of these three processes can also take place synergistically, e.g. certain histone modifications may recruit chromosome remodeling complexes. These epigenetic mechanisms can be used by the cell to both, silence large regions of DNA such that their transcription is turned off, and provide dynamic access to DNA for transcription. Importantly, epigenetic modifications can also be inherited through cell replication.

Regulation of transcription initiation

In this subsection we discuss another way that is important for the establishment of cellular heterogeneity: the regulation of transcription initiation. In addition to the points discussed in the previous paragraph there are additional ways to regulate transcription even when the DNA is already accessible. On the one hand repressor proteins can bind to a DNA sequence in proximity of the transcription initiation site hindering RNA polymerase to bind and therefore repressing transcription (see Figure 2.2 B). On the other hand, there are several transcription factors necessary for the initiation of transcription. A set of general transcription factors is needed for the initiation of transcription of all genes but there may also be transcription factors that are specifically needed for the transcription of a certain gene. These activator proteins bind to an enhancer sequence that is located on the DNA in proximity of the initiation site and only in their presence it is possible to start transcription (see Figure 2.2 C).

Via this mechanism proteins can act as activators or repressors for the expression of genes (incl. their own gene) and regulate the presence of protein species in the cell. This concept is at the foundation of gene regulatory networks, which consist of several genes that activate or repress each other. Gene regulatory networks can – depending on their topology – act as switches or even perform logical operations (Alon, 2006). Moreover, they can become very large and complex; one example for this is the gene regulatory network of the human cell cycle where 4,449 genes are associated to be involved in (Amigo2, 2016; Carbon et al., 2009).

It is the regulation of DNA accessibility and the regulation of transcription initiation that we investigate as the source for cellular heterogeneity in this thesis. While we study ways to infer transcriptional regulation in Chapters 5 and 7, we examine general design principles of histone modifications that are involved in making the DNA accessible in Chapter 6.

2.2 Stochastic gene expression leads to variability

In contrast to cellular heterogeneity, which we argued to arise from differences in gene expression that are due to regulation, there is another process that can cause differences

in gene expression, namely stochastic gene expression leading to variability. For the remainder of this thesis we want to distinguish heterogeneity (caused by regulation) from variability (caused by stochastic gene expression). In the following we outline the origin of variability in more detail.

As outlined in the previous subsection, gene expression can be highly regulated. So far we adopted a biological point of view, where all tasks in the cell are performed by a highly orchestrated and specific interplay of molecular species. However, there is a fundamental limitation to this point of view that is given by the laws of thermodynamics. In fact all molecules in the cell are in constant random motion and have a kinetic energy that is proportional to the temperature of the cell (if all molecules in the cell were non-interacting, stiff particles they would have a velocity that is proportional to the square root of the temperature). As a consequence all molecules in the cell are randomly wobbling around and for biochemical reactions to happen the molecular species have to meet at the same place at the same time. This makes all processes in the cell – including gene expression – fundamentally stochastic.

While we provide a rigorous mathematical formulation of stochastic molecular dynamics in Chapter 3, we want to focus on the biological implications of stochasticity on gene expression, here. As discussed before regulation of gene expression is the source for the formation of different cellular phenotypes and leads to cellular heterogeneity. Stochastic gene expression, however, adds additional differences in gene expression and leads to variability in gene expression, even for cells that are in the same cellular phenotype (this is also known as intrinsic noise; Elowitz et al. (2002)). This variability is especially important for lowly abundant molecular species, since then the stochastic fluctuations in gene expression can have a substantial contribution to their abundance. mRNA species for instance have a median abundance of ~ 10 copies per cell. Small variations in gene expression have a relatively greater contribution on their abundance as compared to protein species, which have a median abundance of $\sim 10,000$ copies per cell (Schwannhäusser et al., 2011).

Therefore, variability in gene expression makes it necessary to describe the abundance of molecular species among cells from the same cellular phenotype with probability distributions (the mathematical framework with which this is possible is presented in

Chapter 3) rather than with a single value. In Section 3.4 we discuss that by investigating the variability of gene expression among cells of an identical cellular phenotypes it is even possible to draw conclusion on wether a gene is regulated or not.

2.3 Single cell and cell population based experiments

There are two possibilities to measure molecular species of cells: (i) It is possible to combine many cells and then extract and measure their cellular content (population based measurement). (ii) It is also possible to extract the cellular content of individual, single cells (single cell measurement). In this Section, we want to briefly discuss advantages and disadvantages of the two approaches.

Let us consider two different scenarios that shall make the advantage of single cell experiments over population based experiments apparent (Hoppe et al., 2014):

i We investigate biological processes in a single cellular phenotype. In this case, when performing many single cell experiment, we can see the variability that is due to stochastic gene expression (see Figure 2.3 A and Section 2.2). A population based measurement coincides with the average value of the many single cell experiments (see Figure 2.3 B) and does therefore also give valuable insight into the abundance of a molecular species of interest. This biological scenario is of importance when we analyze the underlying mechanism that leads to cellular heterogeneity.

ii We investigate a biological scenario where different cellular phenotypes are mixed together. In this case a single cell experiments provide the abundance of the molecular species for every single cell and is therefore suited to resolve the cellular heterogeneity (Figure 2.3 C). A population based experiment, however, coincides again with the average value and is therefore not capable to report about the underlying heterogeneity. Even worse, the measured value even may not be realized in neither of the different phenotypes (Figure 2.3 D). This biological scenario is of importance when we want to dissect cellular heterogeneity and we want to tell different cellular phenotypes apart.

The disadvantage that single cell experiments have as compared to population based measurements is that due to the lower abundance of cellular material in single cell

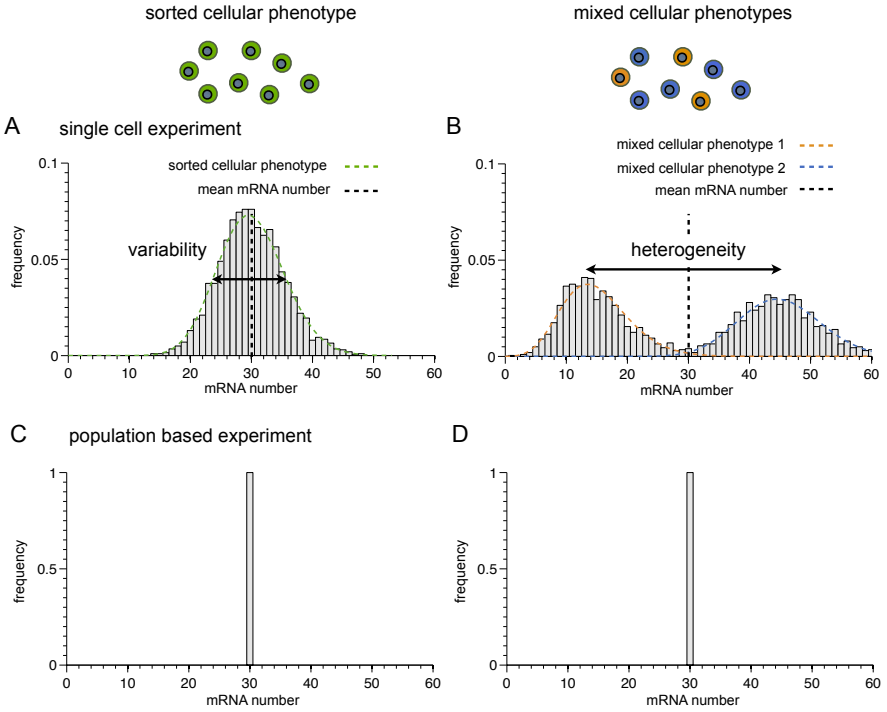


Figure 2.3: Comparison between single cell and population based experiments. (A) When performing single cell experiments (here mRNA measurement) on 2,000 cells of a sorted cellular phenotype (green cells) we obtain the steady-state distribution of mRNA transcript numbers. This is the normalized histogram of the measured transcript numbers of all cells. Since gene expression is intrinsically stochastic (See Chapter 2.3 and Chapter 4) the measured mRNA numbers display a variability. (B) Steady state-distribution obtained by performing single cell experiments on 2,000 cells that consist of a mixture of two different cellular phenotypes (1,000 orange cells and 1,000 blue cells). Here, the two cellular phenotypes express the measured mRNA differently. The single cell experiment can resolve the cellular heterogeneity between the two cellular phenotypes. (C) A population based experiment, where all 2,000 cells are put into one single sample and are then measured only returns one single value, namely the mean of the sample. In case of the sorted cellular phenotype the mean value of population based experiment is biologically meaningful since it corresponds to the mean value of the observed cellular phenotype. (D) A population based experiment for the mixed cellular phenotypes also returns the mean value of the whole sample. Here, however, the mean value is only barely realized in the two distinct cellular phenotypes that are mixed together. In case the sample exhibits cellular heterogeneity, a population based experiment is not capable to resolve the distinct cellular phenotypes and may even give misleading results.

experiments it is often hard to quantify the molecular species inside single cells and in case it is possible to obtain a signal, the level of technical noise is considerably higher than with population based experiments (see e.g. Brennecke et al. (2013)).

2.4 Cellular features

As discussed in the previous Section 2.1 there are multiple scales on which cellular processes take place (e.g. chromatin, mRNA, protein, morphology, etc.; see Figure 2.3). On each scale we can observe certain features (e.g. on the chromatin scale we can observe a set of chromatin modifications). In this Section we discuss the experimental techniques with which features from different scales can be extracted. It is important to note that in this thesis, we are interested in data from cellular processes that happen in living cells (*in vivo*) and not from biochemical processes that have been isolated and are analyzed apart from their cellular environment (*in vitro*).

In the following we want to provide a brief overview over current experimental procedures to access the different scales in the cell (Figure 2.4). We focus on summarizing the available measurement methods and discuss the details of the data properties that are of particular interest for the remainder of the thesis where we apply mathematical models on the data. We note that the experimental processes are a rapidly evolving field (see e.g. Konry et al. (2016) for a review).

DNA and chromatin

With the advent of next-generation sequencing in the 1990s (see Ansorge (2009) for a review) it has become possible to efficiently read out the genetic information contained in the DNA. Since in this thesis we want to focus on cells with the same DNA, we are not interested in genomic differences among organisms. Next generation sequencing, however, can also be used to investigate epigenetic modifications, such as DNA methylation (with DNA bisulfite sequencing; see Plongthongkum et al. (2014) for a review) and histone modifications (e.g. with chromatin immunoprecipitation sequencing (ChIP-seq) or with an assay for transpose-accessible chromatin using sequencing (ATAC-seq); see Meyer and Liu (2014) for a review). Another way to measure histone modifications

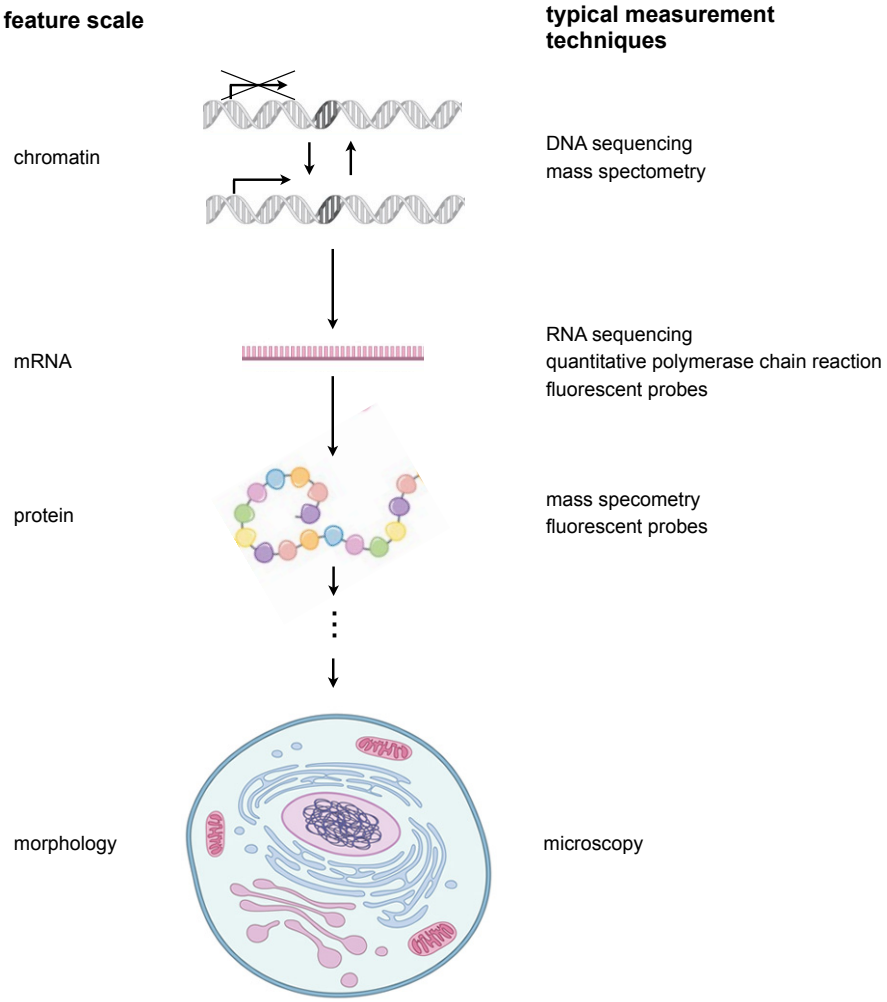


Figure 2.4: Accessible scales to extract cellular features relevant for this thesis. There are many different scales from which cellular features, can be extracted: chromatin, mRNA, proteins but also – on a more coarse scale – morphology. The features (depending on the scale they are extracted from) can correspond to the abundance of a set of chromatin modifications, mRNA or protein species, or to the value of morphological attributes (such as cell size). Typical measurement techniques to obtain features about the chromatin state are DNA sequencing and mass spectrometry. mRNAs can be measured with RNA-sequencing (RNA-seq), quantitative polymerase chain reaction (qPCR) or fluorescent probes, such as fluorescent in situ hybridization (FiSH). The protein scale can be measured with mass spectrometry and also with fluorescent probes. The morphology of a cell is typically measured with microscopes (parts of the Figure are adopted from Clancy (2008) and O'Connor and Adams (2010)).

is given by mass-spectrometry where the histone tails that contain the modifications are disrupted into peptides, which can then be detected with a mass-spectrometer (see e.g. Huang et al. (2015) for a review).

mRNAs

The abundance of mRNA species can be measured with RNA-sequencing (RNA-seq; see Ozsolak and Milos (2011) for a review). There, the mRNAs are inversely transcribed into cDNA (copy DNA), which can then (after multiple steps of amplification) be read out with next-generation sequencing. Another way to measure mRNA is given by quantitative polymerase chain reaction (qPCR; see e.g VanGulder et al. (2008)) where the mRNAs are also inversely transcribed and then subsequently doubled until a fluorescent signal can be detected that appears once a certain threshold of mRNA copies of a species are present. Yet another way to measure the abundance of mRNA is given by fluorescent in situ hybridization (FiSH; see e.g. Levsky and Singer (2003)), where individual mRNAs are tagged with fluorescent probes that can then be optically detected.

Proteins

There are different ways to measure the abundance of proteins in cells. One way is by the aforementioned mass-spectrometry where proteins are broken down into peptides which can then be separated based on their mass (see e.g. Walther and Mann (2010)). Moreover, it is possible to tag proteins with antibodies that are attached different isotopic forms of atoms, which naturally do not occur in the cell and then to perform mass-spectrometry, a method called mass cytometry (see e.g. Bandura et al. (2009)). Another common way to measure protein abundances are antibodies that are attached to fluorophores – the more proteins there are the more antibodies can bind and the higher is the fluorescence intensity of the cell. This approach can be combined with subsequent cell sorting by a method called fluorescence activated cell sorting (see Herzenberg et al. (2002) for a review).

low: 1-100; med: 100-1,000; high: > 1,000

measurement technique	scale	number of features	throughput [number of samples per run]	live possible
DNA sequencing	chromatin	high	med	no
mass spectrometry	chromatin protein	med	low	no
mass cytometry	protein	med	high	no
RNA sequencing	mRNA	high	high	no
quantitative polymerase chain reaction	mRNA	low	med	no
fluorescent probes	mRNA protein	low	high	yes
microscopy	morphology	med	high	yes

Figure 2.5: Overview of data properties for selected experimental techniques. Important properties of measurement techniques contain: the cellular scale from which they can report information, the number of features they are capable to extract, the number of samples they can measure in one run, their capability to perform live cell analysis and their capability to resolve local information about the measured biochemical processes.

Morphology

Finally the morphology of cells can be measured. It is important to note that differences in morphology are usually a consequence of differences in gene expression and are therefore only a downstream effect. Nevertheless, it is often beneficial to obtain images of the cells before other experiments are performed or to focus entirely on imaging. For this purpose, standard microscopy techniques to image live cells have been used for decades (Stephens and Allan, 2003). A recent experimental technique is light sheet microscopy (Keller et al., 2008) that enables the imaging of a whole organisms. In this thesis we use data from another recent experimental technique, imaging flow cytometry (IFC; Basiji et al. (2007)), which delivers robust images of thousands of cells with $40 - 60\times$ magnification. It is possible to obtain information rich features from images by using imaging software tools that quantifies the morphological shape of the cell (Eliceiri et al., 2012).

Data limitations

The experiments with which features from cells can be extracted, however display certain limitations that affect the properties of the obtained data (see Figure 2.5):

i Incompleteness. The features that can be measured from the cell are usually not fully comprehensive, i.e. we cannot observe all properties of the cell that are of interest for an investigated cellular process. Mostly only information about one scale can be obtained.

ii Invasivity. The experiments are often invasive and often involve killing the cell. To access the molecular content in the cell, cells have to be disrupted and the cellular content of interest has to separated with biochemical techniques. As a consequence it is often challenging to perform live cell analysis and the data is often of a so called snapshot nature, which means that the information can only be extracted at one time point, as compared to time-series where we have data from the same cell over an extended period of time.

iii Technical noise. Technical noise directly stems from the way the experiment is performed and it therefore can be unique to the kind of experiment performed. It is a challenge for mathematics and computational biology to find proper ways to take technical noise into account and to normalize the data (we will come back to this in Chapter 6).

2.5 Biological model systems with cellular heterogeneity

We want to conclude this Chapter with the discussion of two biological systems that display cellular heterogeneity: hematopoiesis (the formation of adult blood cells) and cell cycle. While there is already a lot of knowledge about the principle mechanisms of gene expression, its regulation and the cellular phenotypes it leads to, many details of how genes interact and shape cellular phenotypes is yet incomplete. These two systems are already well-studied and are also of interest for the remainder of this thesis.

Hematopoiesis

Hematopoiesis (see Figure 2.6 A), the formation of mature blood cells, serves as a paradigm to study stem cell development (Orkin and Zon, 2008). At the top of its hierarchy is the hematopoietic stem cell (HSC), which is pluripotent (i.e. it can give rise to all mature blood cells) and has the ability to self-renew (i.e. it can remain its pluripotent state). During hematopoiesis a HSC subsequently differentiates into progenitor cells (corresponding to different cellular phenotypes) that are more and more restrictive in their differentiation potential. This process is accompanied by subsequent changes in the protein species that are present in the cells. Thus each progenitor can be categorized by a set of marker proteins that are only present in this particular cellular phenotype. E.g. the common lymphoid progenitor (CLP) gives rise to all the lymphatic cells and the common myeloid progenitor (CMP) gives rise to the other white and the red blood cells, which both have specific marker genes (see e.g. Rieger and Schroeder (2012) for a review on murine hematopoiesis).

In order to better understand the differentiation process it is therefore necessary to study the dynamical changes of protein species. Since malfunction of hematopoiesis is often related to diseases (Whichard et al., 2010) more knowledge about the involved individual processes may increase our ability to find therapeutic targets. Two protein species of particular interest are PU.1 and Gata-1, which were reported to be responsible for the differentiation of the CMP to either the granulocyte-monocyte progenitor (GMP) or the megacaryocyte-erythrocyte progenitor (MEP; Arinobu et al. (2007); Burda et al. (2013)). The common hypothesis that gene regulation between the two protein species is mutually inhibiting and self-activating (see e.g. (Duff et al., 2012) for a review) has been challenged by recent measurements (Hoppe et al., 2016). Thus there is a need for new mathematical methods to find general principles underlying the interaction of the two protein species. In this thesis we therefore investigate the information transfer between these two proteins based on the data measured by Hoppe et al. (2016) (see Chapter 7). Moreover, we use independent mRNA data to check the transcriptional regulation of the two proteins in MEPs (see Chapter 5).

Further recent findings also shine light on the role of chromatin modifications during hematopoiesis (Lara-Astiasio et al., 2014). The quest for answering how chromatin

modifications contribute to differentiation is still at the beginning and many details are lacking (Signolet and Hendrich, 2015). Although we do not analyze chromatin data from hematopoiesis in this thesis, we investigate the general design principles of chromatin modifications to gain insights in their establishment (see Chapter 6).

The cell cycle

Another well-studied process that displays cellular heterogeneity is the cell cycle. Cell cycle is the process cells undergo in order to duplicate their DNA and to divide into two daughter cells. The cell cycle can be divided into 4 major phases (i.e. cellular phenotypes; see Figure 2.6 B). It starts with a first gap phase (G1), in which the cells prepare for the replication of their DNA. It is followed by the DNA synthesis phase (S), in which the DNA then is replicated. Next, there is a second gap phase (G2), where the cell prepares for cell division. The last cell cycle phase, mitosis (M), can be further divided into prophase where chromatin condenses into rod-shaped chromosomes (containing the original and the replicated chromatin fibers), into metaphase where the chromosomes align, into anaphase where the chromosomes split apart and into telophase where the two daughter cells, each with one set of chromosomes divide.

The cell cycle is an important biological process since it governs the replication of all proliferating cells (Nurse, 2000). Investigating the regulatory mechanisms that control the cell cycle therefore leads not only to new insights into a very fundamental process but also helps to understand how diseases that are due to the malfunction of the cell cycle develop (Vermeulen et al., 2003). Moreover, cancer cells are often highly proliferating and arresting them in one of the cell cycle phases offers a prominent therapeutic target (Dickson and Schwartz, 2009). A promising approach is the arrest in the M phase since a prolonged stay in this cell cycle phase is likely to induce cell death (Chan et al., 2012). There, one of the challenges is to identify cells that are in the M phase, which are typically much less frequently occurring than the other cell cycle phases (Filby et al., 2011), in order to closer investigate their response to cell-cycle arrest treatments. In Chapter 9 we present a novel method to identify cell cycle phases that may be of help for future applications.

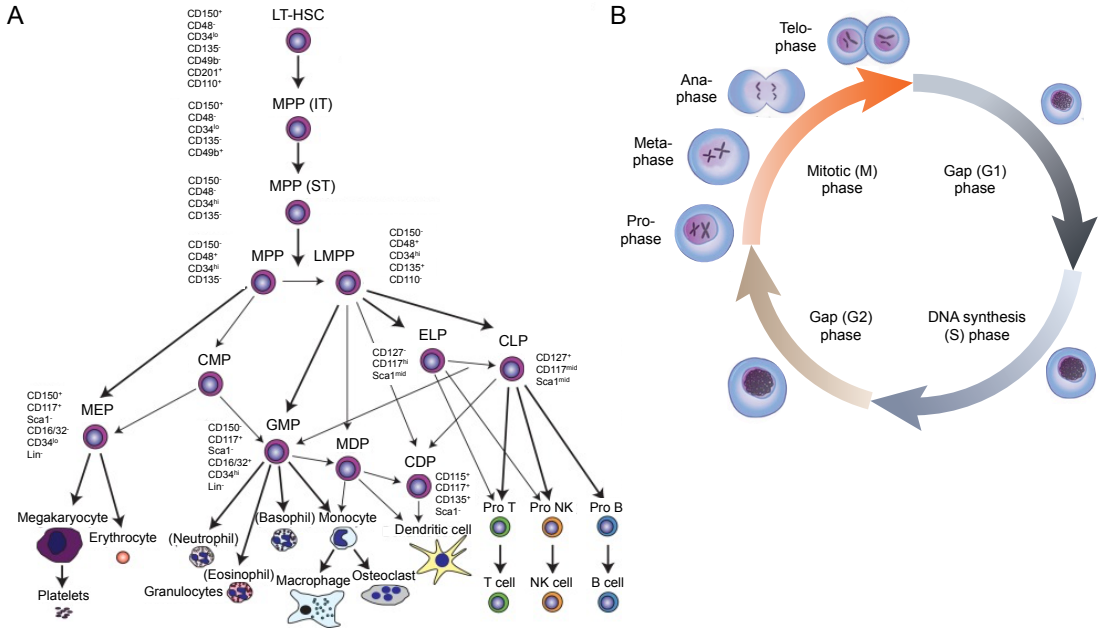


Figure 2.6: Cellular heterogeneity in hematopoiesis and the cell cycle. (A) Hematopoiesis is the formation of mature blood cells. At the top of the hierarchy (depicted here the formation of murine blood cells) is the long-term hematopoietic stem cell (LT-HSC), which differentiate subsequently into intermediate-term and short-term multipotent progenitors (MPP (IT) and MPP (ST)). The multi-potent progenitors split up into a myeloid (common myeloid progenitor (CMP)), which gives rise to all mature red (erythrocytes and megacaryocytes) and white (granulocytes, macrophages, monocytes and dendritic killer cells) blood cells and into a lymphoid linage (common lymphoid progenitor (CLP)), which gives rise to the cells of the immune system (T cells, nature killer cells and B cells. For every cellular phenotype there is a set of protein species on the outer membrane of the cell (listed at the side) that is specific for it, which can be marked with fluorescently tagged antibodies (Figure adopted from Rieger and Schroeder (2012)). (B) Many cells undergo the cell cycle, which is the process of cell division involving the duplication of DNA. The cell cycle starts in the first gap phase (G1), where the cell prepares for DNA replication. The next cell cycle phase is the DNA synthesis phase (S) where the DNA is replicated. It is followed by a second gap phase (G2), in which the cell prepares for division. The last cell cycle phase is mitosis (M), which is further subdivided into prophase, metaphase, anaphase and telophase. In the mitotic phase the cell divides into its two daughter cells (parts of the Figure are adopted from O'Connor and Adams (2010)).

Chapter 3

Stochastic and deterministic modeling of gene expression

Essentially, all models are wrong,
but some are useful.

George E. P. Box [III]

So far the molecular processes discussed in the previous Chapter have been presented from a biologists' point of view in which the involved molecular species are thought to act in a well-defined, deterministic and clearly specific way. In this Chapter we discuss the limitations of this point of view, which are based on the laws of thermodynamics. We describe how gene expression, just as every other chemical reaction is intrinsically stochastic. As a consequence even two hypothetically identical cells display variability in the expression of the same gene. It is important to note that these differences in gene expression do not lead to cellular heterogeneity in the sense introduced in the previous Chapter but to variability among cells that have an identical cellular phenotype.

In Section 3.1 we introduce a probabilistic mathematical framework that is suitable to describe chemical reactions of an idealized system of molecules: the chemical master equation (CME). We show how individual realizations of biochemical processes that obey the CME can efficiently be simulated with the stochastic simulation algorithm (SSA; Section 3.2). In Section 4.3, we then show how the discrete and stochastic CME can be approximated and leads to the chemical Langevin equation and to the reaction

rate equations (RRE). For these Sections we follow the presentation of Gillespie (1992) and Gillespie (2007) and added further references (see also Figure 3.1). In Section 3.4 we demonstrate the implications of stochasticity on biological processes. We simulate a toy system where one gene is regulated by another and use this simplistic gene regulatory network to derive analytic formulas for the steady-state distributions of the variability of gene expression among cells of the same cellular phenotype. Eventually, we discuss the applicability of the CME to biological processes in cells, which do not necessarily obey all the assumptions made to derive the CME.

3.1 The chemical master equation

As outlined in Section 2.1 there are many molecular species involved in gene expression and even the products of gene expression, functional proteins are often present in a vast abundance. Therefore it is an almost infeasible task to keep track of all positions and velocities of all the molecular species involved in biochemical processes. A full simulation of all involved species is computationally simply infeasible. However, based only on a few assumptions about the status of the molecular species, it is possible to derive a suitable mathematical framework to efficiently describe biological processes. In this Section we follow Gillespie (1992) and Gillespie (2007) and present a rigorous derivation of the chemical master equation (CME), which can be used as a probabilistic formulation of (bio-)chemical processes of interest. It is important to note that the CME is a formulation for an idealized system and although it is widely applied to biochemical processes in cells by the community (see e.g. Liang and Qian (2010)) as well as in this thesis, the assumptions that are made for its derivation do not necessarily hold in the context of cells (an issue that we discuss at the end of Section 3.4). The general idea behind the derivation is to:

i Make assumptions about the distribution of the positions and velocities of all species in the cell in order to reach a framework that describes the state of the system based only on the number of molecular species that are involved in a cellular process of interest.

ii Calculate the probability for a particular reaction to happen given the current state of the system.

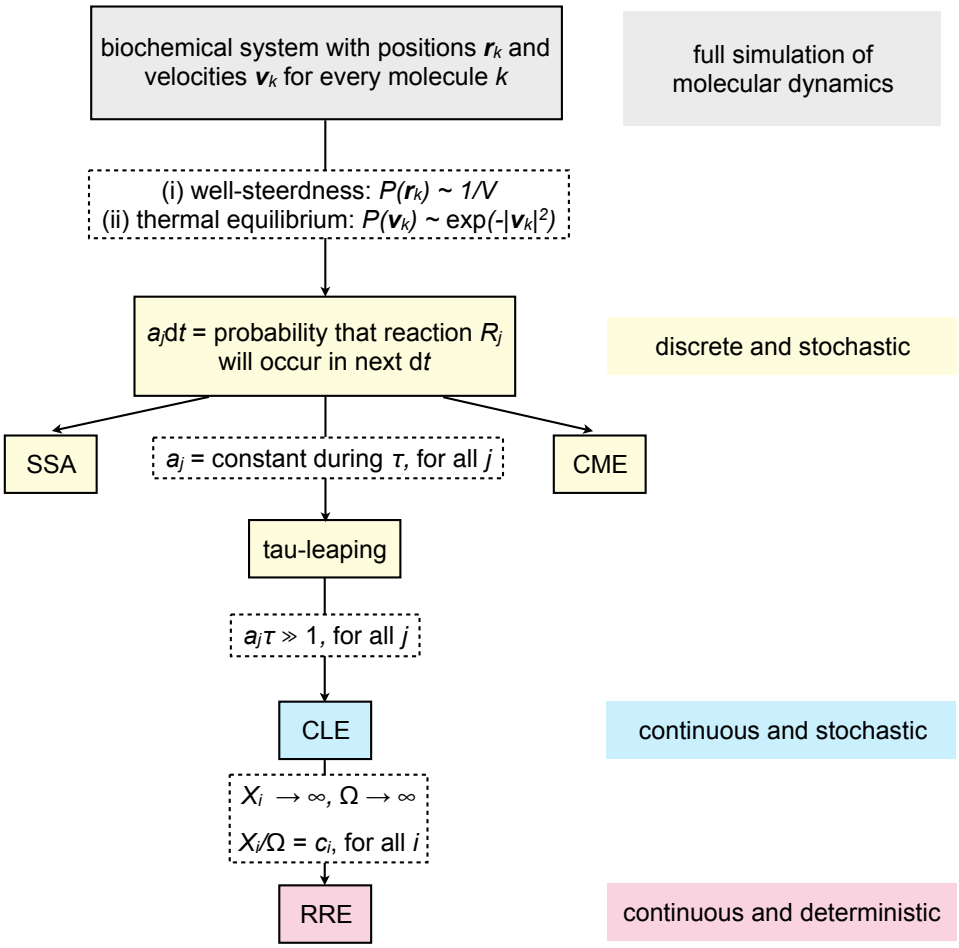


Figure 3.1: Mathematical descriptions of chemical reactions. With assumptions about the positions and velocities of the molecules in a chemical system it is possible to describe the dynamics of the system with propensities $a_j(\mathbf{x})$ that only depend on the current abundance of all species $\mathbf{x}$ instead of simulating the full molecular dynamics. In this case, it can be shown that the probability that reaction R_j will occur in the next infinitesimal time interval dt is proportional to the propensity $a_j(\mathbf{x})$. This is the starting point for the derivation of the chemical master equation (CME; see Section 3.1). While the CME can hardly be solved analytically, individual realizations that obey a given CME can be simulated efficiently with the stochastic simulation algorithm (SSA; see Section 3.2). Both the SSA and the CME describe discrete and stochastic chemical processes. In Section 3.3 we show under which limits we obtain a continuous and stochastic description with the chemical Langevin equation (CLE) and finally a continuous and deterministic description with the reaction rate equation (RRE).

iii Derive a time-evolution equation for the molecular species in the system.

We start with a cellular subsystem of interest (e.g. regulation of gene expression) that is restricted to a particular reaction volume Ω consisting of N biochemical species $\{S_1, \ldots, S_N\}$, which interact through M chemical reactions $\{R_1, \ldots, R_M\}$. $X_i(t)$ indicates the number of molecules of the species S_i and $\mathbf{X}(t) = \{X_1(t), \ldots, X_N(t)\}$ denotes the state vector containing all the molecular species present in the system at a given time t. Every reaction is accompanied by a state-change vector $\boldsymbol{\nu}_j = (\nu_{1j}, \ldots, \nu_{Nj})$ where ν_{ij} is the change in the number of molecules of species S_i due to one R_j reaction, e.g., R_j leads the system from state $\mathbf{X}(t)$ to $\mathbf{X}(t) + \boldsymbol{\nu}_j$ (note that the number of molecules can change through a chemical reaction) Moreover, each molecule has a position $\mathbf{r}_k$ and a velocity $\mathbf{v}_k$. Our goal is to calculate the state vector $\mathbf{X}(t)$ when we know that the system was in state $\mathbf{X}(t_0)$ at the initial time t_0.

We now make certain assumptions about the distributions of the positions and the velocities of the system:

I The system is well-steered: the probability that the position of a randomly selected molecule lies inside a small volume $\Delta\Omega$ of the cell is equal to $\Delta\Omega/\Omega$.

II The system is at thermal equilibrium: the probability that the velocity of a randomly selected molecule is given by the Maxwell-Boltzmann distribution (i.e. every velocity component of a molecule is independent from each other and normally distributed with mean 0 and variance that scales with the temperature of the system).

Note that it is implicitly assumed that the positions and the velocities are statistically independent from each other.

The key for the derivation of the CME is the reaction probability $P(R_i|\mathbf{x})$, given $\mathbf{X}(t) = \mathbf{x}$, that a reaction R_j will occur in the next infinitesimal time interval $[t, t + \mathrm{d}t)$. Given assumptions (I) and (II) one can show that this probability does not depend on neither the species' positions, nor their velocities (Gillespie, 1992)

$$P(R_i|\mathbf{x}) = a_j(\mathbf{x})\mathrm{d}t \tag{3.1}$$

and it is proportional to the propensity function $a_j(\mathbf{x})$. For R_j being a unimolecular reaction ($S_1 \rightarrow$ product(s)) the propensity function is given by $a_j(\mathbf{x}) = c_j^{\mathrm{uni}} x_1$, whereas for R_j being a bimolecular reaction ($S_1 + S_2 \rightarrow$ product(s)) it is given by $a_j(\mathbf{x}) =$

$c_j^{\text{bi}} x_1 x_2$. Here c_j^{uni} and c_j^{bi} are known as reaction rate constants. The fact that the reaction probability $P(R_j|\mathbf{x})$ scales with the number of present species is also known as mass action kinetics. Given the specific form of assumptions (I) and (II) it can be shown (Gillespie, 1992) that in case of unimolecular reactions, the c_j^{uni}'s are independent of the volume, but in case of bimolecular reactions, the c_j^{bi} are inversely proportional to the volume. Furthermore, both $c1_j^{\text{uni}}$ and c_j^{bi} scale with the square root of the system's temperature (i.e. the higher the temperature the faster the dynamics in the system).

By the help of Eq. 3.1 we can calculate a time-evolution equation for $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$, i.e. the probability for finding the system in the state $\mathbf{x}$ at time t, given the system in the initial state $\mathbf{x_0}$ at time t_0:

$$P(\mathbf{x}, t+\mathrm{d}t)|\mathbf{x}_0, t_0) = P(\mathbf{x}, t|\mathbf{x}_0, t_0) \times \left(1 - \sum_{j=1}^{M} a_j(\mathbf{x})\mathrm{d}t + o(\mathrm{d}t)\right) + \sum_{j=1}^{M} P(\mathbf{x} - \boldsymbol{\nu}_j, t|\mathbf{x}_0, t_0) \times (a_j(\mathbf{x} - \boldsymbol{\nu}_j \mathrm{d}t + o(\mathrm{d}t)) + o(\mathrm{d}t). \quad (3.2)$$

The first term on the right hand side of Eq. 3.2 corresponds to the probability for no reaction to happen within the time $\mathrm{d}t$ (i.e. the system remains in the state $\mathbf{x}$) and the second term corresponds to the probability for exactly one reaction R_j to happen that drives the system to $\mathbf{x}$ (i.e. the system has to be in the state $\mathbf{x} - \boldsymbol{\nu}_j$ before). Any reaction that drives the system into the state $\mathbf{x}$ with more than one reaction has a probability that for $\mathrm{d}t \to 0$ goes to zero faster than $\mathrm{d}t$ (indicated by $o(\mathrm{d}t)$).

Subtracting $P(\mathbf{x}, t|\mathbf{x}_0, t)$ from both sides, dividing through $\mathrm{d}t$ and taking the limit $\mathrm{d}t \to 0$ we obtain the chemical master equation (CME):

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x_0}, t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(\mathbf{x} - \boldsymbol{\nu}_j) P(\mathbf{x} - \boldsymbol{\nu}_j, t|\mathbf{x_0}, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t|\mathbf{x_0}, t_0)]. \quad (3.3)$$

Note that the CME is a discrete and stochastic differential equation that describes how a biochemical system that starts in an initial state $P(\mathbf{x}, t_0)$ propagates in time. At every point in time the system is described by a probability distribution for $\mathbf{x}$ rather than by a deterministic value and at a given time point t individual realizations of the CME display variability with respect to their state $\mathbf{x}$ (see also Section 2.2).

The CME provides a mathematical foundation to calculate biochemical processes. Although it can only be solved analytically in rare cases (Jahnke and Huisinga, 2007) that are often too simplistic for real biochemical systems of interest, there are many ways to find approximate solutions. One way to obtain approximate solutions for the CME is to solve the time-evolution for its moments (Engblon, 2006). Another way is given by the finite state projection (Munsky and Khammsh, 2006) where the number of species in the system is restricted to a finite upper value. This upper value is limited because of computational reasons: the dimensions of the matrix that results from the finite state projection scales with the number of molecules in the system, which becomes too large to computationally keep track of very fast. Furthermore the CME can be expanded in a Taylor series for the systems volume Ω, known as system size expansion (Van Kampen, 1997) that leads to the linear noise approximation (Elf and Ehrenberg, 2003) when only the first two terms in the Taylor series are considered.

3.2 The stochastic simulation algorithm

In the previous Section we derive the CME and point out that it is hard to find analytical solutions for it. Often, however, it is not necessary to solve the CME, but it is sufficient to simulate a biochemical process of interest. In this Section we follow Gillespie (2007) and show that it is possible to efficiently simulate realizations of a biochemical process that obeys the CME.

Simulating trajectories of $\mathbf{X}(t)$ is possible by considering the probability that, given the current state of the system $\mathbf{X}(t) = \mathbf{x}$ at time t, a reaction will occur in time $\tau > t$, and that it will be the reaction R_j:

$$p(\tau, j|\mathbf{x}, t) = a_0(\mathbf{x}) \exp(-a_0(\mathbf{x})\tau) \cdot \frac{a_j(\mathbf{x})}{a_0(\mathbf{x})}, \tag{3.4}$$

where $a_0(\mathbf{x}) = \sum_{j'=1}^{M} a_{j'}(\mathbf{x})$ is the sum over all propensity functions. This probability is obtained by the product of an exponential distribution (with parameter $a_0(\mathbf{x})$) for the time τ, and a weighting factor $a_j(\mathbf{x})/a_0(\mathbf{x})$ for the reaction being of type j.

Here, we present the 'direct method' to generate realizations that obey the CME (Gillespie, 2007):

0. Fix the initial time $t = t_0$ and the initial state of the system $\mathbf{x} = \mathbf{x_0}$.

1. Evaluate all propensities $a_j(\mathbf{x})$ and calculate their sum $a_0(\mathbf{x})$.

2. Generate two random numbers r_1 and r_2 from the uniform distribution that lie in the interval $[0, 1]$.

3. Given the random numbers r_1 and r_2 calculate the time τ for the next reaction to happen and the integer j that indicates what R_j reaction it is

$$\tau = \frac{1}{a_0(\mathbf{x})} \log\left(\frac{1}{r_1}\right), \tag{3.5}$$

$$j = \arg\min_{j \in \mathbb{N}} \sum_{j'=1}^{j} a_{j'}(\mathbf{x}) > r_2 a_0(\mathbf{x}). \tag{3.6}$$

4. Perform the next reaction by updating $t \leftarrow t + \tau$ and $\mathbf{x} \leftarrow \mathbf{x} + \boldsymbol{\nu}_j$

5. Save $(\mathbf{x}, t)$ and return to Step 1 or end the simulation if sufficient reactions or time points have been simulated.

We give an example for individual realizations for the CME of a biological toy system in Section 3.4.

3.3 From discrete and stochastic to continuous and deterministic models

In this Section we want to show under which limits the CME can be approximated by continuous and/or deterministic instead of a discrete and stochastic equations. As outlined in Section 2.2 stochastic gene expression is mostly relevant for lowly expressed molecular species (e.g. mRNA species can be present with a median of ~ 10 copies per cell; see Section 3.4). This Section provides a clear mathematical formulation for what 'lowly expressed' means quantitatively.

The chemical Langevin equation

We start by approximating the CME with the chemical Langevin equation (CLE). To this end, we make the following assumption, known as the leap condition. We suppose

there exists a $\tau > 0$, for which

$$a_j(\mathbf{x}) = \text{ constant during } [t, t+\tau], \text{ for all } j \tag{3.7}$$

holds. Then the number of reaction R_j to happen in the time interval $[t, t+\tau]$ is given as a Poisson distribution $P(j|\mathbf{x}, t, \tau) = \text{Pois}_j(a_j(\mathbf{x})\tau)$ and we can leap the system from time t to $t+\tau$ by

$$\mathbf{X}(t+\tau) = \mathbf{x} + \sum_{j=1}^{M} \text{Poisrnd}_j(a_j(\mathbf{x})\tau)\boldsymbol{\nu}_j, \tag{3.8}$$

where $\text{Poisrnd}_j(a_j(\mathbf{x})\tau)$ are statistically independent random numbers drawn from the Poisson distribution with mean $a_j(\mathbf{x})\tau$ (note that for the Poisson distribution the mean equals the variance). Eq. 3.8 is known as the tau-leaping formula.

Additionally to the leap condition, Eq. 3.7, we assume

$$a_j(\mathbf{x})\tau \gg 1, \text{ for all } j, \tag{3.9}$$

i.e. the product of all propensities with the time τ during which we assume the propensities to be constant is supposed to be large. Then we can make use of the fact that a Poisson distributed random variable with a large mean value ($\gg 1$) can be approximated as a normal distributed random variable with the same mean and variance, giving the Langevin leaping formula

$$\begin{aligned}\mathbf{X}(t+\tau) &= \mathbf{x} + \sum_{j=1}^{M} \text{Normrnd}_j(a_j(\mathbf{x})\tau, a_j(\mathbf{x})\tau)\boldsymbol{\nu}_j \\ &= \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{x})\tau + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{x})} \text{Normrnd}_j(0,1)\sqrt{\tau},\end{aligned} \tag{3.10}$$

where $\text{Normrnd}_j(\mu, \sigma^2)$ denotes statistically independent random numbers drawn from the normal distribution with mean μ and variance σ^2. Note that by approximating the Poisson random variables with normal random variables the Langevin leaping formula now describes the state of the system $\mathbf{X}(t)$ in a continuous way.

It can further be shown (Gillespie, 2000) that, in the limit of $\tau \to 0$, Eq. 3.10 leads to the chemical Langevin equation (CLE)

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}(t)) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \sqrt{a_j(\mathbf{X}(t))}\Gamma_j(t), \tag{3.11}$$

which is a continuous, stochastic differential equation for the time-evolution of the system's state $\mathbf{X}(t)$. Here $\Gamma_j(t)$ are statistically independent and temporally uncorrelated Gaussian white processes defined by $\Gamma_j(t) = \lim_{\tau \to 0} \mathcal{N}(0, 1/\tau)$.

The CLE, Eq. 3.11, offers a nice way to interpret biochemical processes. The first term on the right hand side corresponds to a deterministic part of the time-evolution whereas the second term incorporates the stochastic part. If the stochastic part of the CLE was neglected (we will see in the next Section under what conditions this can be done) the state of the system $\mathbf{X}(t)$ could be described purely by an ordinary differential equation (ODE) that is governed by the topology of the reactions and the time-evolution would deterministically follow the solution of the ODE. The second term, however, adds stochastic fluctuations that let the state of the system fluctuate around the ODE time-evolution (Huang, 2009).

The reaction rate equation

Finally the CLE can be approximated by a deterministic equation when

$$X_i \to \infty \,\text{and}\, \Omega \to \infty \,\text{such that}\, X_i/\Omega = \text{const}, \tag{3.12}$$

which is known as the thermodynamic limit and is given by the system's size as well as the species becoming very large, but in a way that the concentration of the species X_i remains constant.

To assess the implications of this approximation, we need to remember that both the propensity functions for unimolecular and bimolecular reactions scale with the system size (see Section 3.1). In the thermodynamic limit, Eq. 3.12, the second term on the right hand side of the CLE, Eq. 3.11, becomes negligibly small compared to the deterministic part and can be neglected, leaving us with the reaction rate equation (RRE)

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \sum_{j=1}^{M} \boldsymbol{\nu}_j a_j(\mathbf{X}(t)), \tag{3.13}$$

which is a deterministic and continuous formulation for biochemical processes. Note that only for unimolecular reactions the mean of the CME and the RRE coincide.

3.4 Application to biological processes

In this Section we begin by formulating the CME for a simple gene regulatory network where a constantly expressed gene represses the expression of another gene and simulate its gene expression using the SSA. We then focus on the steady-state limit of the CME, in which analytical solutions for the CME can obtained for both regulated and unregulated genes. We conclude this Section with considerations on the applicability of the CME to cellular processes where we discuss to what extend the assumptions made for the derivation of the CME are biologically reasonable.

Simulation of a simple toy example of regulated gene expression

In general we could simulate every individual step of transcription and translation, which we described in Section 2.1, with the CME. While both processes by itself are very complex and it currently proves to be infeasible to observe all the molecular species involved in gene expression *in vivo*, it is possible to measure the abundance of mRNA and protein species in a cell (see Section 2.3). Since the aim of this thesis is to infer cellular properties in a data-driven way (i.e. we need data to compare our model with), we treat, for the remainder of this thesis, transcription and translation, each as one single process neglecting their molecular sub-steps.

Let us now consider a concrete biological process, namely a gene regulatory network, where a gene A represses another gene B (see Figure 3.2 A and B). Gene A is constantly transcribed into mRNA A with a rate constant $\nu_{\text{mRNA,A}}$ and subsequently translated into protein A with a rate constant $\nu_{\text{protein,A}}$. The transcription initiation of gene B is repressed by protein A, which binds and unbinds on the repression site of gene B with rate constants k_{off} and k_{on}, respectively. When the repression site of gene B is not bound by protein A, gene B is transcribed into mRNA B with rate constant $\nu_{\text{mRNA,B}}$, which then is translated into protein B with rate constant $\nu_{\text{protein,B}}$. Once the mRNA and proteins A and B are produced they can degrade with rate constants $\gamma_{\text{mRNA,A}}$, $\gamma_{\text{protein,A}}$, $\gamma_{\text{mRNA,B}}$ and $\gamma_{\text{protein,B}}$, respectively. The current state of this system at time t is described by the vector $\mathbf{X}(t) = (x_{\text{mRNA,A}}, x_{\text{protein,A}}, x_{\text{DNA,B,off}}, x_{\text{DNA,B,on}}, x_{\text{mRNA,B}}, x_{\text{protein,B}})$ and obeys the CME, Eq. 3.3, which in this case reads

$$
\begin{aligned}
\frac{\partial P(\mathbf{x}, t|\mathbf{x_0}, t_0)}{\partial t} =& \nu_{\mathrm{mRNA,A}} P(x_{\mathrm{mRNA,A}} - 1, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& - \nu_{\mathrm{mRNA,A}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \nu_{\mathrm{protein,A}} x_{\mathrm{mRNA,A}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}} - 1, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& - \nu_{\mathrm{protein,A}} x_{\mathrm{mRNA,A}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \gamma_{\mathrm{mRNA,A}} (x_{\mathrm{mRNA,A}} + 1) P(x_{\mathrm{mRNA,A}} + 1, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& - \gamma_{\mathrm{mRNA,A}} x_{\mathrm{mRNA,A}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \gamma_{\mathrm{protein,A}} (x_{\mathrm{protein,A}} + 1) P(x_{\mathrm{mRNA,A}} + 1, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& - \gamma_{\mathrm{protein,A}} x_{\mathrm{protein,A}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + k_{\mathrm{off}} x_{\mathrm{DNA,B,on}} (x_{\mathrm{protein,A}} + 1) P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}} + 1, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& - k_{\mathrm{off}} x_{\mathrm{DNA,B,on}} x_{\mathrm{protein,B}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + k_{\mathrm{on}} x_{\mathrm{DNA,B,off}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}} - 1, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& - k_{\mathrm{on}} x_{\mathrm{DNA,B,off}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \nu_{\mathrm{mRNA,B}} x_{\mathrm{DNA,B,on}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}} - 1, x_{\mathrm{protein,B}}) \\
& - \nu_{\mathrm{mRNA,B}} x_{\mathrm{DNA,B,on}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \nu_{\mathrm{protein,B}} (x_{\mathrm{mRNA,B}}) P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}} - 1) \\
& - \nu_{\mathrm{protein,B}} x_{\mathrm{mRNA,B}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \gamma_{\mathrm{mRNA,B}} (x_{\mathrm{mRNA,A}} + 1) P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}} + 1, x_{\mathrm{protein,B}}) \\
& - \gamma_{\mathrm{mRNA,B}} x_{\mathrm{mRNA,B}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \\
& + \gamma_{\mathrm{protein,B}} (x_{\mathrm{protein,A}} + 1) P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}} + 1) \\
& - \gamma_{\mathrm{protein,B}} x_{\mathrm{protein,B}} P(x_{\mathrm{mRNA,A}}, x_{\mathrm{protein,A}}, x_{\mathrm{DNA,B,off}}, x_{\mathrm{DNA,B,on}}, x_{\mathrm{mRNA,B}}, x_{\mathrm{protein,B}}) \qquad (3.14)
\end{aligned}
$$

Before we can simulate this system with the SSA, we need to specify the reactions' rate constants for transcription, translation as well as mRNA and protein degradation. To this end, we adopt typical values for mammalian cells from Schwannhäusser et al. (2011) who analyzed the genome-wide abundance of all mRNA and protein species in mouse fibroblast cells. They report that there are $\sim$ 6,000 different species of proteins and their corresponding mRNAs present with a median abundance of 16,000 proteins and 17 mRNAs per species. Moreover, they measured the median half-life times among all protein species to be 46 h and the median half-life time among all mRNAs to be 9 h. Given the abundance and the half-life time of all the mRNA and protein species they determined a median transcription rate constant to be 2 mRNA transcripts per hour and the median translation rate constant to be 40 proteins per mRNA transcript per hour. The degradation rate constants can be obtained as the inverse half-life times. We chose the reaction rate constant for the binding and unbinding of protein B, such that transcription can remain in the on state longer than the typical time that it takes for one mRNA transcript to be produced.

For the simulation of the described gene regulatory network displayed in Figure 3.2 we used the SSA as implemented by the StochKit2 toolbox (Sanft et al., 2011) with the following reaction rate constants: $\nu_{\mathrm{mRNA,A}} = 5.6 \times 10^{-4}\mathrm{s}^{-1}$, $\nu_{\mathrm{protein,A}} = 1.1 \times 10^{-2}\mathrm{s}^{-1}\mathrm{mRNA}^{-1}$, $\gamma_{\mathrm{mRNA,A}} = 3.1 \times 10^{-5}\mathrm{s}^{-1}$,$\gamma_{\mathrm{protein,A}} = 6.0 \times 10^{-6}\mathrm{s}^{-1}$, $\nu_{\mathrm{mRNA,B}} =$

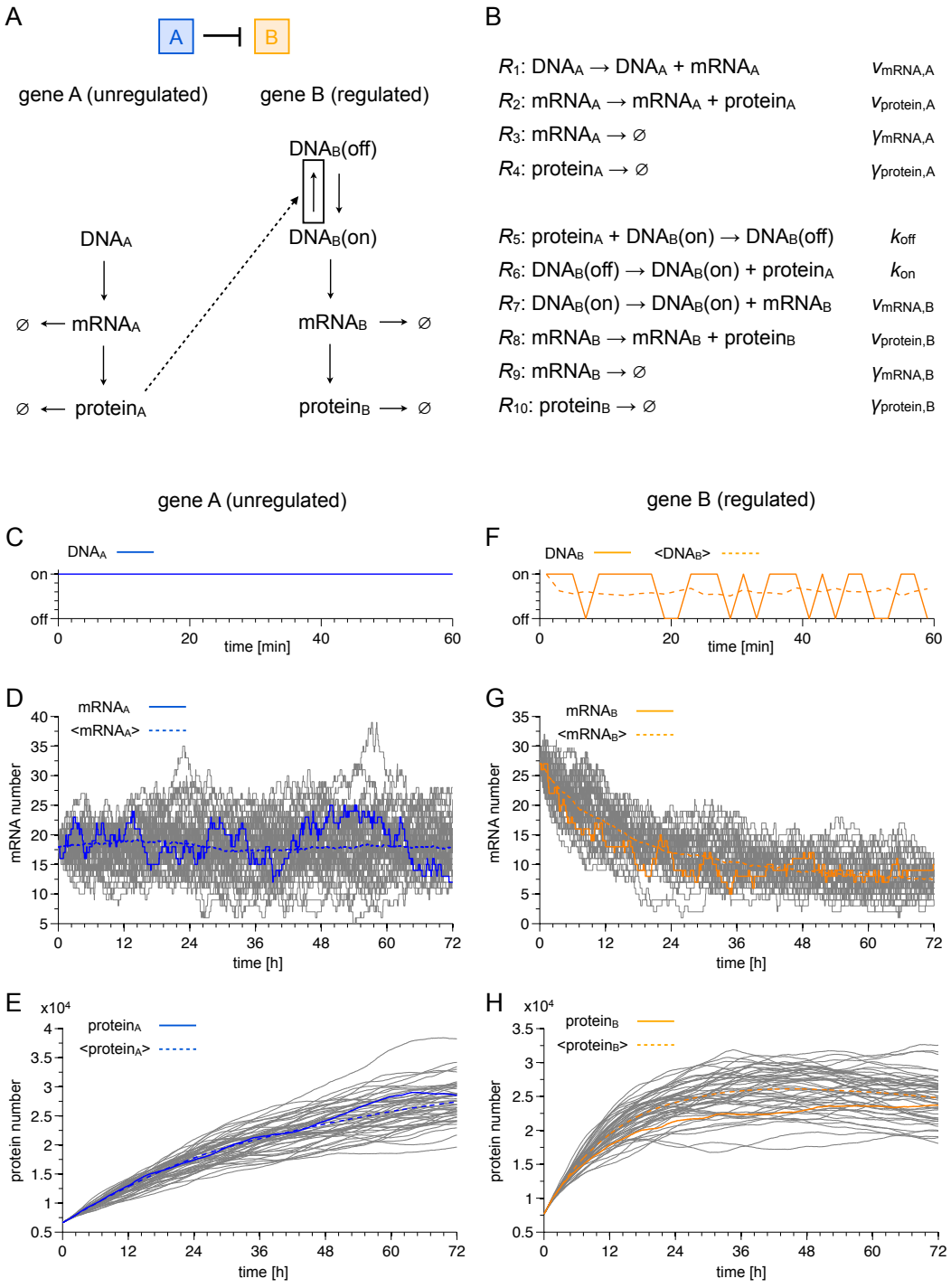


Figure 3.2: Simulated realizations of gene expression using the stochastic simulation algorithm (Figure legend on next page).

Figure 3.2: (From previous page). Simulated gene expression using the stochastic simulation algorithm. (A) We consider a simple gene regulatory network, where gene A is a repressor protein for gene B (see also Figure 2.2 C). Gene A is constantly transcribed into mRNA and translated into protein. Protein A can bind on the repressor site of gene B and can turn gene B into an off state, where no transcription occurs. When protein B unbinds. gene B is turned on and can be transcribed into mRNA, which in turn is translated into proteins. (B) All ten reactions that can occur within the gene regulatory network depicted in (A). Every reaction is governed by its own reaction rate constant. We used values for the reaction rate constants that are similar to the values obtained by Schwannhäusser et al. (2011). (C-H) Time evolution of 100 individual realizations of the CME simulated with the SSA. A single realization is highlighted with a colored solid line; the mean value of the 100 realizations is displayed with colored dashed lines. While gene A remains constantly in the on state, gene B is repressed by protein A and can switch between the on and the off state.

$8.4\times10{-}4\mathrm{s}^{-1}$, $\nu_{\mathrm{protein,B}} = 1.7\times10^{-2}\mathrm{s}^{-1}\mathrm{mRNA}^{-1}$, $\gamma_{\mathrm{mRNA,B}} = 3.1\times10^{-5}\mathrm{s}^{-1}$,$\gamma_{\mathrm{protein,B}} = 6.0\times10^{-6}\mathrm{s}^{-1}$, $k_{\mathrm{on}} = 1.0\times10^{-1}\mathrm{s}^{-1}$ and $k_{\mathrm{off}} = 1.0\times10^{-5}\mathrm{s}^{-5}$.The system starts with initial conditions $x_{\mathrm{mRNA,A}}(t=0) = 18$ mRNA transcripts of gene A, $x_{\mathrm{mRNA,A}}(t=0) = 6,624$ proteins of species A, $x_{\mathrm{mRNA,B}}(t=0) = 27$ mRNA transcripts of gene B, $x_{\mathrm{mRNA,B}}(t=0) = 7,677$ proteins of species B and the DNA of gene B in the on state.

Steady-state distributions of gene expression

We now turn to a very important limit of the CME, namely the limit $t \to \infty$, in which we reach the steady-state where the system approaches is chemical equilibrium and $\partial P(\mathbf{x}, t|\mathbf{x_0}, t_0)/\partial t = 0$. In this case we can obtain analytical solutions of the CME's steady-state distribution for biologically relevant systems. In the previous paragraph we discussed the case where one gene is constantly transcribed whereas the other gene can switch between a state where the DNA is off and a state where it is on and only then can be transcribed. The steady-state of these two cases corresponds to the two-stage model (i.e. constant gene expression) and the three-stage model (i.e. regulated gene expression) that have been introduced in literature (Peccoud and Ycart, 1995; Raj et al., 2006; Raj and van Oudenaarden, 2009; Shahrezaei and Swain, 2008).

For large time scales $t \gg t_0$ the dynamics of the previously described system (see Figure 3.2) approaches its steady-state where the abundances of mRNA A and B stochastically fluctuate around their expectation value (see Figure 3.3 A and C). Then the

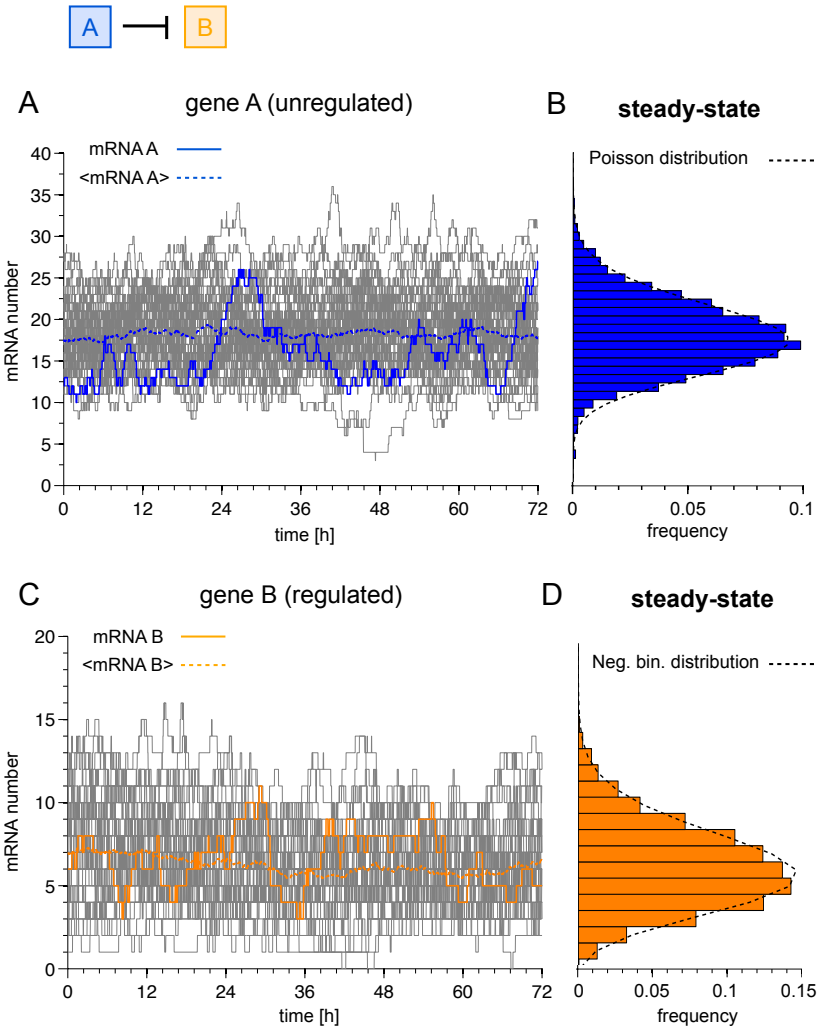


Figure 3.3: Steady-state distributions for the two-stage and three-stage model. (A) Stochastic simulation of 100 realizations of mRNA A transcripts that stem from an unregulated gene once the system is in chemical equilibrium. An individual realization that obeys the CME is displayed with a solid blue line; the mean of all 100 realizations is depicted with a dashed blue line. (B) Steady-state distribution of the unregulated mRNA transcripts of gene A that follows the Poisson distribution (dashed black line). (C) Stochastic simulation of 100 realizations of mRNA B transcripts that come from a regulated gene described by the system displayed in Figure 3.2 for large times (i.e. the system is in its steady-state). An individual realization that obeys the CME is displayed with a solid orange line; the mean of all 100 realizations is depicted with a dashed orange line. (D) The state-state distribution of mRNA B that is transcribed from a regulated gene rather follows the over-dispersed negative binomial distribution (dashed black line).

steady-state distribution can be obtained by sampling from an individual time-series (see Figure 3.3 B and D). Note, however, that the steady-state distributions of molecular species are experimentally often obtained by a snapshot measurement of many single cells (see also Figure 2.3 A and B) rather than by time-dependent observations.

We now describe how the steady-state distributions for the two-stage and the three-stage model can be obtained analytically. First, it is important to note that for the three-stage model we do not consider the coupled gene regulatory network from the previous Section, but look at a regulated gene individually of the abundance of its regulatory protein. By assuming a constant abundance of the regulatory protein the propensity for the off-switching rate of the regulated gene turns into an effective rate constant $k^*_{\text{off}} = x_{\text{protein,B}} k_{\text{off}}$.

The steady-state distribution of mRNA numbers m for the three-stage model contains four reaction rate constants: k_{on}, the rate constant at which the DNA transitions from the off to the on state; k^*_{off}, the (effective) rate constant at which the DNA transitions from the on to the off state; ν, the transcription rate constant; and γ, the mRNA degradation rate constant. It can be shown that the CME has the following steady-state distribution for mRNAs (Raj et al., 2006; Shahrezaei and Swain, 2008)

$$\mathcal{P}(m|\boldsymbol{\theta}) = \frac{\theta_1^m \mathrm{e}^{-\theta_1}}{\Gamma(m+1)} \frac{\Gamma(\theta_2+m)\Gamma(\theta_2+\theta_3)}{\Gamma(\theta_2+\theta_3+m)\Gamma(\theta_2)} \times {}_1F_1(\theta_3, \theta_2+\theta_3+m, \theta_1), \tag{3.15}$$

where $(\theta_1, \theta_2, \theta_3) = (\lambda, \kappa_{\text{on}}, \kappa_{\text{off}})$ and only ratios of the rate constants $\lambda = \nu/\gamma$, $\kappa_{\text{on}} = k_{\text{on}}/\gamma$ and $\kappa_{\text{off}} = k^*_{\text{off}}/\gamma$ enter. Here, $\Gamma(.)$ denotes the Gamma function and ${}_1F_1(.)$ describes the Kummer confluent hypergeometric function.

In the limit of large $\kappa_{\text{off}} \gg 1$ and constant $\xi = \nu/k_{\text{off}}$, Eq. 3.15 turns into a negative binomial distribution

$$\mathcal{P}(m|\boldsymbol{\theta}) = (1+\theta_1)^{-\theta_2} \frac{\Gamma(\theta_2+m)}{\Gamma(\theta_2)\Gamma(m+1)} \left(\frac{\theta_1}{1+\theta_1}\right)^m \tag{3.16}$$

where $(\theta_1, \theta_2) = (\xi, \kappa_{\text{on}})$. In this case the DNA stays only very shortly in the on state and transcription takes place in instantaneous bursts, with an average burst size ξ and a burst frequency κ_{on}.

For $\kappa_{\text{off}} = 0$ the DNA remains constantly in the on state corresponding to the two-stage model. In this case the mRNA steady-state distribution is given by a Poisson distribution

$$\mathcal{P}(m|\theta) = \frac{\theta^m}{m!} \mathrm{e}^{-\theta} \tag{3.17}$$

where $\theta = \lambda$ denotes the average number of mRNA molecules over many single cells. Similar solutions can also be obtained for the protein number steady-state distributions of the two- and three-stage model of gene expression. Since for this thesis only the mRNA distributions are of interest, we refer to Shahrezaei and Swain (2008).

To conclude, the mRNA numbers of gene A, which is unregulated are distributed according to a Poisson distribution where the mean equals the variance and the mRNA numbers of gene B, which is regulated have an over-dispersed steady-state distribution (i.e. the mean exceeds the variance of the distribution, which it does e.g. in case of the negative binomial by a factor of $(1 + \xi) > 1$). The absence of over-dispersion can therefore be used as an indicator for unregulated genes; the presence of over-dispersion, however, can also be due to other confounding sources of variability (such as technical noise) and is not necessarily sufficient to conclude that a gene is regulated (see also Chapter 5).

Considerations on the applicability of the chemical master equation to biological processes

It is worth mentioning that the derivation of the CME as carried out by Gillespie (1992) is intended to hold for chemical systems that behave like an ideal gas. An ideal gas is a hypothetical system where all species are perfectly well mixed and not subjected to any affinity to react with each other, except that they randomly collide in an elastic way. This corresponds to a system of non-interacting particles (or molecules) with no internal degrees of freedom (a real world example where these conditions are approximately fulfilled is a low-density noble gas). In this case the assumptions that were made to integrate out the positions and velocities of the molecules, (I) uniformly distributed positions (well-steardness) and (II) normally distributed velocities (thermal equilibrium; see Figure 3.1 and Section 3.1) are well justified (see e.g. Schroeder (2000)).

The cell, however, has a high degree of internal organization and substructure and the molecular species in the cell are not elastic and have clear affinities to react with other molecules (as described in Section 2.1). It is important to be aware of this fact and the limitations that it implies: a system of interacting molecules with internal degrees of freedom exhibits potential energy that couples the positions and velocities of the molecules (depending on the specific form of the interaction) and gives rise to an altered and joined probability distribution $f(\mathbf{r}_k, \mathbf{v}_k)$ for the positions and velocities.

The ideal gas assumptions (I) and (II) were taken to obtain propensity functions that

are constant with respect to the molecules' positions and velocities. Only with constant propensity functions it is possible to avoid simulating the positions and velocities individually and to work with the probabilistic framework of the CME instead. Note, however, that the derivation of Eq. 3.1 does not necessarily depend on the specific form of the position- and velocity-distributions of the ideal gas. It might therefore be possible to find other distributions $f(\mathbf{r}_k, \mathbf{v}_k)$ for the positions and velocities that could be derived from a more suitable description of the cellular environment, which could also give rise to constant propensity functions.

Chapter 4

Statistical learning methods

Frustra fit per plura,
quod potest fieri per pauciora.

William of Ockham [IV]

Causa latet, vis est notissima fontis.

Ovid [V]

Niemand urteilt schärfer als der Ungebildete;
er kennt weder Gründe noch Gegengründe [...].

Anselm Feuerbach [VI]

In the last Chapter we presented a mathematical framework to describe and simulate biological processes. A mere simulation of a biological system, however, is only of limited predictive power since many different parameters and model topologies could in principle give rise to qualitatively very similar biological observables. Here, we provide the mathematical formulation with which we can quantitatively assess the underlying model topologies and parameters and which conjoins the mathematical formulation of biological systems with the data we can measure.

In Chapter 3 we discussed the properties of data that can be obtained from biological systems and concluded that their information content is limited in the sense that the data is incomplete, subjected to technical noise and often suffers from an averaging out of relevant information. We therefore need mathematical methods that are capable to deal with these limitations and that are often intimately related to the properties of the data. One major challenge of biomathematics and computational biology is the

development of novel methods and the adaptation of existing methods to handle data limitations and to infer new knowledge about biological systems.

In this Chapter, we provide several probabilistic learning methods that are ideally suited to (i) conjoin biological data and the mathematical formulation of biological systems in a probabilistic way to investigate biological mechanisms that lead to cellular heterogeneity and (ii) discover information in data sets in an unbiased way to dissect cellular heterogeneity. The task of these methods is to infer biological relevant information given data from the system. Moreover, a probabilistic approach provides straightforward ways to quantify the error of the estimated conclusions, which is of fundamental importance given the aforementioned data limitations.

From a mathematical point of view biological data can often be understood as a three-dimensional data tensor whose dimensions equal the number of features times the number of samples times the number of time-points (see Figure 4.1). As mentioned before, it is not always possible to observe all three dimensions depending on the particular experiment that is performed. Nevertheless, it is possible to use methods from statistics and machine learning even if only one of these dimensions is observed. For instance if only cellular features are provided for one sample, we can still analyze the mechanisms that lead to the observed features (we do this in Chapter 6, where we analyze the mechanism behind histone tail modification). We can also perform an analysis if only one feature for many samples is given and investigate the steady-state distributions of gene expression (we do this in Chapter 5, where we perform a steady-state distribution analysis for key genes during hematopoiesis). Moreover, there are also measures to analyze data sets with only one or a few different features, but with many time points (we analyze time-series of two proteins during hematopoiesis in Chapter 7). Lastly, we can use machine learning methods for the multivariate analysis of data with a large number of samples and features (we use this approach in Chapter 8 to classify cells based on their morphological features into their cell cycle phase). One of the major tasks in the application of these methods is to find suitable biological questions that can be posed to given data and that can actually be answered with methods from statistics and machine learning.

We start in Section 4.1 by providing the mathematical framework to perform parameter estimation and model selection. In Section 4.2 we discuss ways to analyze time-series data. Section 4.3 contains methods for clustering and dimension reduction of high-dimensional data sets. We conclude with Section 4.4 where we provide an overview over classification and regression methods. In this Chapter we introduce all methods

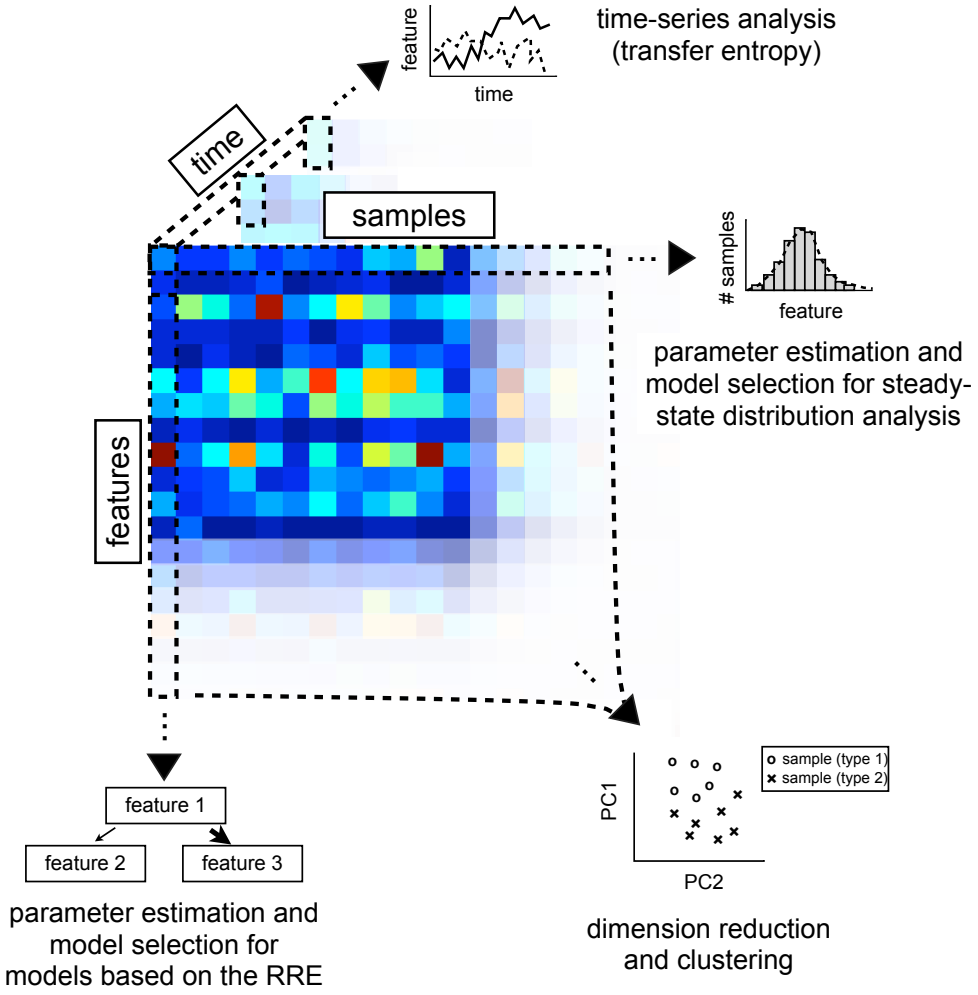


Figure 4.1: Statistical learning methods to extract information from biological data. From a mathematical viewpoint the biological data relevant for this thesis can be understood as a tensor where the dimensions are given by the number of measured features times the number of measured samples times the number of measured time points. There are different mathematical tools to extract information from either of these dimensions individually (indicated by the blacked dashed boxes around vectors from the three mentioned dimensions), such as parameter estimation and model selection for models based on the reaction rate equation (RRE) with data from different features, parameter estimation and model selection on steady-state distribution of many samples, analyzing the time-evolution of one or two features with transfer entropy, and clustering and dimension reduction methods on the samples times features matrix. We present these methods in this Chapter and apply them in the original contributions of this thesis (Chapters 5-8).

that are necessary for the original contributions of this thesis (Chapters 5-8), where we apply them to biological data. The interested reader finds references for further approaches at the end of each Section.

4.1 Parameter estimation and model selection

In this Section we show how we can estimate parameters and perform model selection. As discussed in Chapter 3, we can formulate mathematical models that describe biochemical systems. Here, we present how the parameters of these models can be estimated and how even the model topology can be inferred. This method is particularly suitable if we already have some knowledge about the biological process of interest and we can formulate a model for it. We use parameter estimation and model selection in Chapter 5 on individual features (across samples) and in Chapter 6 on individual samples (across features) to infer model parameters for transcription and histone modifications, respectively (see Figure 4.1).

The likelihood

If we regard both the measured data points D and the model parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ as random variables with a certain probability distribution, we can use Bayes' theorem (see e.g. Murphy (2012)) to denote

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int \mathrm{d}\boldsymbol{\theta} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})} \tag{4.1}$$

for the posterior probability $P(\boldsymbol{\theta}|D)$ of the parameters given the data, with the prior distribution $P(\boldsymbol{\theta})$, the evidence $P(D) = \int \mathrm{d}\boldsymbol{\theta} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})$ and the likelihood $P(D|\boldsymbol{\theta})$. This formulation provides us with a probabilistic framework that incorporates the data into the model in a natural way.

In case we choose a uniform prior distribution, which is done when there is no a priori information about the model parameters available, the posterior distribution is directly proportional to the likelihood $P(\boldsymbol{\theta}|D) \propto P(D|\boldsymbol{\theta})$. We can then work directly with the likelihood instead of the posterior distribution and we do not have to evaluate the potentially high-dimensional integral of the evidence. The explicit formulation of the likelihood depends on both the model formulation and an error model that is chosen to incorporate measurement noise (we formulate likelihood functions in Chapters 5 and 6).

Once we formulated the likelihood of a particular model we can perform maximal likelihood estimation by finding

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} P(D|\boldsymbol{\theta}), \tag{4.2}$$

where, $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate (MLE) of the model parameters.

In some cases it is possible to find closed solutions for the MLE. In this thesis, however, we use local optimization strategies to minimize the negative log-likelihood

$$\mathcal{J}(\boldsymbol{\theta}) = -\log P(D|\boldsymbol{\theta}). \tag{4.3}$$

Additional material on working directly with the posterior probability distribution, which requires solving the integral of the evidence can be found in Hug (2015). Typical methods to numerically calculate the posterior probability distribution are the posterior harmonic mean estimate (Newton and Raftery, 1994), Chib's method (Chib and Jeliazkov, 2001) and thermodynamic integration (Gelman and Meng, 1998).

Uncertainty analysis

As mentioned in Section 2.4 the data of biological systems is often not comprehensive and subjected to technical noise. It is therefore important to have a way to quantify the error in the prediction that can come from this limited amount of information. Furthermore, the model can have a structure that does not allow to identify some of its parameters. A practical way to obtain error bars to the MLE and to assess the identifiability of the parameters is the profile likelihood (PL; Venzon and Moolgavkar (1988)). It allows us to obtain confidence intervals in which the estimated parameters lie with a given significance level.

To determine CIs and to address identifiability of the MLE of the parameters, we can calculate the profile likelihood (PL) for each parameter θ_n (Raue et al., 2009):

$$\mathrm{PL}(\theta_n) = \max_{\boldsymbol{\theta_{n' \neq n}}} \mathcal{J}(\boldsymbol{\theta}). \tag{4.4}$$

This corresponds to maximizing the log-likelihood with respect to all parameters $\boldsymbol{\theta_{n' \neq n}} = (\theta_1, \ldots, \theta_{n-1}, \theta_{n+1}, \ldots, \theta_N)$ for each parameter value θ_n. Point-wise CIs of the parameters to a significance level α can be obtained as

$$\mathrm{CI}(\theta_n) = \{\theta_n | \mathrm{PL}(\theta_n) - \mathcal{J}(\hat{\boldsymbol{\theta}}) < \Delta_\alpha\} \tag{4.5}$$

with the threshold $\Delta_\alpha = (\chi^2)^{-1}(\theta_n \leq \alpha, 1)/2$ determined by the chi-square inverse cumulative distribution function (Raue et al., 2009). In case the confidence intervals are finite the parameters are identifiable.

There are also different approaches to perform uncertainty analysis in the context of parameter estimation. For instance, one can obtain asymptotic confidence intervals with the Fisher information matrix (which corresponds to the Hessian matrix of the negative log-likelihood) or directly from the posterior distribution as posterior profile CIs (see e.g. Hasenauer and Theis (2013) for more details).

Model selection

In many cases it is not the MLE of the model parameters that is of particular interest, but the model topology, which is a more general property of the biological system of interest. To compare two models j and k with different model topology we can calculate the ratio of the posterior probability of two models, which (in case of uniform prior probabilities for both models, i.e. $P(j) = P(k)$) equals the Bayes factor

$$\begin{aligned} B_{jk} &= \frac{P(j|D)}{P(k|D)} \\ &= \frac{\int \mathrm{d}\boldsymbol{\theta}^j P(D|\boldsymbol{\theta}^j) P(\boldsymbol{\theta}^j|j)}{\int \mathrm{d}\boldsymbol{\theta}^k P(D|\boldsymbol{\theta}^k) P(\boldsymbol{\theta}^k|k)}. \end{aligned} \tag{4.6}$$

It can be shown (Kass and Raftery, 1995) that in the limit of large samples (i.e. $\dim D \gg 1$) the Bayes factor of a model j with parameters $\boldsymbol{\theta}^j$ can be approximated by the Bayesian information criterion

$$\mathrm{BIC}_j = -2 \max_{\boldsymbol{\theta}^j} P(D|\boldsymbol{\theta}^j) + \dim \boldsymbol{\theta}^j \log(\dim D), \tag{4.7}$$

via the following limit

$$\lim_{\dim D \to \infty} \frac{-2 \log B_{jk} - (\mathrm{BIC}_j - \mathrm{BIC}_k)}{-2 \log B_{jk}} = 0. \tag{4.8}$$

The BIC is given as the sum of twice the log-likelihood evaluated at the MLE and a term that scales proportional to the number of model parameters and the logarithm of the number of observed data points.

Kass and Raftery (1995) also specified for which values of the Bayes factor (or the BIC) model j should be rejected in favor for model k. With $\Delta\mathrm{BIC} = \mathrm{BIC}_j - \mathrm{BIC}_k$ the

evidence against a model with higher BIC is considered to be 'not worth more than a bare mention' if the difference in the BIC value of the two models $\Delta\text{BIC} < 2$, 'positive' if $2 < \Delta\text{BIC} < 6$, 'strong' if $6 < \Delta\text{BIC} < 10$, and 'very strong' if $\Delta\text{BIC} > 10$.

While in the following we exclusively use the BIC for model selection, there is also another criterion, known as the Akaike information criterion (AIC) based on which models can be selected. The AIC is derived as the Kullback-Leibler divergence between a candidate model and the true underlying model. Moreover, there are different approximations to the Bayes factors available than the BIC (such as the Laplace approximation) or it is possible to directly evaluate the Bayes factor without further approximations (Hug, 2015). In case we have to select between nested models (model j is nested in model k when model j can be obtained from model k by fixing some of model k's parameters) we can perform the likelihood ratio test, which is based on Wilks' theorem (Wilks, 1938). Additional reading on other model selection strategies can be found in Hasenauer and Theis (2013).

4.2 Time-series analysis

Time-series can be obtained by observing cellular features and samples over time t (see Figure 4.1). We have seen in the previous Chapter how time-dependent biological processes can be simulated in an efficient way. In the context of cellular processes, however, it is often challenging to obtain time-dependent measurements from the same cell, since experiments tend to be invasive (see Section 2.4). In general, there are many time-series analysis methods that can be applied to biological data (see e.g. Bar-Joseph et al. (2012) for a review). For instance it is straight-forward to formulate the likelihood for a time-dependent biological model given measured data and to infer its systems parameters with the methods introduced in the previous Section (we formulate the likelihood for models based on the RRE in Chapter 6). Another way to learn about the biological process given measured data is by exploiting its correlation structure. In this thesis, we want to focus on one particular sort of time series analysis that relies on such correlation structure: transfer entropy (Schreiber, 2000). In Chapter 7 we use transfer entropy to infer the information transfer between two protein species.

Transfer entropy

We start by introducing Shannon entropy, which measures the amount of information in a random variable (Shannon, 1948). Given a discrete random variable X that can take values $X \in \{x_1, \ldots, x_N\}$, each with probability $p(x_i)$ (where $i = 1, \ldots, x_N$) it is defined by

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log p(x_i). \tag{4.9}$$

From an information theoretic point of view it quantifies the average information contained in the discrete random variable X that is distributed according to $p(X)$. In other words, it measures the uncertainty in the random variable X. In case of a certain event where $p(X = x_{i^*}) = 1$ the Shannon entropy becomes zero, whereas it reaches its maximum if X is uniformly distributed.

The Shannon entropy can also be formulated for two discrete random variables X and Y leading to the joint entropy

$$H(X, Y) = -\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} p(x_i, y_j) \log p(x_i, y_j) \tag{4.10}$$

with the joint probability distribution $p(x_i, y_j)$ for X being in state $X = x_i$ while Y being in state $Y = y_j$ corresponding to the joint information content in both random variables X and Y.

Moreover, we can compute the conditional entropy

$$H(X|Y) = -\sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} p(x_i, y_j) \log p(x_i|y_j), \tag{4.11}$$

which corresponds to the information content of the random variable X when the other random variable Y is already known.

Let $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ denote two time series then the transfer entropy can then be obtained by

$$\mathrm{TE}_{X \to Y} = H(Y|Y_t^{(l)}) - H(Y|Y_t^{(l)}, X_\tau^{(k)}), \tag{4.12}$$

where $X_\tau^{(k)}$ and $Y_t^{(l)}$ denote the elements of the time series X and Y that are shifted by τ and t and of length k and l, respectively. More explicitly, the transfer entropy reads

$$\mathrm{TE}_{X \to Y} = \sum_{i > \max(t,\tau)}^{N} p(y_i, y_{i-t}^{(l)}, x_{i-\tau}^{(k)}) \log \frac{p(y_i|y_{i-t}^{(l)}, x_{i-\tau}^{(k)})}{p(y_i|y_{i-t}^{(l)})} \tag{4.13}$$

with $x_{i-\tau}^{(k)} = (x_{i-\tau}, \ldots, x_{i-\tau-k+1})$ and $y_{i-t}^{(l)} = (y_{i-t}, \ldots, y_{i-t-l+1})$.

For the remainder of this thesis we follow Schreiber (2000) and set $k = l = 1$ and $\tau = t$. In this case Eq. 4.13 simplifies to

$$\mathrm{TE}_{X\to Y}(\tau) = \sum_{i>\tau}^{N} p(y_i, y_{i-\tau}, x_{i-\tau}) \log \frac{p(y_i|y_{i-\tau}, x_{i-\tau})}{p(y_i|y_{i-\tau})}. \tag{4.14}$$

Transfer entropy quantifies the amount of information in time-series Y that is already contained in time-series X. In both limits where either $X = Y$ or X is statistically independent from Y the transfer entropy between them becomes $\mathrm{TE}_{X\to Y}(\tau) = 0$, as in both limits $P(y_i|y_{i-\tau}, x_{i-\tau}) = P(y_i|y_{i-\tau})$ and the argument of the logarithm in Eq. 4.14 becomes 1.

4.3 Dimension reduction and clustering

Dimension reduction and clustering are methods from the field of unsupervised machine learning (see e.g. Hastie et al. (2009) or Murphy (2012)). While dimension reduction methods are used to map a high dimensional data set into a low dimensional representation with the aim to preserve the information content in the data, clustering methods are used to group objects into distinct subsets or 'clusters' that share certain similarities. These methods are often used to investigate data sets in an explorative way and to generate new hypotheses. In the context of dissecting cellular heterogeneity, clustering and dimension reduction algorithms can be used to search in gene expression data sets for new cellular phenotypes (see e.g. Sandberg (2014) for an overview and Buettner and Theis (2012) for an application). In Chapter 6 we show how new mathematical methods can further contribute to improve this quest for identifying cellular phenotypes. In Chapter 7 we use hierarchical clustering to find similarities not among cells but among different candidate models.

Dimension reduction

Dimension reduction methods are mappings $f(.)$ from a high dimensional data space with $\dim D = N \times M$ (where N equals the number of features and M equals the number of samples; see Figure 4.1) to a low dimensional representation space D' with $\dim D = N' \times m$, where $N' \leq N$, $f(D) \to D'$. The aim of the mapping $f(.)$ is to preserve the information content from the data D. The dimension N' of the low dimensional data

space D' is often chosen to be $N' = 2$ in order to to visualize the M samples in a two-dimensional figure. It is often beneficial to have a one-to-one correspondence between features in the original data space D and in the mapped data space and/or a one-to-one correspondence between the features (also known as loadings) in the low dimensional space. A well-known method that performs this task is principal component analysis (PCA; Hotelling (1933)).

Latent variable models

One way to formulate a probabilistic framework for dimension reduction mappings is given by latent variable models (see e.g. Murphy (2012)). Let y_m denote the N-dimensional feature vector of one observed data sample (i.e. $D = \{y_1, \ldots, y_M\}$) and let x_m denote a N'-dimensional latent variable vector of the mapped data space (i.e. $D' = \{x_1, \ldots, x_{M'}\}$). For now, we assume there is a linear mapping W between the feature vector and the latent variable vector of the form

$$y_m = Wx_m + \eta_m, \tag{4.15}$$

that is additionally subjected to Gaussian noise

$$P(\eta_m) = \mathcal{N}(\eta_m|0, \beta^{-1}I) \tag{4.16}$$

with zero mean and variance β^{-1}. Here, I denotes the N-dimensional unity matrix.

The likelihood (see Section 5.1) for the observed data sample y_m is then given by

$$P(y_m|W, x_m, \beta) = \mathcal{N}(y_m|Wx_m, \beta^{-1}I). \tag{4.17}$$

This likelihood is the starting point for latent variable models (see discussion below).

For the purpose of this thesis we proceed by marginalizing the likelihood, Eq. 4.17, over the latent variables

$$P(y_m|W, \beta) = \int \mathrm{d}x_m P(y_m|x_m, W, \beta)P(x_m). \tag{4.18}$$

By choosing a normal distributed prior distribution for the latent variables $P(x_m) = \mathcal{N}(x_m|0, I)$ we end up with

$$P(y_m|W, \beta) = \mathcal{N}(y_m|0, WW^T + \beta^{-1}I). \tag{4.19}$$

The mapping W^{MLE} we obtain by maximizing the likelihood, Eq. 4.19, for all data samples M is known as probabilistic principal component analysis (PPCA) introduced by Tipping and Bishop (1999). They showed that the maximization corresponds to finding the eigenvectors of the $N \times M$ dimensional data matrix. To eventually obtain a lower dimensional representation of the data we have to calculate $x_m = (W^{\mathrm{MLE}})^{-1}$. PPCA is characterized by a linear mapping W and a covariance matrix $\beta^{-1}I$ with scalar β. If we allow for different values along the diagonal of the covariance matrix in Eq. 4.19 (i.e. β is given by a vector) the mapping is known as factor analysis (see e.g. Murphy (2012)).

Another approach for LVMs is to marginalize Eq. 4.17 not over the latent variables x_m but over the parameters W. By assuming a Gaussian prior for the parameters W one obtains a likelihood $P(y_m|x_m, \beta)$ corresponding to Eq. 4.19 (a method known as dual PCA (Lawrence, 2004)). If we replace the linear covariance matrix by a non-linear kernel we end up with a mapping known as Gaussian process latent variable model (GP-LVM; see Lawrence (2004) and Lawrence (2005)).

Since there is no clear way of evaluating the goodness of a dimension reduction mapping a priori, there is a multitude of different methods known in literature (for an overview see e.g. Fan and Kamath (2014)), many of which have been successfully applied to dissect heterogeneity in biological data (see e.g. Haghverdi et al. (2015), who also present a comparison between the applicability of different approaches on biological data). Alternative methods to the LVMs introduced above that are commonly applied are t-SNE (t-distributed stochastic neighbor embedding; Van der Maaten and Hinton (2008)) and diffusion maps (Coifman et al., 2006).

Clustering

Often dimension reduction and clustering are performed in combination, when first a high-dimensional data set is mapped into lower dimensions and then the resulting data points are grouped into clusters based on their closeness in the mapped space. This clustering is known as 'partitional clustering' (see Murphy (2012)). Typical methods are mixture models (see e.g. McLachlan and Peel (2000)) and 'k-means clustering MacQueen (1967). In this thesis we use hierarchical clustering, which – in addition to partitional clustering – aims to find nested pairs of relations among the data points. The following paragraph is adopted from chapter 25.5 of Murphy (2012).

Hierarchical clustering

The aim of hierarchical clustering is to merge data points into clusters whilst keeping a nested hierarchy that indicates the degree of similarity between the data points. For this purpose a dissimilarity matrix is calculated that quantifies how different individual data points are. Then subsequently pairs of data points are merged into clusters based on them being minimally dissimilar. The algorithm can be denoted as follows:

0. (a) initialize clusters as single data points $C_m = \{m\}$ for all data points m.
0. (b) initialize the set of clusters that can be merged: $S = 1, \ldots, M$.
0. (c) Calculate the initial dissimilarity matrix $d(j, k)$
1. Find the 2 most similar clusters and merge them: $(j, k) = \arg\min_{j,k \in S} d(j, k)$.
2. Define new cluster $C_l = C_j \cup C_k$.
3. Remove j and k from the set of clusters that can be merged $S = S \setminus \{j, k\}$.
4. If $C_l \neq \{1, \ldots, M\}$ mark l as available, $S = S \cup \{l\}$.
5. Calculate the dissimilarity matrix $d(i, l)$ for each $i \in S$.
6. Go back to 1. and repeat until no more clusters can be merged.

There are different ways to define the dissimilarity matrix $d(j, k)$. A popular choice is average linkage clustering (Sokal and Michener, 1958), where the average distance of all pairs is measured:

$$d_{\text{avg}}(G, H) = \frac{1}{n_G n_H} \sum_{i \in G} \sum_{i' \in H} d_{i,i'} \tag{4.20}$$

with n_G and n_H denoting the number of elements in each group and $d_{i,i'}$ a distance measure (usually a metric) between two data points in the high-dimensional data space.

4.4 Regression and classification

After we have discussed methods from the field of unsupervised machine learning in the previous Section, we now present supervised machine learning methods, namely regression and classification. Although both concepts fulfill distinct purposes, i.e. classification aims to predict a class label and regression aims to predict a continuous variable, they can be derived from the same premises as we detail below. In supervised machine learning we use a subset of the data (known as 'training data set') to

minimize a target variable (also known as 'training'), which can be continuous (in the case of regression) or discrete (in the case of classification) before we use the trained algorithm to predict the target variable on an independent subset of the data for which the target variable is unknown. During the training step the algorithm can find the most important features in the data set, which then can be used to make predictions for the independent data set. Note that now we specify a target variable that we like to predict whereas in the case of unsupervised machine learning the algorithms try to find similarities among data points in an entirely unspecified and unbiased way.

When using regression and classification methods it is very important to ensure that the algorithms do not overfit the training data set and lack predictive power on the independent data set. One way to prevent overfitting is by applying cross-validation. There the predictive power of the algorithm is tested on another subset of the data (known as 'test data set') and the training and testing steps are repeated multiple times before applying the trained algorithm to make predictions on the independent data set. We use regression and classification in Chapter 9 to predict the position of cells in their cell cycle. In this Section we start our presentation with the formulation of regression in a general way, before we outline the specific method, classification and regression trees that we applied in Chapter 9. The following parts of this Section are extracted from chapters 1, 7, 14 and 15 of Murphy (2012).

Linear regression and classification

Given a data set $D = \{y_1, \ldots, y_M\}$ (corresponding to the test data set) of M data samples y_m containing $\dim y_m = N$ features, the likelihood for the linear regression of a target variable z_m ($m = 1, \ldots, M$), which is not among the features of y_m is given by

$$P(z_m|y_m, \boldsymbol{\theta}) = \mathcal{N}(z_m|w^T\phi(y_m), \sigma^2). \tag{4.21}$$

Here $\boldsymbol{\theta}$ indicates a vector of parameters containing the linear weights w of the regression, the parameters of the mapping $\phi(.)$, which can be a non-linear function of the feature vector y_m and the variance σ^2. Note that Eq. 4.21 is called linear regression since it is still linear with respect to the weights w, even though non-linear interdependencies in the data set can be incorporated with the non-linear mapping $\phi(.)$.

In case $\phi(.) = \mathrm{Id}$ (the identity matrix), we obtain the simplest form of linear regression,

where the expectation value of the normal distribution in Eq. 4.21 turns into

$$\begin{aligned} E[z_m|y_m] &= \bar{z}_m(\boldsymbol{\theta}) \\ &= w_0 + \sum_{n=1}^{N} w_n \cdot y_{n,m}. \end{aligned} \tag{4.22}$$

Then the negative log-likelihood $\mathcal{J}(\boldsymbol{\theta}) = -\log P(D|\boldsymbol{\theta})$ takes the following form

$$\mathcal{J}(\boldsymbol{\theta}) = \frac{M}{2}\log(2\pi\sigma^2) + \frac{\sum_{m=1}^{M}(z_m - \bar{z}_m(\boldsymbol{\theta})^2}{2\sigma^2}. \tag{4.23}$$

Similar closed form solutions can be obtained also for non-linear mappings $\phi(.)$, since the negative log-likelihood still remains linear with respect to the parameters w.

To render the mapping $\phi(.)$ non-linear, often kernel functions $\kappa(y_m, y'_m)$ can be chosen such that $\phi(y_m) = [\kappa(y_m, y_1), \ldots \kappa(y_m, y_M)]$. One class for the choice of kernels, e.g., are radial basis functions $\kappa(y_m, y'_m) = g(||y_m - y'_m||, \boldsymbol{\theta})$, which only depend on the distance between y_m and $y_{m'}$. Since the kernels also depend on the parameters $\boldsymbol{\theta}$ that have to be estimated from the data it is often beneficial to enforce sparsity on the parameters. This can be achieved by the selection of suitable priors for the parameters, in case of sparse vector machines (see e.g. Krishnapuram et al. (2005) for a Laplacian prior) or by modifying the likelihood term with additional constraints, in case of the support vector machine (SVM; Cortes and Vapnik (1995)).

The same approach can be taken for linear classification, where we only have to alter two things: (i) the normal distribution in the likelihood, Eq. 4.21, has to be changed to a Bernoulli distribution and (ii) we ensure the prediction to be between 0 and 1 by applying a sigmoid function to the prediction

$$P(z_m|y_m, \boldsymbol{\theta}) = \mathrm{Ber}(z_m|\mathrm{sigm}(w^T\phi(y_m))), \tag{4.24}$$

with $\mathrm{sigm}(y_m) = 1/(1 + \exp(-y_m))$.

In contrast to regression there is no closed form solution for the maximum likelihood in the case of classification. However, there are several iterative algorithms that can fulfill this task. E.g. one class of algorithms, such as the steepest descent or the conjugate gradient method, use the gradient at the current step of the algorithm, which corresponds to the direction that points towards the maximum of the likelihood in order to find the next step in an iterative way until the maximum is reached.

Classification and regression trees

A different approach to predict the target variable z_m is by learning the basis functions from the data itself – instead by specifying them in advance as was done in the previous subsection – known as adaptive basis function models. The prediction then takes the form

$$\begin{aligned} E[z_m|y_m] &= f(y_m, \boldsymbol{\theta}) \\ &= w_0 + \sum_{k=1}^{K} w_k \phi(y_m, v_k) \end{aligned} \tag{4.25}$$

where $\phi_k(.)$ is the kth basis function learned from the data and v_k are the parameters of the basis function.

In this thesis we use classification and regression trees (CARTs) as the basis functions. There, the data set is split into a tree structure with K different regions R_k ($k = 1, \ldots, K$). In this case, Eq. 4.25 becomes

$$f(y_m, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k I(y_m \in R_k), \tag{4.26}$$

The regions are obtained by growing a tree where subsequently maximally distinguishing features n^* and thresholds t_{n^*} on them are chosen according to

$$\begin{aligned} (n^*, t_{n^*}) = \arg \min_{n \in 1,\ldots,N} \min_{t \in \mathcal{T}_n} &\text{cost}(\{D, z_1, \ldots, z_M : y_{n,m} \leq t\}) \\ &+ \text{cost}(\{D, z_1, \ldots, z_M : y_{n,m} > t), \end{aligned} \tag{4.27}$$

where $y_{n,m}$ is the nth feature value of data sample m and $\mathcal{T}_n$ is the set of possible threshold for feature n.

The choice of the cost-function depends on whether we want to perform regression or classification. In case of regression we define

$$\text{cost}(D, z_1, \ldots, z_M) = \sum_{m=1}^{M} (f(y_m) - z_m)^2, \tag{4.28}$$

whereas in classification we use

$$\text{cost}(D, z_1, \ldots, z_M) = \frac{1}{M} \sum_{m=1}^{M} I((\text{sig}mf(y_m)) = z_m). \tag{4.29}$$

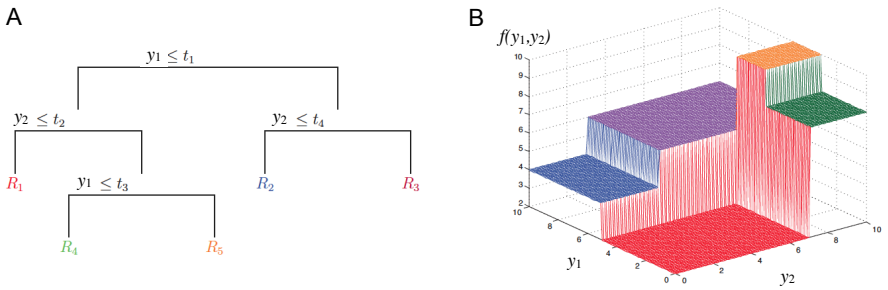


Figure 4.2: Example of a simple regression tree. (A) Graphical representation and (B) prediction of a simple regression tree. The regression tree divides the two dimensional feature space of y_1 and y_2 into $K = 5$ five regions R_k $(k = 1, \ldots, K)$. Subsequently it finds the maximally distinguishing feature $(n = 1, 2)$ and the threshold t_n value that minimize the cost function. The prediction for each region corresponds to the z-axis of (B) are given by the weights w_n. Figure adopted and modified from Murphy (2012).

An example for a simple regression tree on a two dimensional feature space is given in Figure 4.2.

There is a variety of choices for the basis functions. For instance the discussed classification and regression trees can also be used for training random forests (Breiman, 2001). A different choice for basis functions are splines resulting, e.g., in multivariate regression splines (Hastie et al., 2009). A very fashionable choice of basis functions is taken in the framework of feedforward neuronal networks as so called hidden layers $\phi(y_m, v_k) = a(v_k^T y_m)$ with non-linear activation functions $a(.)$ where $k = 1, \ldots, H$ runs over all hidden units H and $V = (v_1, \ldots, v_H)$ denotes the weight matrix (see (Bengio et al., 2015) for an introduction to neuronal networks and deep learning).

Boosting

Boosting is an algorithm that allows to fit adaptive basis function models, such as CARTs, which we discussed in the previous subsection, in a greedy way (Schapire, 1990). In boosting weak learners, which are adaptive basis functions that are efficient to fit and that have only a poor performance on their own are subsequently fitted to the data. While there are different choices for the weak learners, we focus on CARTs, Eq. 4.26, with only a few regions $K \lesssim 10$ for the remainder of this thesis.

The objective of boosting is to minimize a loss-function

$$\min_{f} \sum_{m=1}^{M} L(z_m, f(y_m, w_k)) \tag{4.30}$$

that depends on the aim of the prediction.

In case of least-squares boosting the loss-function is given by

$$L^{(l)}_{\text{L2B}}(z_m, f_l(y_m, w_k^{(l)})) = (z_m - f_l(y_m, w_k^{(l)}))^2. \tag{4.31}$$

At each iteration step l a new learner $f_l(y_m, w_k^{(l)})$ with parameters $w_k^{(l)}$ is fit to the difference between the data and the aggregated prediction of all previously grown weak learners $f_l(y_m, w_k^{(l)})$. Once a pre-defined number of maximal weak learners L is reached the prediction of least-squares boosting for the target variable is given by the expectation value of all learners. The algorithm that solves this objective is known as L2boosting (see Hastie et al. (2009) for details).

Another loss-function is chosen for binary classification

$$L^{(l)}_{\text{AM1}}(z_m, f_l(y_m, w_k^{(l)}) = \sum_{m=1}^{M} g_m^{(l)} I(z_m \neq f_l(y_m, w_k^{(l)})) \tag{4.32}$$

where at each iteration step l the loss-function data points y_m, which were classified incorrectly before become an increased weights $g_m^{(l)}$ to ensure that they are also fitted properly. After a pre-defined number of maximal weak learners L is fitted the average of all weak learners decides about the prediction of the class label also known as majority vote. Boosting for binary classification is solved by the AdaBoost.M1 algorithm (see Freund and Schapire (1996) where also a generalization for multi-class classification can be found).

Chapter 5

cgCorrect: correcting single-cell gene expression data for confounding cell growth effects

Transcriptional regulation of gene expression is one of the core principles that allow cellular heterogeneity to occur among cells that share the same DNA (see Section 2.1). A mathematical way to infer information about the underlying regulation status of a gene is given by the analysis of steady-state distributions of mRNA transcript numbers. Moreover, it is possible to dissect the cellular heterogeneity by applying dimension reduction methods and by looking for cells that form distinct clusters. However, there can be different confounding factors that the measured data from cells may exhibit. In Section 2.4 we discuss that due to the experimental techniques applied to measure the data, technical noise is present that leads to additional confounding variability. Furthermore, additional variability in the data (beyond cellular heterogeneity) can be caused by differences in cell sizes. Therefore there is a need for the adaption and extension of existing methods to analyze gene expression data.

In this Chapter we introduce a novel mathematical framework that is capable to correct confounding effects that come from technical noise and from different cell sizes. We outline the problems that arise when cells of different sizes are used for steady-state distribution analysis in Section 5.1. In Section 5.2 we show how current analysis methods can be extended to correct for confounding cell size effects for both dimension reduction mappings and the analysis of steady-state distributions. In the following we apply our

new mathematical framework to artificially simulated data (Section 5.3) and to data from a single cell qPCR experiment (Section 5.4) before we discuss the assumptions and possible limitations of the presented method in Section 5.5.

We find that many genes may misleadingly be interpreted to originate from the three-stage model of gene expression (also known as bursty gene expression) corresponding to a regulated gene when differences in cell size are not taken into account. When correcting for differences in cell size many genes can actually be understood in terms of the simpler two-stage model of gene expression (also known as simple gene expression) corresponding to an unregulated gene (see Section 3.4).

This Chapter is based on and in parts identical with the following manuscript that is currently under preparation:

Blasi, T., Buettner, F., Strasser, M.K., Marr, C. and Theis, F.J. cgCorrect: A method to correct for confounding cell-cell variation due to cell growth in single-cell qPCR data. *In preparation.*

5.1 Biological background and problem statement

Recent technical advances allow for the analysis of single cells with high-throughput omics technologies (Wang and Steven, 2010). Investigating transcripts of single cells with both quantitative real-time PCR (qPCR; Citri et al. (2012); Ståhlberg and Martin (2010)), and single-cell RNA sequencing (RNA-seq; Islam et al. (2011, 2014); Tang et al. (2009); Yan et al. (2013)) has become possible. But new experimental methods bring new challenges with them: biological variability among single cells, which remained hidden in population based approaches now becomes evident. One major challenge of bio-mathematics and computational biology is the development of new and the adaptation of existing methods for single-cell gene expression data (Buettner and Theis, 2012; Kim and Marioni, 2013).

Gene expression is a stochastic process (see Section 2.2 and Chapter 3) and the abundance of mRNA transcripts (of an individual gene) among many single cells (of the same cell type) can be formulated in terms of steady-state probability distributions (see Section 3.4). Analyzing these steady-state probability distribution can yield new insights into the underlying gene expression mechanism (Kim and Marioni, 2013; Larson, 2011; Shahrezaei and Swain, 2008).

There are two well-studied mechanisms of gene expression that have been serving as a paradigm (Raj and van Oudenaarden, 2008): Simple, constitutive gene expression (also known as the two-stage model), where DNA is continuously transcribed to mRNA (see Figure 5.1 A) and bursty gene expression (also known as the three-stage model), where the promoter of the DNA successively switches between an active and inactive state and transcripts are produced in episodic bursts (see Figure 5.1 B). The steady-state distributions of simple gene expression follows the Poisson distribution (Peccoud and Ycart (1995); Thattei and van Oudenaarden (2001); see also Figure 3.3 B) whereas the steady-state distribution of bursty gene expression follows the over-dispersed negative binomial distribution ((Raj et al., 2006); see also Figure 3.3 D), which allows for more variability among the transcript numbers.

Besides the stochastic nature of gene expression that gives rise to this insightful biological variability, there are also other, confounding sources of variability, such as technical noise (Brennecke et al., 2013; Buettner et al., 2014; Ramsköld et al., 2012; Vallejos et al., 2015) and cell cycle effects. Especially the influence of the latter on the interpretation of gene expression data based on steady-state probability distributions has not been sufficiently investigated so far, even though confounding cell cycle effects appear in all proliferating cells (such as stem and progenitor cells). During cell cycle, the cell grows and the number of transcripts within a cell doubles on average (Mitchison, 2003). Recently Padovan-Merhar et al. (2015) found experimental evidence for the compensation of differences in cell size and suggest that the concentration of transcripts within a cell is maintained constant. This means that measuring the abundance of a particular transcript in two identical cells with different cell sizes will yield different results. The differences in cell size cause a broadened, over-dispersed steady-state distribution of transcript numbers, which may be mistakenly interpreted in an upstream analysis.

To illustrate this issue we consider the following scenario (illustrated in Figure 5.1 C): Assume we measure the mRNA transcripts of a particular gene from several single cells, which have the same volume. The gene of interest is subjected to simple, constitutive gene expression and follows the Poisson distribution. In a typical experiment, however, cells are not synchronized and single cells with different sizes are pooled together (see Figure 5.1 C) leading to an over-dispersed steady-state distribution. Performing model selection (see Section 5.1) on the steady-state distribution of transcript numbers obtained by this type of experiment incorrectly favors the negative binomial over the Poisson distribution and therefore the gene expression mechanism would be interpreted to be bursty.

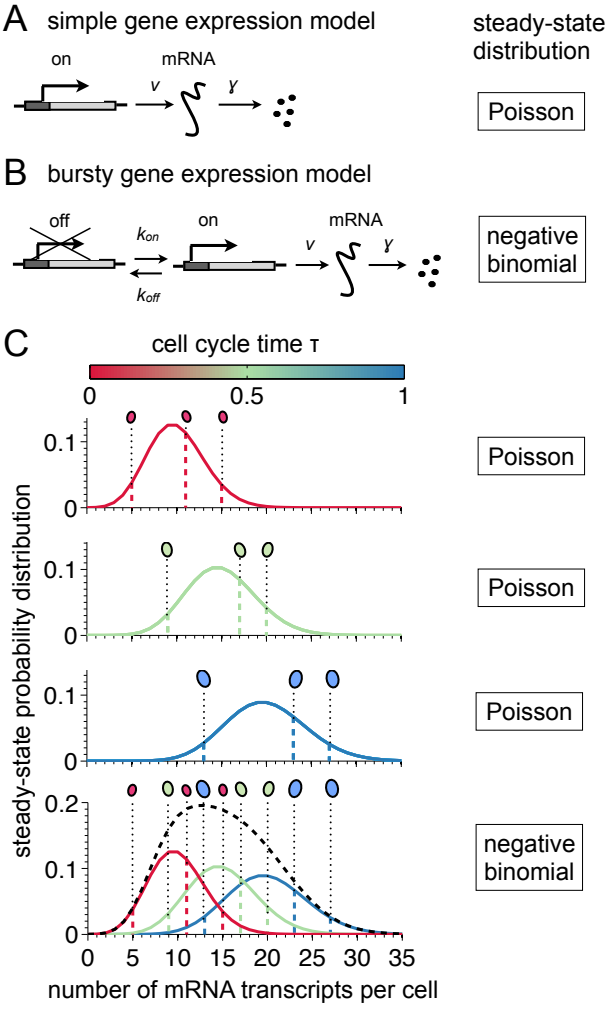


Figure 5.1: Differences in cell size lead to a broadened mRNA distribution and can lead to incorrect identification of the underlying gene expression mechanism (Figure legend on next page).

5.2 cgCorrect: A probabilistic method to correct for confounding cell growth effects

Here, we introduce cgCorrect (cell growth correction), a statistical method to correct single-cell transcriptomics data for latent differences in cell size. cgCorrect can be used for both normalizing single-cell gene expression data sets, and for parameter estimation and model selection on steady-state distributions of gene expression. Our approach is based on the assumption that the average number of mRNA transcripts within the cell

Figure 5.1: (From previous page). Differences in cell size lead to a broadened mRNA distribution and can lead to incorrect identification of the underlying gene expression mechanism. (A) Simple gene expression mechanism. The promoter of a particular gene is always active transcribing its associated DNA to mRNA with rate ν. The degradation rate of the mRNA is given by γ. (B) Bursty gene expression mechanism. Additionally to the simple gene expression mechanism, the promoter can perform transitions between the active and inactive state with rates k_{on} and k_{off}, respectively. (C) Observing mRNA transcripts from single cells with different cell sizes obscures their true underlying steady-state distribution of transcript numbers. We display the Poisson distribution of steady-state transcript numbers for three generic cells with different cell sizes (increasing volume from top to bottom) that are all subjected to constitutive, simple gene expression. By pooling these 9 cells together and ignoring their different volumes, a broadened mRNA distribution is observed (bottom panel, dashed black line) that does not follow a Poisson distribution any more. This will in turn lead to wrong conclusions about the underlying gene expression mechanism.

increases proportionally to the volume as the cell grows during cell cycle, leaving the concentration of transcripts constant (Padovan-Merhar et al., 2015).

We calculate the cell growth correction probability, which corrects for differences in transcript numbers that are due to differences in cell size. This is the conditional probability, for finding the corrected, cell growth independent number of mRNA transcripts of a particular gene, given the measured, cell growth dependent number of mRNA transcripts of this gene. cgCorrect can include information on the cells' volume, but, more strikingly, it can also be applied if there is a total lack of additional information on the cell's volume. Since the cell volume is typically not observed, we marginalize this latent variable out, which corresponds to a blind deconvolution problem.

cgCorrect is based on discrete molecule numbers of individual mRNA transcripts in single cells. Discrete molecule numbers are essential for the interpretation of the underlying mechanism of gene expression (Raj and van Oudenaarden, 2009). There are two high throughput transcriptomics techniques, qPCR and RNA-seq, which both hold the ability to measure discrete molecule numbers in single cells (e.g. via digital PCR (Vogelstein and Kinzler, 1999), droplet digital PCR (Hindson et al., 2011), direct RNA sequencing (Ozsolak et al., 2009) or strand-specific single-cell sequencing (Islam et al., 2011)). Especially the use of unique molecular identifiers for quantitative RNA-seq (Islam et al., 2014) offers a powerful method to perform this task. If the experiment does not provide discrete molecule numbers, the data can be converted to such by

matching the measured value (e.g. cycle time (ct) values in qPCR experiments or reads per kilo-base of transcript per million mapped reads (RPKM) values in RNA-seq) to known absolute molecule numbers of a particular gene in the same cell type.

Current state-of-the-art normalization techniques to account for confounding variability are based on scaling the measured number of mRNA transcripts with reporters that should correlate with the confounding variability. In qPCR where the mRNA transcripts of only a few genes are observed the measured number of transcripts is scaled with the abundance of house-keeping gene transcripts from the same single cell (Guo et al., 2010; Liviak et al., 2013; Moignard et al., 2013). In RNA-seq experiments where the whole transciptome is measured the sum of all mRNA transcripts or rank statistics thereof can be used as an estimator for the cell size of each single cell (Brennecke et al., 2013; Glusman et al., 2013; Sasagawa et al., 2013; Vallejos et al., 2015). However, scaling does not account for the discreteness of mRNA numbers.

Scaling normalization strategies can also be performed based on genes selected from the data as has been pointed out for bulk measurements (Glusman et al., 2013). Whereas this approach is infeasible for single cell qPCR, it is applicable for single cell RNA-seq data since there the whole genome is measured. For instance, it has been shown that the covariance of cell cycle related genes can be used to correct for specific gene expression during cell cycle phases (Buettner et al., 2015). However, this is not the focus of this work where we introduce a correction scheme that is based on a global characteristic of each sample, namely the cells' volume, rather than on the correlations among the expression of different genes.

The cell growth correction probability

Measuring the abundance of a particular mRNA in a single cell during its cell cycle yields a discrete transcript number m, which is generally greater than the transcript number m_0 that we would find at the beginning of the cell's cell cycle ($\tau = 0$). During cell cycle the size of the cell increases from its initial volume $V_0 = V(\tau = 0)$ (at the beginning of its cell cycle) to $V(\tau > 0)$. Cell cycle and cell growth are intimately related (Kafri et al., 2013; Mir et al., 2011) and the number of mRNA transcripts within the cell increases as the cell volume increases. Therefore, we assume the concentration of mRNA transcripts m/V to remain constant during cell cycle. To render the numbers of mRNA transcripts from single cells with different cell sizes comparable, we introduce the volume-dependent cell growth correction probability $P_{\mathrm{cgc}}(m_0|m, V)$. This is the

probability of finding m_0 mRNA transcripts within a cell's initial volume V_0 given a measured number of mRNA transcripts m within a cell's total volume V. The volume-dependent cell growth correction probability is described by a binomial distribution

$$P_{\mathrm{cgc}}(m_0|m,V) = \mathrm{Bi}(m_0|m, V_0/V), \tag{5.1}$$

since this is the discrete probability distribution for finding m_0 transcripts inside the initial volume V_0 given the number of transcripts m present in the total volume V with success rate $p = V_0/V$ (see Figure 5.2 A). In the limit of high mRNA transcript numbers the binomial distribution tends to a normal distribution. In this limit cell growth correction corresponds to scaling the measured number of mRNA transcripts m with the normalized volume of the cell V_0/V. Therefore, the volume-dependent cell growth correction probability, Equation (5.1), contains the commonly performed scaling correction in the limit of high mRNA transcript numbers.

If the single cell's volume V and its initial volume V_0 are measured, we can evaluate $P_{\mathrm{cgc}}(m_0|m,V)$ directly. In many experimental applications (such as qPCR), however, measuring each single cell's volume is not performed or impossible. In this case, we treat the volume as a latent variable and marginalize over it to obtain the cell growth correction probability

$$P_{\mathrm{cgc}}(m_0|m) = \int \mathrm{d}V \mathcal{P}_{\mathrm{cgc}}(m_0|m,V)P(V). \tag{5.2}$$

To evaluate this we require the probability distribution of the cells' volumes $P(V)$ (i.e. the volume distribution over the cell population). This may be determined experimentally, or we can use generative models to simulate $P(V)$ computationally. In the following we use a linear growth model to generate $P(V)$.

There are two common growth scenarios in literature, linear and exponential growth (Mitchison, 2003, 2005), which both can be used for our method. Here, we use a linear growth model, which has been reported to be appropriate for rat Schwann cells (Conlon and Raff, 2003) to computationally simulate the distribution of cell volumes. Moreover the linear growth model does not depend on additional parameters. The cell cycle time τ indicates the time-point of a single cell during its cell cycle and is normalized to the cell cycle length, such that $0 \leq \tau \leq 1$. For linear growth, the cell increases its mean volume from its initial value $\langle V(\tau = 0)\rangle = 1$ (without loss of generality, we will set $V_0 = 1$) at the beginning of cell cycle according to $\langle V(\tau)\rangle = 1 + \tau$ and will have doubled its mean volume at the end of cell cycle. The conditional probability distribution $P(V|\tau)$ of the cells' volume V given its cell cycle time τ is generated by a

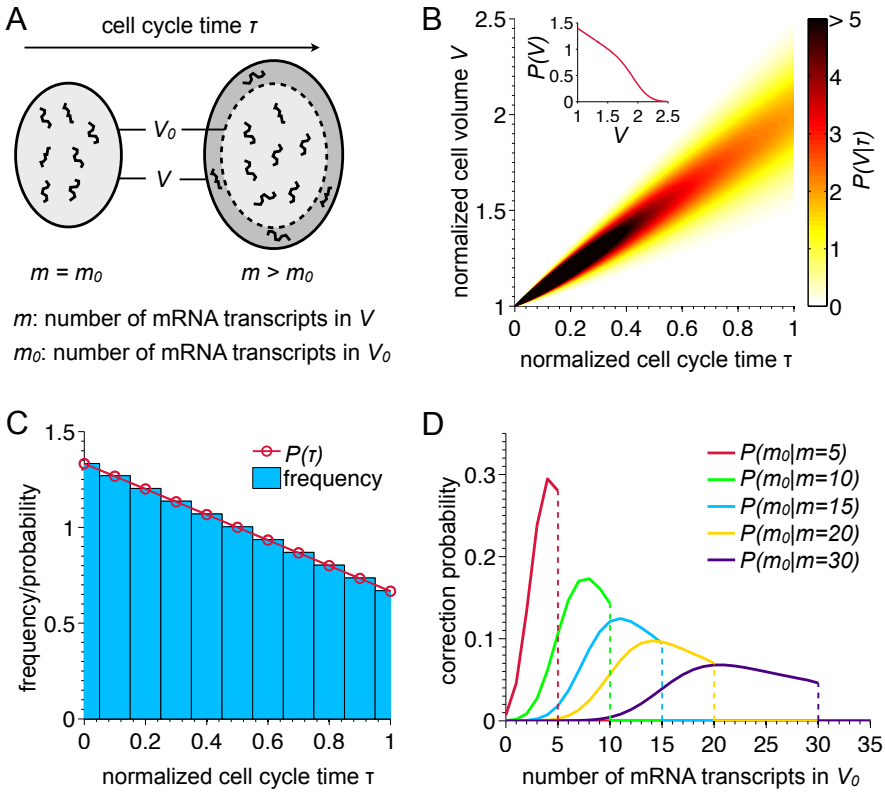


Figure 5.2: Cell growth model and correction probability (Figure legend on next page).

Gaussian density

$$P(V|\tau) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(V - \langle V(\tau)\rangle)^2}{2\sigma(\tau)^2}\right), \tag{5.3}$$

where we used a standard deviation that increases linearly with the cell cycle time $\sigma(\tau) = \sigma_0 \cdot \tau$. The resulting conditional probability distribution $P(V|\tau)$ is displayed in Figure 5.2 B. The marginal probability distribution yields

$$P(V) = \int \mathrm{d}\tau P(V|\tau)P(\tau). \tag{5.4}$$

To determine the probability of the cells' cell cycle time $P(\tau)$, we simulated entirely asynchronous cells that proliferate with known cell cycle time τ (see Figure 5.2 C). The resulting probability of the cells' volume $P(V)$ is displayed in the inset of Figure 5.2 B. The cell growth correction probability $P_{\mathrm{cgc}}(m_0|m)$ (Eq. 5.2) for this growth model is displayed in Figure 5.2 D for several values of observed molecule numbers m .

Figure 5.2: (From previous page). Cell growth model and correction probability. (A) During cell cycle, a cell will increase its initial volume V_0 to $V > V_0$. We assume the cell to increase its molecular content accordingly keeping the concentration of mRNA transcripts constant. Then the number of mRNA transcripts m measured at a latter time in cell cycle ($\tau > 0$) is greater than the number of mRNA transcripts m_0 at the beginning of the cell cycle ($\tau = 0$). To render measured mRNA transcript numbers m from single cells at different time points in their cell cycle comparable, we calculate the cell growth correction probability $P_{\text{cgc}}(m_0|m, V)$. This is the probability to find m_0 transcripts within the cell's initial volume V_0 (light grey area) given the cell's current volume V and the measured number of transcripts m. (B) Linear generative growth model. The colorbar indicates the probability $P(V|\tau)$ for the cell to be found with a particular volume V given the cell's time during cell cycle τ. For each time point τ the probability distribution is given as a Gaussian with a variance that increases during cell cycle. On average the cell doubles its volume during one cell cycle. Inset: Marginalized probability distribution for the cells' volume $P(V)$ when integrating over the the probability of the cells' cell cycle time $P(\tau)$. (C) Probability of the cells' cell cycle time(red line) that was determined by simulating a population of totally asynchronous cells with known cell cycle time τ and randomly picking cells out of this population. The obtained histogram of their cell cycle times is depicted with blue bars. (D) Cell growth correction probability $P_{\text{cgc}}(m_0|m)$ obtained after marginalizing over a linear growth-model (see text and Supplementary Figure S1) for several values of measured mRNA transcript numbers m. Notice that $P_{\text{cgc}}(m_0|m) = 0$ for $m_0 > m$ resulting in the displayed discontinuities.

cgCorrect for normalization of data sets

The cell growth correction probability $P_{\text{cgc}}(m_0|m)$ can be used to correct measured mRNA transcript numbers m directly to cell growth independent mRNA transcript numbers m_0^* by determining its mode

$$m_0^* = \arg\max_{m_0} P_{\text{cgc}}(m_0|m). \tag{5.5}$$

For instance, measuring $m = 15$ transcript numbers in a single cell, the most likely value for the transcript number, which we corrected for differences in cell size is $m_0^* = 11$ (see blue line in Figure 5.2 D). This approach offers a rank-conserving, one-to-one correspondence between measured and cell growth corrected mRNA transcript numbers, as needed for normalization of a data set.

cgCorrect for steady-state distribution analysis of gene expression

When using point estimates (such as the mode of a probability distribution) many alternative mRNA transcript numbers m_0 with non-negligible probability are ignored (see Figure 5.2 B). However, we can also exploit the full distribution of the correction probability $P_{\text{cgc}}(m_0|m)$: The number of mRNA transcripts of a particular gene is measured in many single cells. This yields a set of measured mRNA transcript numbers, which we use to obtain the steady-state probability distribution $P(m)$ of measured mRNA transcript numbers of this gene. We then sum over the correction probability of all measured transcript numbers m multiplied by the steady-state probability distribution to gain the cell growth corrected steady-state distribution

$$P_{\text{cgc}}(m_0) = \sum_m P_{\text{cgc}}(m_0|m)P(m). \tag{5.6}$$

The correction probability can also be used to account for differences in cell size when performing parameter estimation and model selection. Given the mRNA transcript numbers m of a particular gene from several single cells, the likelihood $P(m|\boldsymbol{\theta})$ for the kinetic parameters $\boldsymbol{\theta}$ of the underlying gene expression mechanism can be calculated at steady-state (see Section 3.4). Neglecting differences in cell size, however, can lead to incorrect parameter estimation and identification of the underlying gene expression mechanism (as already demonstrated in Figure 5.1 C) and has not been considered within this context so far.

Using the correction probability, it is straightforward to incorporate cell growth correction into the existing framework,

$$P_{\text{cgc}}(m|\boldsymbol{\theta}) = \sum_{m_0} P_{\text{cgc}}(m|m_0)P(m_0|\boldsymbol{\theta}), \tag{5.7}$$

allowing us to obtain the likelihood for the measured mRNA transcript numbers m from cells that differ in cell size given the parameters $\boldsymbol{\theta}$ of the gene expression mechanism under consideration. To obtain $P_{\text{cgc}}(m|m_0)$ from the correction probability $P_{\text{cgc}}(m_0|m)$, we use Bayes' theorem with uniform prior on the measured transcript numbers numbers m.

In case of simple gene expression where the steady-state distribution is given by a Poisson distribution there is one kinetic parameter: The mean expression level among all cells $\lambda = \nu/\gamma$. In case of bursty gene expression where the steady-state distribution is given by a negative binomial distribution there are two kinetic parameters: The

burst size $\xi = \nu/k_{\text{off}}$ and the burst frequency $\kappa_{\text{on}} = k_{\text{on}}/\gamma$ (see Figure 5.1). The model parameters can then be found via maximum likelihood estimation (MLE) $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} P_{\text{cgc}}(m|\boldsymbol{\theta})$.

We evaluate if the parameters of both gene expression mechanisms are identifiable and therefore capable of describing the data by calculating their profile likelihoods. If both the parameters of both mechanisms are identifiable, we perform model selection using the Bayesian information criterion (BIC) to select between the simple and bursty mechanism. In case this test is inconclusive ($\Delta\text{BIC} \leq 10$), we call the underlying gene expression mechanism inconclusive due to model selection (see Section 4.1 for details on parameter estimation and model selection).

cgCorrect and technical noise correction

In general, cell growth correction can also be combined with technical noise correction. To incorporate technical noise correction into the likelihood, Equation (5.7), the technical noise has to be measured in the experiment (e.g. with external spike-in controls) and the probability distribution of the technical noise $P_{\text{tn}}(m|m_{\text{t}})$ has to be determined experimentally. This is the conditional probability for the number of mRNA transcripts m that would be measured without technical noise given the number of mRNA transcripts m_{t} that are measured and are subjected to technical noise. The likelihood of cell growth correction and technical noise correction can then be calculated as

$$P_{\text{cgc,tn}}(m_{\text{t}}|\boldsymbol{\theta}) = \sum_{m} P_{\text{tn}}(m_{\text{t}}|m) P_{\text{cgc}}(m|\boldsymbol{\theta}). \tag{5.8}$$

To obtain $P_{\text{tn}}(m_{\text{t}}|m)$ from the probability distribution of the technical noise $P_{\text{tn}}(m|m_{\text{t}})$, Bayes' theorem can be applied with uniform prior on the measured mRNA transcript numbers with technical noise m_{t}.

Validation of cgCorrect on simulated gene expression data

To validate cgCorrect, we applied it to mRNA transcript numbers that we simulated from the simple gene expression mechanism. We generated mRNA transcript numbers m of 100 single cells with different cell sizes. We started by simulating the cell's volumes by the inverse transform method (Gentle, 2004) such that the cell's volumes are distributed according to the probability we obtained by the generative model for cell growth $V \sim P(V)$ (see Figure 5.2 B). The measured mRNA transcript number m

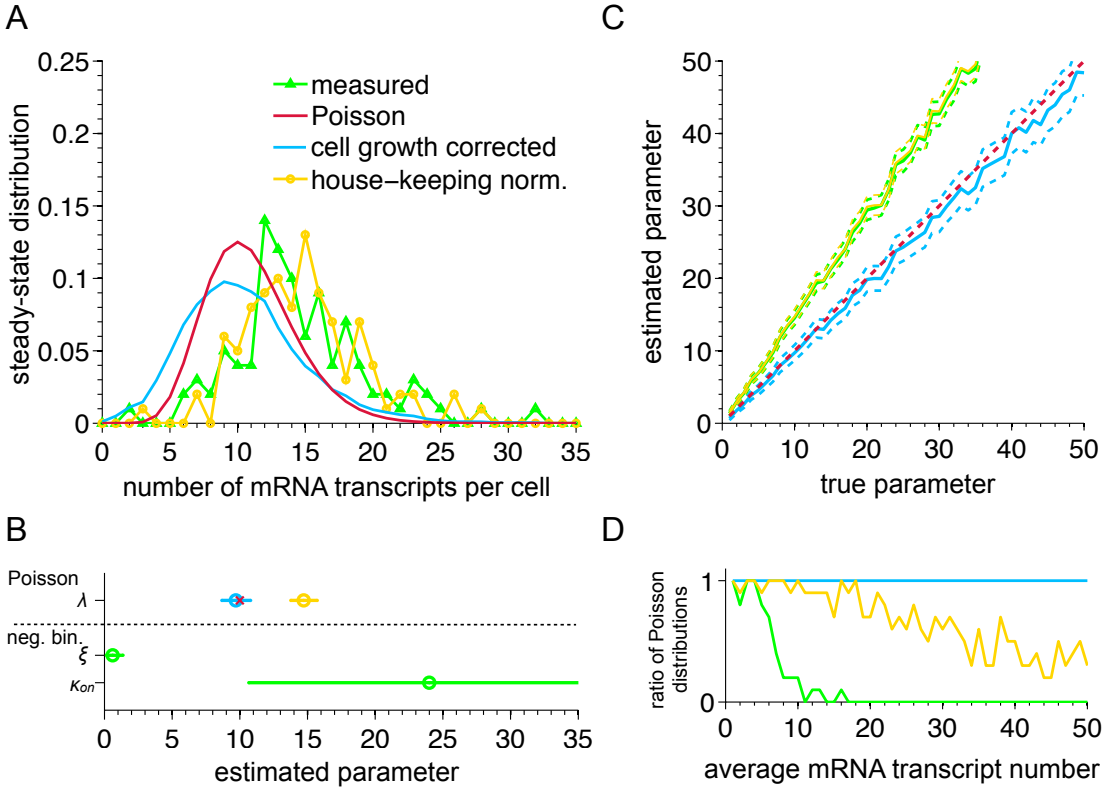


Figure 5.3: Cell growth correction of simulated gene expression data leads to the correct identification of parameters and the underlying gene expression mechanism (Figure legend on next page).

for a cell with volume V was then simulated by randomly drawing from the Poisson-distribution $m \sim \text{Pois}(m|\lambda)$ with $\lambda = V \cdot \lambda_0$, where λ_0 indicates the average mRNA transcript number per cell if the cell's did not grow and not differ in size. During cell growth we assume the transcription rate ν ($\lambda = \nu/\gamma$) to increase proportional to the cell's volume V. This particular choice of the parameter λ ensures that the concentration of mRNA transcripts remains constant during cell growth, whereas the number of mRNA transcripts increases and reflects the differences in cell size.

Without differences in cell size the mRNA transcript numbers would be Poisson-distributed $m_0 \sim \text{Pois}(m_0|\lambda_0)$ with the average number of mRNA transcripts per cell chosen to be $\lambda_0 = 10$ (see red line in Figure 5.1 C and 5.3 A). Due to differences in cell size the steady-state distribution of measured mRNA transcript numbers $P(m)$ is shifted towards higher transcript numbers (green line in Figure 5.3 A). We can correct for latent differences in cell size by calculating the corrected steady-state distribution

Figure 5.3: (From previous page). Cell growth correction of simulated gene expression data leads to the correct identification of parameters and the underlying gene expression mechanism. (A) Steady-state probability distribution of the measured mRNA transcript numbers m (green line) and the cell growth corrected mRNA transcript numbers m_0 (blue line). The underlying gene expression mechanism in the absence of differences in cell size is given by a Poisson distribution with an average mRNA molecule number per cell $\lambda_0 = 10$ (red line). The cell growth corrected probability distribution resembles the Poisson distribution closer than the house-keeping normalized distribution (yellow line). (B) Estimated parameters and identified models for cell growth corrected, house-keeping normalized and measured mRNA transcript numbers. The gene expression mechanism is identified to be simple for house-keeping normalization and cell growth correction; for the measured data it is identified to be bursty. The simple mechanism is governed by one parameter: the mean expression level among all cells $\lambda = \nu/\gamma$. The bursty mechanism is governed by two parameters: the burst size $\xi = \nu/k_{\text{off}}$ and the burst frequency $\kappa_{\text{on}} = k_{\text{on}}/\gamma$. Only cell growth correction is in the range of the true parameter (red x). Error bars indicate 0.99 confidence intervals of the estimated parameters. (C) To explore the parameter range we performed parameter estimation and model selection for several values of the true kinetic parameter λ_0. We find that cell growth correction (blue line) is capable of correctly inferring the true parameter (red dashed line) for the whole parameter range whereas inferring the parameter on the measured (green line) and the house-keeping normalized mRNA transcript numbers (yellow line) fails. (D) Ratio of gene expression mechanisms that were identified to be simple for 10 independently simulated data sets. Model selection (based on the BIC) on the cell growth corrected data (blue line) identifies the true gene expression mechanism correctly over the whole parameter range, in contrast to model selection on the measured data (green line) and on the house-keeping normalized data (yellow line).

of transcript numbers $P_{\text{cgc}}(m_0)$, Equation (5.6) (blue line in Figure 5.3 A). Since we ignored the cell's volumes by marginalizing the volume out (cf. Equation (5.2)) the corrected steady-state distribution of transcript numbers does not entirely coincide with the Poisson distribution but has slightly larger tails. To compare cgCorrect with conventional house-keeping normalization, we scaled the measured number of transcripts m with the transcript number of an additionally simulated house-keeping gene m_{hk}, which we chose to have an average number of transcripts $\lambda_{0,\text{hk}} = 100$ (yellow line in Figure 5.3 A). Visual comparison of the two normalization strategies shows that cell growth correction for normalization outperforms house-keeping normalization for this data set.

Model selection between the simple and bursty gene expression mechanism reports

very strong evidence that the measured steady-state distribution of mRNA transcript numbers is bursty. When correcting for cell growth, model selection correctly chooses the simple expression mechanism. Performing parameter estimation the true gene expression parameter can only be inferred when using cgCorrect (see Figure 5.3 B), confirming that cgCorrect outperforms house-keeping normalization in recovering the true underlying distribution. To test cgCorrect for a broad parameter range we simulated additional mRNA data sets for several average numbers of mRNA transcripts per cell (Figure 5.3 C and D): Only when we apply cgCorrect we are able to infer the underlying gene expression mechanism and its parameters for the whole parameter range correctly.

Moreover, we verified that after applying cgCorrect on transcript numbers that were simulated from the negative binomial distribution the inferred steady-state distribution is negative binomial: To this end, we simulated mRNA transcript numbers from the negative binomial distribution $m_0 \sim \mathrm{NB}(m_0|\xi, \kappa_{\mathrm{on}})$ for a wide range of average numbers of mRNA transcripts $\langle m_0 \rangle = \xi \cdot \kappa_{\mathrm{on}}$. When applying cgCorrect we found that model selection for $m_0\rangle \geq 3$ correctly identifies the underlying steady-state distribution to be negative binomial. For very small average numbers of mRNA transcripts $\langle m_0 \rangle \leq 2$ the obtained distribution of transcript numbers is very narrow and we find cases (20% for $\langle m_0 \rangle = 2$ and 90% for $\langle m_0 \rangle = 1$), where the underlying steady-state distribution is identified to be Poisson (data not shown). In summary, cgCorrect is capable of both, successfully inferring the underlying system parameters from the simulated, cell growth dependent transcript numbers and correctly specifying the steady-state distribution of transcript numbers.

5.3 Application of cgCorrect to biological data

cgCorrect on qPCR data suggests that many genes rather follow the simple than the bursty gene expression mechanism

To reveal the gene expression mechanism during hematopoiesis, we applied cgCorrect to a recently published single-cell qPCR data set of hematopoietic stem and progenitor (HSP) cells (Moignard et al., 2013). In this experiment 18 transcripts of key hematopoietic genes (and six additional transcripts of house-keeping genes) were measured in 597 single cells of five different HSP cell types. To transform the measured data from ct-values into discrete numbers of mRNA transcript we use results from digital qPCR (Warren et al., 2006), where the discrete number of one of the 18 transcripts, PU.1, was

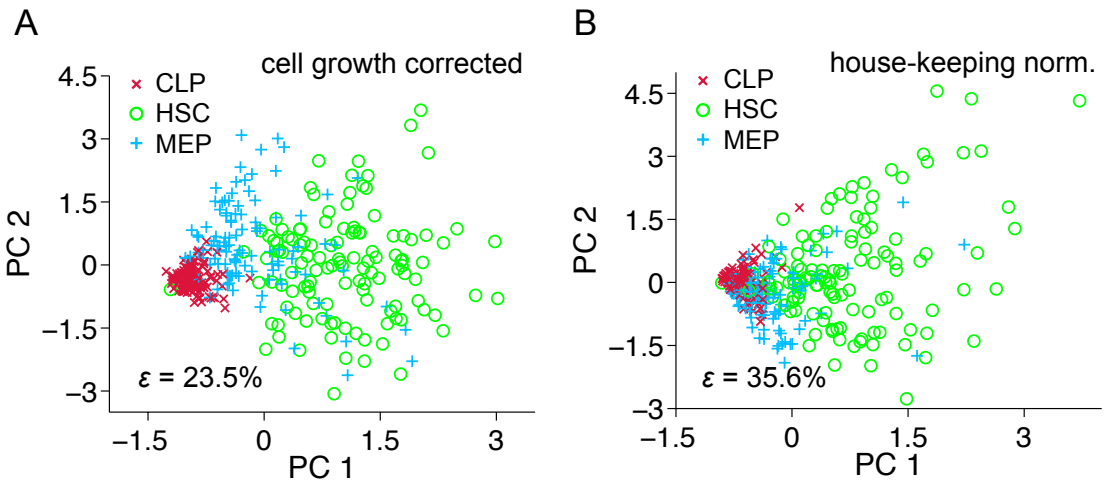


Figure 5.4: Probabilistic principal component analysis (PPCA) of single-cell qPCR data resolves hematopoietic sub-populations better when using cell growth correction. (A) PPCA of cell growth corrected and (B) PPCA of house-keeping normalized single-cell qPCR data of 18 transcripts. The nearest neighbor error ϵ decreases by 12.1% when using cell growth correction compared to house-keeping normalization.

measured for hematopoietic stem cells (HSCs), common lymphoid progenitors (CLPs) and common myeloid progenitors (CMPs), all of them found among the HSPs (see Supplementary Material S6 for details on the data pre-processing).

Since in this experiment neither technical noise nor information about the cells' volume was measured we apply cgCorrect without technical noise correction and with marginalized volume (Equation 5.2). To compare cgCorrect with conventional house-keeping normalization we normalized the data set with the house-keeping genes *Ubc* and *Polr2a* as described by Moignard et al. (2013). cgCorrect is better suitable to resolve distinct cell types than house-keeping normalization, as can be visualized by a probabilistic principal component analysis (PPCA) (see Figure 5.4 and Section 4.3): The nearest neighbor error of finding two differing cell-types next to each other is decreased by 12.1%.

Applying cgCorrect to all measured mRNA transcripts, we find that 18 out of 54 ($\sim$ 33.0%) gene/cell type combinations can be explained by simple rather than by bursty gene expression, whereas this is the case for only 3 out of 54 ($\sim$ 5.6%) without cgCorrect (see Figure 5.5 C).

To further illustrate the effect of cgCorrect, we focus on one particular transcript, PU.1 in one cell type (CLP) (see Figure 5.5 A). We analyze the Fano factor $\mathcal{F}$ =

σ^2/μ defined as the ratio between the variance σ^2 and the mean μ of the steady-state distribution of mRNA transcript numbers. The Fano factor is a key parameter to quantify deviations from a Poisson distribution (Munsky et al., 2012) and it equals 1 if the values are Poisson-distributed. cgCorrect alters the Fano factor from $\mathcal{F}(m) = 2.29$ for the measured PU.1 transcript numbers to $\mathcal{F}(m_0^*) = 1.32$. Parameter estimation for the measured and the corrected transcript numbers is depicted in Figure 5.5 B. Model selection between the simple and bursty gene expression mechanism shifts strong evidence for PU.1 expression in CLP from bursty without correction to simple with cgCorrect.

5.4 Discussion

In this work we present cgCorrect, a statistical method for the correction of latent differences in cell size. We show that differences in cell size may lead to an over-dispersed steady-state distribution of transcript numbers, which may be misleadingly interpreted in a computational analysis. cgCorrect can be used for data normalization before visualization as well as for a steady-state distribution analysis of the data. It can incorporate information about the cell size on different levels: (i) If the size of each cell or an estimator for the size is known, we can use this information to obtain the volume-dependent cell growth correction probability. (ii) If only the probability distribution of the cells' volume among the whole population is known we can use this distribution to marginalize the volume out. (iii) If there is a total lack of information about the cells' volume (as is typically the case for qPCR data), we can use generative growth models to simulate the cells' volume distribution computationally and use this for marginalization. Moreover, we showed how cgCorrect can in principle be combined with the correction of technical noise, if the technical noise of the experiment is measured.

We validated cgCorrect on simulated mRNA data, where we could show that it is only possible to infer the true steady-state distribution and its parameters when cgCorrect was applied. To show that cgCorrect is generally applicable and independent of the experimental setup that was used to measure the data it was applied on transcriptomics data from single-cell qPCR. Analyzing steady-state distributions of transcript numbers from the qPCR data set we found that cgCorrect changed the identified steady-state distribution in 27.4% of the measured cell/gene combinations in HSPs from an over-dispersed negative binomial distribution to the Poisson distribution.

In contrast to conventional normalization techniques cgCorrect takes the discreteness

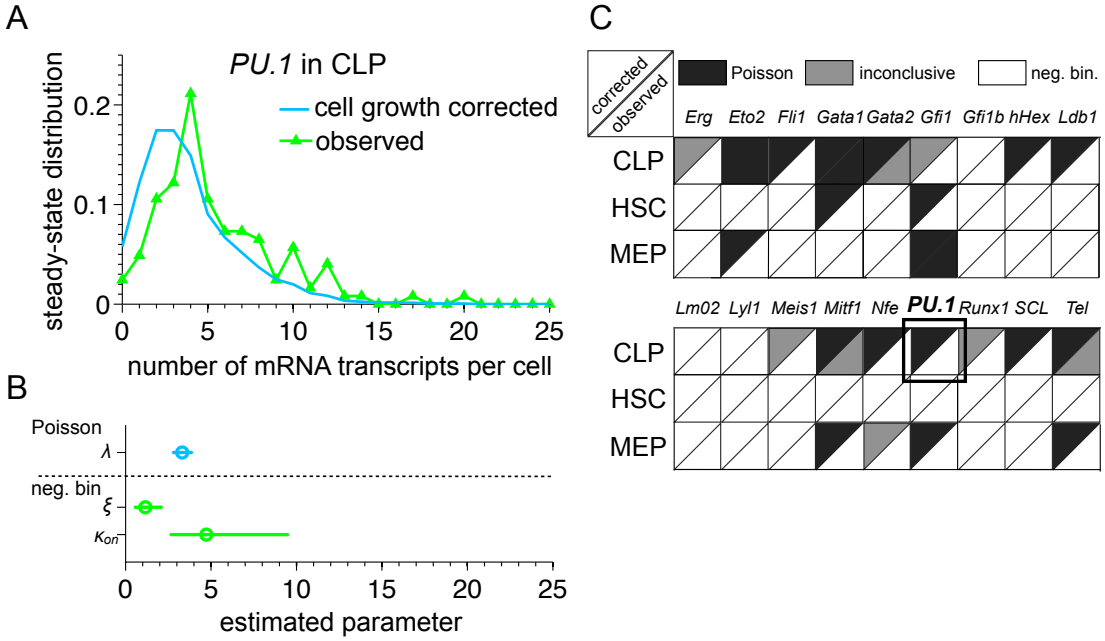


Figure 5.5: Parameter estimation and model selection on cell growth corrected single-cell qPCR data reveals that 15 out of the 56 hematopoiesis genes are more likely to origin from simple gene expression than from transcriptional bursting. (A) Probability density of corrected (blue) and measured (green) PU.1 transcript numbers within CLPs. cgCorrect renders the observed distribution narrower. (B) Estimated kinetic parameters and identified gene expression mechanism for PU.1 mRNA in CLPs. After cell growth correction (blue) the gene expression mechanism is identified to be simple whereas without correction (green) it is bursty. (C) The result of model selection between the simple (black) and bursty (white) gene expression mechanism is visualized. Inconclusive model selection is indicated in grey. The lower right triangle indicates the identified mechanism of the measured transcript numbers, the upper left triangle of the corrected transcript numbers. The identified gene expression mechanism is altered in 20 cases after having performed cgCorrect.

of mRNA transcript numbers into account. For the analyzed qPCR data set we showed that cgCorrect outperforms traditional house-keeping gene normalization resulting in a better separation of known cell types in a principal component analysis (PCA). House-keeping genes underlie stochastic gene expression themselves and may therefore not suit as reliable reporters for cell size.

In previous analysis the steady-state distribution of a gene is used to interpret its gene expression mechanism (Kim and Marioni, 2013; Larson, 2011; Raj et al., 2006; Shahrezaei and Swain, 2008). The Poisson steady-state distribution corresponds to the

simple gene expression mechanism and the negative binomial distribution corresponds to the bursty gene expression mechanism. However, there are several assumptions involved that are important to consider for this interpretation.

First, it is assumed that the reaction rates that govern the gene expression mechanism remain constant during cell cycle. Here, we do not consider transcriptional changes during cell cycle that may alter the reaction rates and have been reported to affect the measured number of mRNA transcripts (Bertoli et al., 2013; Zopf et al., 2013). In order to assess the effect of cell cycle specific gene expression, we modeled transcriptional changes of the reaction rates by an activation function reaching its maximum in the S phase of cell cycle.The resulting steady-state distribution is identified to follow the over-dispersed, negative binomial distribution and would therefore be interpreted to origin from the bursty gene expression mechanism both with and without applying cgCorrect (data not shown). Cell cycle specific gene expression corresponds to a highly orchestrated on and off switching of the promoter region. For a sample of unsynchronized cells that are pooled together, however, the resulting steady-state distribution of mRNA transcript numbers exhibits over-dispersion.

The second assumption that is made when analyzing steady-state distribution of mRNA transcript numbers is that the kinetic parameters that govern gene expression are equal for all cells of the same cell type (Kim and Marioni, 2013; Raj and van Oudenaarden, 2008; Shahrezaei and Swain, 2008; Thattei and van Oudenaarden, 2001), which does not necessarily have to reflect the biological reality. We tested the effect on the steady-state distribution analysis when neglecting this assumption by simulating mRNA transcript numbers from a cell population with varying transcription rates expressing mRNAs with the simple mechanism and showed that this effect can also lead to over-dispersed steady-state distributions (data not shown). A final conclusion on the gene expression mechanism cannot be made based on steady-state distributions of gene expression alone but needs techniques that also allow for spatial resolution such as fluorescence in situ hybridization (FiSH) (Battich et al., 2013; Raj et al., 2006).

Finally, we made assumptions concerning the cell growth parameters for the generative growth model that we used to obtain the correction probability. The question whether mammalian cells grow linearly or exponentially is still under debate (Cooper, 2004; Popescu et al., 2014). Here, we used a linear growth model, which has been reported to be appropriate for rat Schwann cells (Conlon and Raff, 2003) to computationally simulate the distribution of cell volumes. Moreover, we performed a sensitivity analysis (data not shown) that investigates the effect of different linear cell growth scenarios on

the correction probability and indicates that our findings are robust with respect to the growth scenario. As already discussed, cgCorrect does not rely on a generative growth model as it allows to include additional information on either each single cell's volume or the distribution of the cells' volume, if they are measured.

To summarize, we identified differences in cell size of proliferating cells to be a latent cause of confounding variability. We introduced cgCorrect, a statistical method that is capable to correct for this confounding cell cycle effect in gene expression data, which can be used for data normalization, parameter estimation and model selection. We validated cgCorrect on a simulated data set and applied it to single-cell qPCR gene expression data (Moignard et al., 2013) from mouse HSPs where we could show that the interpretation of the underlying gene expression mechanism could be influenced by cell growth effects.

Chapter 6

Model comparison between histone acetylation scenarios reveals motif-specificity

In Section 2.1 we discuss how histone modifications influence transcription initiation and how they therefore serve as an important ingredient for the emergence of cellular heterogeneity. In this Chapter our focus lies on histone acetylation: so far high levels of histone acetylation have mainly been associated with an increased transcriptional activity whereas low levels of histone acetylation have been associated with gene silencing (see e.g. Grunstein (1997)).

It only recently became possible to experimentally resolve the combinatorial patterns of histone acetylation using a novel approach in liquid chromatography mass spectrometry (LC-MS; Feller et al. (2015)). However, the complexity of the new data made it unfeasible to analyze it with standard means and called for a new bio-mathematical and computational framework.

The question we answer with our new approach is if combinatorial modification patterns ('motifs') are specifically set or if they arise merely by random events (Section 6.1). In Section 6.2 we formulate the likelihood (see Section 4.1) for all possible acetylation scenarios that we describe using the reaction rate equation (RRE; see Section 3.3) and compare the models using the Bayesian information criterion (BIC; see Section 5.1). Our new mathematical framework reveals that histone acetylation is highly motif-specific (Section 6.3). Moreover we can use our findings for the prediction of acetylation path-

ways (Section 6.4), which we subsequently validate qualitatively (Section 6.5) using an independent data set where all enzymes that are known to be involved in the acetylation of histones were systematically depleted (Feller et al., 2015).

The data we analyze here is from untargeted mass spectrometry measurements on bulk samples. As detailed in Section 2.3 it is not possible to observe cell to cell variability with bulk data on a single cell level. The mechanism, histone acetylation, we study in this Chapter, however, allows cells to regulate transcription initiation and is therefore a key for the establishment of cellular heterogeneity that – as we show here – can also be analyzed with bulk data. While targeted mass spectrometry approaches where antibodies with unique isotope masses are attached to different protein species has become widely applied for single cell measurements (see e.g. Bjornson et al. (2013)), the ability to quantify the abundance of proteins in single cells with an untargeted approach (as it was chosen for the bulk measurements analyzed in this Chapter) has been reported only very recently (Lombard-Banek et al., 2016). We anticipate the quest for single cell based techniques to measure the abundances of histone modification states in single cells to be highly beneficial to further resolve cell to cell heterogeneities. Single cell techniques have already been attracting more and more attention in other fields of epigenetics (Bheda and Schneider, 2014), such as the study of DNA methylation, where single cell data can readily be obtained via bisulfite sequencing (see e.g. Farlik et al. (2015)).

This Chapter is based on and in parts identical with the following article:

Blasi, T., Feller, C., Feigelman, J., Hasenauer, J., Imhof, A., Theis, F. J., Becker, P. B. and Marr, C. (2016). Combinatorial histone acetylation patterns are generated by motif-specific reactions. *Cell Systems* 2:49–58.

The designed workflow and the used data is open-source and freely available on the publisher's homepage.

6.1 Biological background and problem statement

In the nucleosome, the fundamental structuring unit of DNA in eukaryotes, DNA winds around an octamer formed by four pairs of core histones, H2A, H2B, H3, and H4. Each histone has an N-terminal tail extending from the otherwise compact nucleosome (Figure 6.1 A). These tails are subject to a large number of post-translational modifications (PTMs) of fundamental importance for gene regulation and gene regulatory networks

(see Section 2.1 and Kouzarides (2007); Suganuma and Workman (2011); Tan et al. (2011)). One important type of histone PTM is lysine acetylation, which often prevents tight folding of the nucleosome fiber through charge neutralization, and thereby increases the accessibility of DNA to regulatory proteins Shahbazian and Grunstein (2007).

The histone N-terminal tails have many lysines that can be acetylated, and although some functions for acetylation at specific lysines ('sites') are known (Lucchesi and Kuroda, 2015; Shahbazian and Grunstein, 2007; Straub and Becker, 2011) it is assumed that acetylation at neighboring sites may serve largely redundant functions based on simple charge neutralization Allahverdi et al. (2011); Dion et al. (2005). However, the observation that nearby lysines may be acetylated in combinatorial patterns – forming 'motifs' – led to the hypothesis that such acetylation motifs may interact and lead to functions beyond those of their single acetylations (Jennuwein and Allis, 2001; Smith and Shilatifard, 2010; Strahl and Allis, 2000; Suganuma and Workman, 2011; Turner, 2000). We thus differentiate between 'sites' which denote acetylation of a single lysine residue independent of its adjacent modifications, and 'motifs' which denote lysine acetylation in the context of the modification state from neighboring lysines (e.g. the motif 'K5K12' refers to acetylations at lysines K5 and K12).

Understanding how acetylation motifs arise is key to elucidating their function. The prevailing paradigm of histone tail acetylation is that KATs target either a single site with high specificity (e.g. MOF acetylates unmodified histone H4 at lysine 16) or multiple residues with relaxed specificity (e.g. CBP for lysine residues K5, K8, and several sites on histone H3). Whether combinatorial motifs arise through the action of dedicated enzymes or are the result of independent, uncoordinated activities is poorly understood. The only well-characterized acetylation motif so far is K5K12, which is generated by the lysine acetyltransferase HAT1 (Parthun, 2012). *In vitro*, HAT1 was shown to serially acetylate the unmodified H4 tail peptide (referred to as 0ac) first at the lysine K5 site followed by K12 to yield the motif K5K12 (Benson et al., 2007; Makowski et al., 2001; Richman et al., 1988), although other modes have also been reported (Dose et al., 2011; Parthun et al., 1996; Sobel et al., 1994).

In order to evaluate whether motif-specific acetylation constitutes an inherent property of the histone acetylation system, i.e. a 'general design principle', Feller et al. (2015) recently began to study the network components at the system scale (including all KATs, KDACs, their regulators and all acetylation motifs. Key to the analysis was a liquid chromatography-mass spectrometry (LC-MS)-based workflow that enables the

precise and accurate quantification of most histone acetylation motifs, including the 16 motifs formed on the H4 tail by combinatorial acetylation of lysines K5, K8, K12 and K16 (Figures 6.1 A, B). They found that *Drosophila melanogaster* KC cells show substantially skewed motif abundances within layers as well as across layers (see Figure 6.1 B). They also assessed systematically perturbed cells by individually depleting all known and suspected KATs and KDACs expressed in these cells and quantifying the resulting changes in the histone acetylation distribution. The response of the system to depletion of individual enzymes provided insight into specific enzyme-substrate relationships but a deeper understanding was hindered by the complex interplay between the network components. Hence the principles underlying the regulation of acetylation motif abundances could not be deduced with standard analysis.

We surmise that a theoretical framework based on simple, biologically reasonable assumptions should be helpful for identifying the mechanisms giving rise to the observed motif abundances and therefore for identifying the underlying properties of the histone H4 acetylation system. In the past, computational modeling of genomic datasets was able to explain the emergence of heritable chromatin states that can lead to bistable gene expression (Angel et al., 2011; Dodd et al., 2007; Sneppen and Dodd, 2012; Zerihun et al., 2015) and gene silencing (Mukhopadhyay and Sengupta, 2013; Sedighi and Sengupta, 2007). Others reported turnover rates for histone modifications using metabolic labelling and LC-MS time-course strategies (Evertts et al., 2013; Zheng et al., 2012).

Here, we developed a computational framework to investigate the genesis of combinatorial histone acetylation motifs, and to assess the relative importance of dedicated synthesis pathways versus uncoordinated enzymatic activity. We trained our model on published data from unperturbed cells and validated predicted pathways using an independent KAT depletion dataset Feller et al. (2015). Our modeling strategy represents a novel approach for providing insight into the design principles of PTM motifs, which will become increasingly relevant as further LC-MS datasets become available.

6.2 A mathematical framework for modeling acetylation motif abundances

The 16 histone H4 acetylation isoforms (motifs) show different abundances within layers and across layers (Fig. 6.1 B, data from LCMS measurements from unperturbed Drosophila melanogaster KC cells). However, the general design principles underly-

ing the observed complex acetylation patterns remain elusive. We hypothesized that motif-specific reactions contribute to the skewed abundance distribution, where enzymes are sensitive to adjacent modifications (or their absence), and hence catalyze (de)acetylations only in the context of a distinct pre-modification state (Feller et al., 2015). To test this hypothesis, we assessed three hypothetical acetylation scenarios (Fig. 6.1 C):

1. Lysine acetylation could be unspecific, i.e. not dependent on the site or motif, and therefore be governed by a single, basal acetylation rate constant $\alpha_r = \alpha_b$ for all reactions r (Fig. 6.1 C, left).

2. Acetylation could be site-specific with some or all of the four lysine sites K5, K8, K12, K16 being acetylated by site-specific enzymes, resulting in a common acetylation rate for each reaction that targets that site independent of neighboring modifications (e.g. $\alpha_{K5} \neq \alpha_{K16} \neq \alpha_b$ in Fig. 6.1 C, middle).

3. Acetylation could be motif-specific with enzymes being sensitive to the modification state of nearby amino acids. In the depicted example, the acetylation rate at site K12 is different from the basal rate when K16 (but not K5 or K8) is already acetylated ($\alpha_{K16} \neq \alpha_b$; Fig. 6.1 C, right).

We thus develop different mathematical models encompassing these three acetylation scenarios. For each model, we predict the abundance of all acetylation motifs as a function of the acetylation rates and infer the most likely acetylation rates by fitting the predictions to the measured abundances (see Figure 6.1 D).

In the following, we use the symbol m to denote a motif (e.g. m = 0ac for the state with no acetylations, m = K5 for acetylated lysine 5 and no modifications at lysines 8, 12, 16, etc.) and xm for the relative abundance of m, for each of the 16 acetylation motifs. In our model we assume:

- Acetylation and deacetylation occurs stepwise, i.e. only one acetylation or deacetylation event can occur at a time. Thus, a motif can be generated via a single acetylation of a less acetylated state, or a single deacetylation of a more acetylated state (Figure 6.1 C).

- Deacetylation is unspecific, such that the rate constant is the same for all reactions r connecting two motifs, consistent with current views on deacetylation and measured response upon KDAC depletion in Drosophila cells (Feller et al., 2015; Seto and Yoshida, 2014).

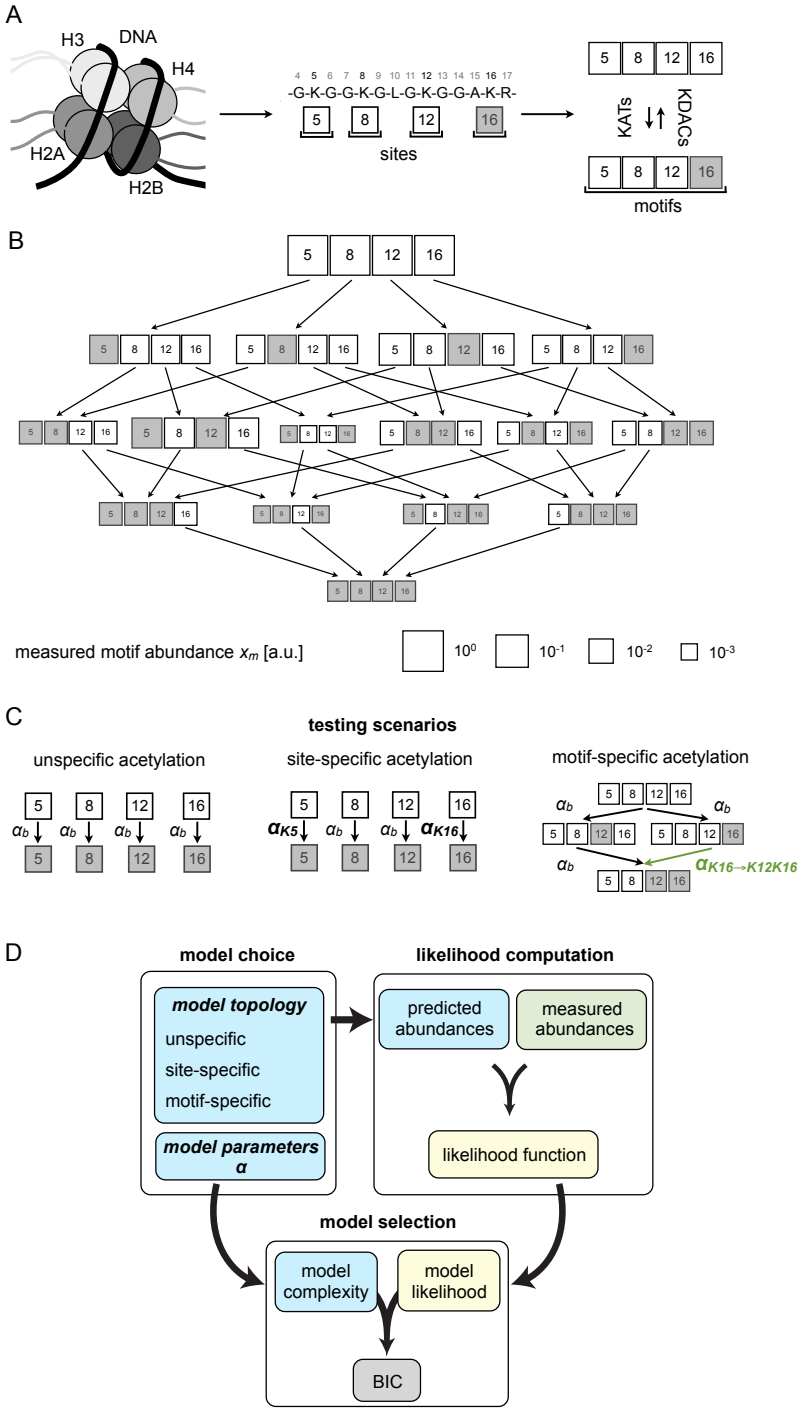


Figure 6.1: Overview on biological background and possible histone H4 acetylation scenarios (Figure legend on next page).

Figure 6.1: (From previous page). Overview on biological background and possible histone H4 acetylation scenarios. (A) Nucleosome with protruding N-terminal histone 'tail' domains (left), with four potential lysine (K) acetylation residues ('sites') on the histone H4 peptide G4R17 (center) which together form the histone H4 N-terminal acetylation 'motifs' (right). The motif K16 (right bottom) arises by acetylation of lysine 16 from the unmodified 0ac motif (right top). (De)acetylation reactions are mediated by lysine acetyltransferases (KATs) and lysine deacetylases (KDACs). Only two out of 16 possible acetylation motifs are shown (unmodified: white, acetylated: grey). (B) Skewed abundance distribution of the histone H4 acetylation motifs in *Drosophila* cells. Network representation of the abundances for the 16 histone H4 acetylation motifs as determined by liquid chromatography-mass spectrometry (LCMS) (data from Feller et al. (2015)). Box sizes depict log10 abundances. Note that the box size for K5K16 was set to scale with the lowest quantifiable H4 motif (K5K8K16), because the K5K16 motif is below the quantification limit (see Supplementary Experimental Procedures 1). (C) Testing hypothetical acetylation scenarios: unspecific acetylation scenario with the same unspecific, basal reaction rate α_b for all lysines (left); site-specific acetylation scenario with site-specific reaction rate(s) for individual site(s) (blue, center);and motif-specific acetylation scenario with rate(s) specific to the modification context of lysines 5, 8, 12 and 16. In the depicted case, acetylation of lysine 12 requires pre-acetylated lysine 16 (green). (D) Outline of computational modeling framework. All models are based on mass action kinetics with rate constants α_r for the r reactions. The models are fitted to the LC-MS data from (B) via maximum likelihood estimation (MLE). To find the model with an optimal trade-off between complexity and goodness of fit, we perform model selection based on the Bayesian Information Criterion (BIC), a score which penalizes more complex models. Applying this approach allows to test the different acetylation scenarios shown in (C) with a rigorous quantitative assessment. This Figure and its legend is adopted from Figure 1 of the author's following publication Blasi et al. (2016a).

- Histone H4 acetylation follows mass action kinetics: the rate of each acetylation or deacetylation reaction is proportional to the abundance of the substrate of that reaction.

- The measured motif abundances are assumed to be in steady state. This is consistent with the fact that acetylation and deacetylation progresses much faster than the cell cycle and histone protein turnover (Katan-Khaykovich and Struhl, 2002; Toyama et al., 2013).

See the next subsection for more details.

For the site-specific and/or motif-specific scenarios, it is unknown which of the sites or motifs are acetylated with a non-basal rate, hence we consider models containing any combination of site-specific reactions or any combination of motif-specific reactions. In case of site-specific acetylation there are 11 different combinations of site-specific and basal acetylation rates: 4 possibilities with one site-specific rate, 6 possibilities with two site-specific rates, and one possibility where every site has a specific acetylation rate. For the motif-specific acetylation scenario, there are a total of $2^{32} = 4.295 \times 10^9$ possible combinations of motif-specific reactions among the 32 acetylation reactions. In each case, we formulate the likelihood of the model given the data and optimize the model parameters to obtain the best fit, given by the maximum likelihood estimator (see Section 4.1). To ensure identifiability of the model parameters we calculate the profile likelihoods. We then compare all models using the Bayesian Information Criterion (BIC) based on their likelihood. This allows us to penalize highly complex models and thereby prevents overfitting, thus permitting the identification of the simplest models capable of describing the data.

Reaction rate equation of the acetylation network

To model the reaction network of histone H4 tail acetylation and deacetylation we use a stepwise reaction network (Fig. 6.1 B) wherein only a single acetyl group can be added or removed at a time. We use the reaction rate equation (RRE; Section 3.3) to describe the acetylation and deacetylation reactions with acetylation rate constants a_r and deacetylation rate constants d_r. The change of the abundances of a particular motif x_m is given by the difference between influx from and the outflow to neighboring motif abundances

$$\frac{\mathrm{d}x_m}{\mathrm{d}t} = \sum_{m'} R_{m,m'}(a_r, d_r) \cdot x_{m'} \tag{6.1}$$

The reaction matrix $R_{m,m'}$ incorporates all possible reactions between the motifs within the stepwise reaction network. For example, the change of the abundance of motif K8

$$\frac{\mathrm{d}x_{K8}}{\mathrm{d}t} = a_{K8}x_{0ac} + d_{K5}x_{K5K8} + d_{K12}x_{K8K12} + d_{K16}x_{K8K16} \tag{6.2}$$

$$-(a_{K5} + a_{K12} + a_{K16} + d_{K8})x_{K8} \tag{6.3}$$

is given as the difference between the influx from (terms with positive sign) and the outflow to (terms with negative sign) neighboring motifs (i.e. motifs differing by exactly one acetyl group).

It was previously found that deacetylation occurs broadly with only little evidence for motif-specificity in the analyzed cell system (Feller et al., 2015). Thus, we additionally assume that deacetylation is unspecific, and therefore proceeds with the same rate constant d for all motifs. We then simplify by dividing Eq. 6.1 by the deacetylation rate constant d , thus converting all acetylation rate constants a_r to rescaled, effective rate constants $\alpha_r = a_r/d_r$.

The time scale of acetylation and deacetylation is much faster than the cell cycle and measured histone protein half-lifes. Therefore we assume the reactions to be at steady state

$$\frac{\mathrm{d}x_m}{\mathrm{d}t} = 0. \tag{6.4}$$

To take into account that the abundances of the motifs are not measured in total numbers but are given as fractions that are normalized to one, we solve for the steady states subjected to the constraint of unity total abundance, i.e. we solve

$$\sum_{m'} R_{m,m'}(a_r, d_r) \cdot x_{m'} = 0 \quad s.t. \quad \sum_m x_m = 1. \tag{6.5}$$

Likelihood for the acetylation network

In order to fit our model to the measured static dataset (Feller et al., 2015) we need to define the likelihood function (Section 4.1). By using an error model with the differences between the logarithms of estimated and measured motif abundances being normally distributed we ensure that low and high motif abundances (the measured abundances vary over three orders of magnitude) enter the likelihood with equal weights (Kreutz et al., 2007). This choice of the error model corresponds to a multiplicative lognormal distributed error. Given the data D of measured abundances $x_{m,i}$ of all motifs m and biological replicates i we can calculate the log-likelihood $\mathcal{J}(\boldsymbol{\alpha}) = \log \mathcal{P}(\mathcal{D}|\boldsymbol{\alpha})$ for all effective reaction rate constants $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{32})$

$$\mathcal{J}(\boldsymbol{\alpha}) = -\frac{1}{2}\left(n \log(2\pi\sigma^2) + \sum_{m,i} \frac{(\log x_{m,i} - \log x_m(\boldsymbol{\alpha}))^2}{\sigma^2} \right). \tag{6.6}$$

Here $n = \sum_{m,i} 1$ denotes the number of measured data points (including all motifs and replicates) and

$$\sigma^2 = \frac{1}{n-1} \sum_{m,i} (\log x_{m,i} - \log x_m(\boldsymbol{\alpha}))^2 \tag{6.7}$$

indicates the empirical variance. The dependency on the effective reaction rates $\boldsymbol{\alpha}$ enter via the model motif abundances $x_m(\boldsymbol{\alpha})$. The maximum likelihood estimate (MLE) of the reaction rate constants can be obtained by maximizing the log-likelihood

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \mathcal{J}(\boldsymbol{\alpha}). \tag{6.8}$$

In this Chapter, all optimizations were performed via multi-start local optimization using the lsqnonlin function from Matlab with the boundaries of the reaction rate constants set to $10^{-12} < \alpha_r < 10^3$ (for all reactions r).

6.3 Histone H4 acetylation is motif specific

We start by analyzing the simplest scenario, which contains only one basal reaction rate governing all acetylation reactions. In this unspecific model all motifs with the same total number of acetylations (0, 1, 2, 3 and 4) will have the same abundance due to symmetry of the model. Using MLE we determine the basal rate constant $\alpha_b = 0.182$ [0.170, 0.195] (95% confidence interval estimated using the profile likelihood) that fits the data best (Fig. 6.2 A right). The basal reaction rate is hence approximately 18% of the deacetylation rate leading to an overall higher abundance of motifs with a lower degree of acetylation. This simple scenario already suffices to capture the overall trend of decreasing abundance with increased degree of acetylation apparent in the data (Fig. 6.2 A left). However, comparison of the predicted and the experimentally determined values shows that the abundance for motifs with a single or two acetylated lysines (except K5K12) is overestimated. Moreover this very simple model cannot capture the large abundance variability present for the multiply acetylated motifs. In summary, we find that 86.6% (13 out of 15 measured abundances) are not explained by the unspecific scenario, because they fall outside the estimated 95% confidence intervals of the model. We therefore conclude that the unspecific scenario is insufficient to explain the observed abundances.

Next, we analyze the site-specific acetylation scenario, in which the acetylation rates of each site K5, K8, K12 and K16 are allowed to differ. We construct all 11 possible models and in each case obtain the MLE of the parameters. We compare the models using the BIC, and find that the most complex model with four different site-specific reaction rates best explains the measured abundances (Fig. 6.2 B left). This model yields the site-specific acetylation rate constants $\alpha_{K5} = 0.225\,[0.181,\ 0.280]$, $\alpha_{K8} = 0.126\,[0.104,\ 0.154]$, $\alpha_{K12} = 0.417\,[0.331,\ 0.528]$ and $\alpha_{K16} = 0.107\,[0.088,\ 0.131]$. We note that as

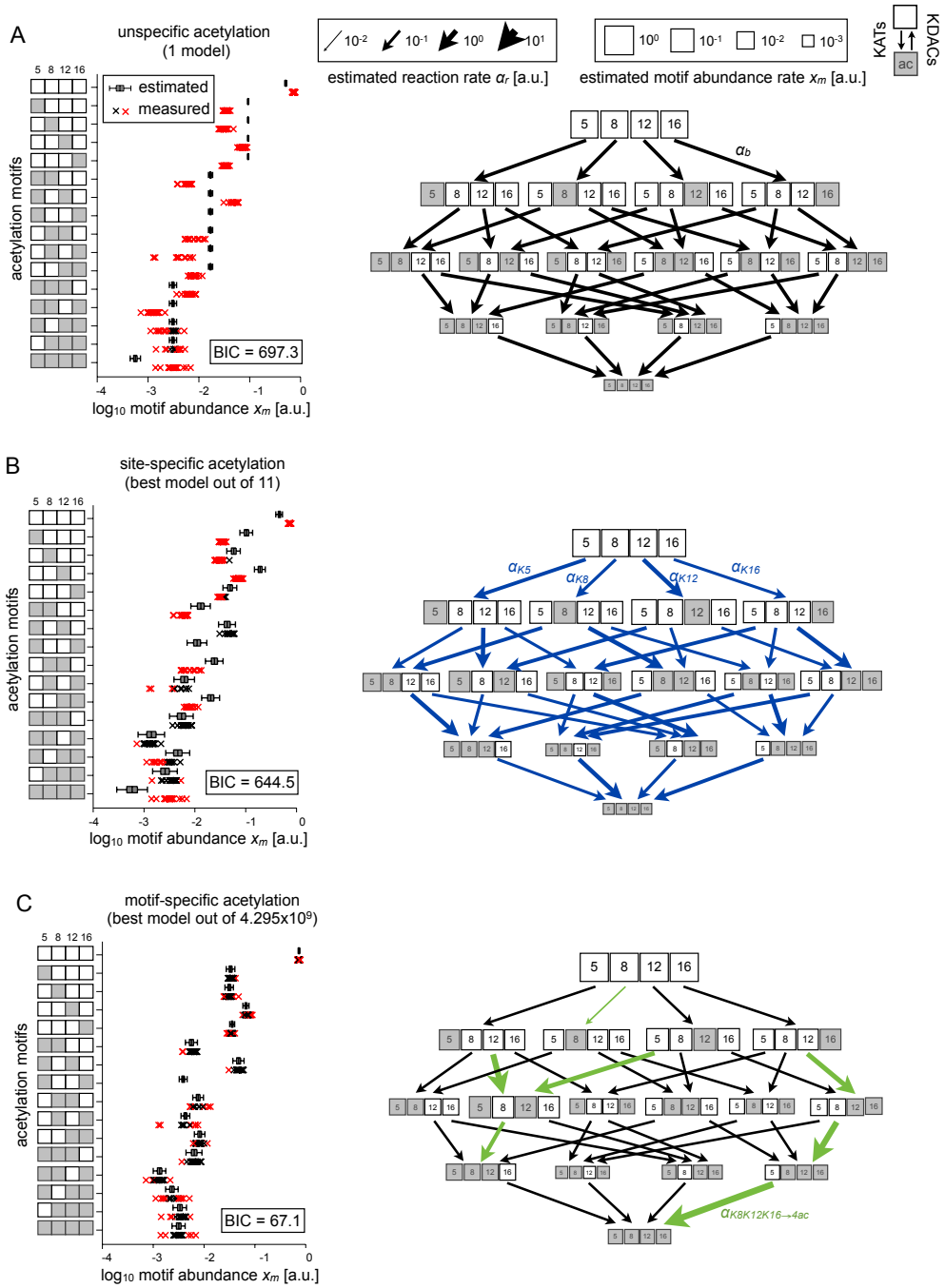
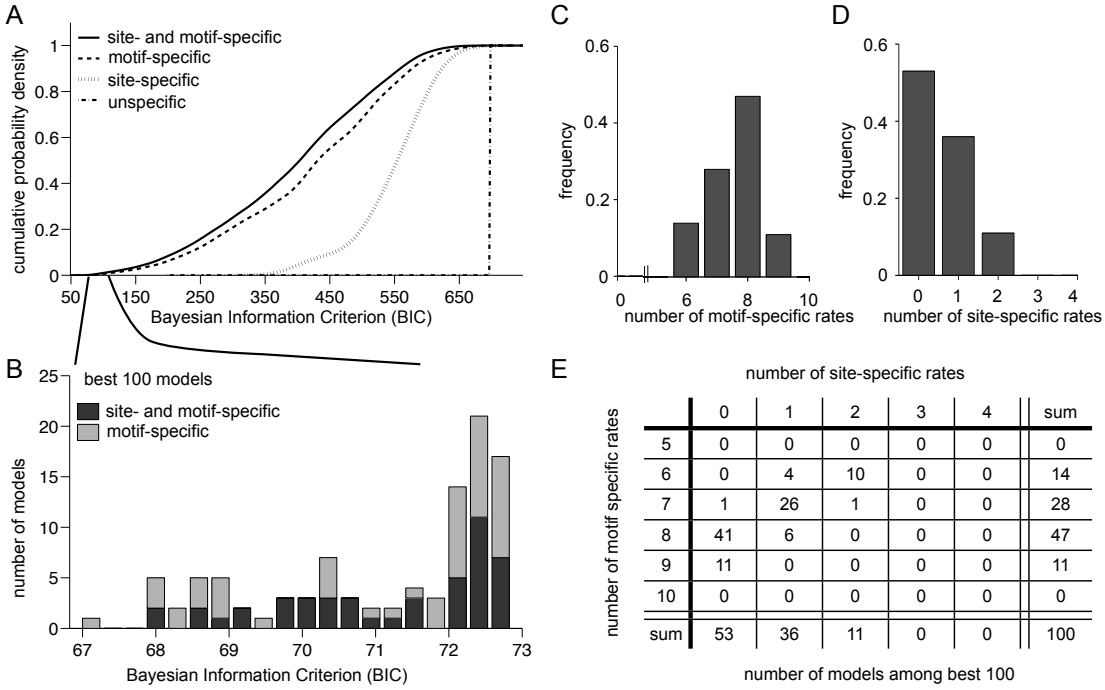


Figure 6.2: Testing different histone acetylation scenarios: a motif-specific model is preferred over unspecific and site-specific models (Figure legend on next page).

Figure 6.2: (From previous page). Testing different histone acetylation scenarios: a motif-specific model is preferred over unspecific and site-specific models. (A) Left: Fitted basal reaction rates of the unspecific model after fitting its parameter α_b to the data. The size of the arrows corresponds to the magnitude of the basal reaction rate α_b. Right: Measured (indicated with $\times$ symbols, 18 replicates) and estimated (indicated with bars and boxes) abundances of acetylation motifs. For 13 motifs the measured abundances do not fall within the 95% confidence intervals of the model, indicating that the unspecific model is inadequate to explain the data. (B) Left: Reaction rates of the best site-specific model, where 4 sites are acetylated with different rates α_{K5}, α_{K8}, α_{K12}, α_{K16}. K12 is acetylated fastest followed by K5, K8 and K16. Right: For 7 motifs all abundances deviate significantly from the measured abundances indicating that the site-specific model is also inadequate to explain the data. (C) Left: Reaction rates of the best motif-specific model, characterized by 7 motif-specific reactions. The acetylations K5 $\rightarrow$ K5K12, K12 $\rightarrow$ K5K12, K16 $\rightarrow$ K12K16, K5K12 $\rightarrow$ K5K8K12, K12K16 $\rightarrow$ K8K12K16 and K8K12K16 $\rightarrow$ 4ac are increased compared to the basal reaction rate, whereas the acetylation 0ac $\rightarrow$ K8 is decreased. Right: The motif-specific acetylation model explains the mean of all measured acetylation motif abundances. Corresponding to the BIC there is very strong evidence that the best motif-specific model describes the data significantly better than both the best site-specific model (ΔBIC = 577.4) and the best unspecific model (ΔBIC = 630.2). This Figure and its legend is adopted from Figure 2 of the author's following publication Blasi et al. (2016a).

for the unspecific scenario, the acetylation rates are all less than 1, indicating that deacetylation is faster than acetylation in this dynamic equilibrium. Since in this model each site possesses a different acetylation rate, it is possible that motifs within a single layer can achieve different abundances. Indeed, this extra flexibility allows the model to achieve better agreement with the data compared to the unspecific model (Fig. 6.2 B right), and is favored by the BIC (644.5 for the site-specific model as compared to 697.3 f or the unspecific model). We find that 46.6% of all replicates of the measured motif abundances (0ac, K5, K12, K5K8, K8K12, K12K16 and 4ac) fall outside the 95% confidence intervals of the predicted abundances, representing a roughly twofold improvement compared to the purely unspecific model. However, the substantial fraction of measurements not explained by the model suggests that a more complex model, which accounts for differences among individual motifs, is required.

Finally, we analyze the motif-specific acetylation scenario which takes into account the context of nearby modifications. We compute the BIC score of each model and find that the BIC increases monotonically for models with more than 7 motif-specific rates



	0	1	2	3	4	sum
5	0	0	0	0	0	0
6	0	4	10	0	0	14
7	1	26	1	0	0	28
8	41	6	0	0	0	47
9	11	0	0	0	0	11
10	0	0	0	0	0	0
sum	53	36	11	0	0	100

Figure 6.3: Model selection on the tested acetylation scenarios. (A) The cumulative distribution of Bayesian Information Criterion (BIC) scores shows a strong preference (i.e. left shift) of motif-specific and combined motif- and site-specific models over unspecific and purely site-specific models. Shown are all tested models ($> 10^9$). (B) Comparison of the best 100 models shows BIC scores ranging from 67.1 to 73.2 with little difference between motif-specific and combined motif-and- site-specific models; the BIC difference of $\Delta\text{BIC} < 6$ indicates no individual model is singularly the best. (C) The best 100 models all exhibit motif-specific acetylation with 6 to 9 motif-specific rates. (D) The majority of the best 100 models are not site-specific and no model is characterized by more than 2 site-specific rates. (E) All models in the ensemble contain between 6 and 9 motif-specific rates. However, most models show no site-specificity, and no model contains more than 2 site-specific rates. This Figure and its legend is adopted from Supplementary Figure 2 of the author's following publication (Blasi et al., 2016a).

(data not shown). Thus, we limit the analysis to maximally 11 motif-specific rates, reducing the number of candidate models from 4.295×10^9 to 2.366×10^8 models. Of the motif-specific models, the model that best fits the data has 7 motif-specific reaction rates, while all other reactions are governed by the same basal reaction rate (Fig. 6.2 C left). Six motif-specific rates are faster than the basal rate of 0.067 ± 0.004 and three reactions (K5 $\rightarrow$ K5K12, K12K16 $\rightarrow$ K8K12K16 and K8K12K16 $\rightarrow$ 4ac) have an acetylation rate constant $\alpha_r > 1$, indicating faster acetylation than deacetylation.

The obtained BIC of 67.1 indicates a strong preference of the motif-specific scenario over the unspecific (ΔBIC = 631.2) and the best site-specific scenario (ΔBIC = 577.4), which is corroborated by the observation that the best motif-specific model captures all measured mean motif abundances (Fig. 6.2 C right).

Our analysis of the three hypothetical scenarios provides strong evidence that motif-specificity is a necessary component to accurately recapitulate measured motif abundances. Having identified the single best model, we next analyze the robustness of this candidate model. Until now we considered only unspecific or exclusively site- or motif-specific models. However, by allowing both site- and motif-specificity simultaneously, it might be possible to obtain an even better candidate model. Thus we again fit all models containing up to 11 motif or site-specific acetylation rates (9.3×10^8 total models) and compare the models using the BIC. Interestingly, we find that permitting both site- and motif-specific rates improves the corresponding model BICs only minimally (Figure 6.3 A). The model comparison reveals that the best model in this mixed scenario is exactly the same candidate found in the motif-specific scenario, i.e. no site specific reactions included (Figure 6.2 C). Taken together, this implies that site-specificity is not essential to describe the observed abundances and confirms our previous finding (Figures 6.2A-C) that motif-specificity strongly contributes to the acetylation network.

We compare the best model with all other models by examining the distribution of BIC scores over all scenarios. This analysis reveals that i) the BIC scores of motif-specific models are vastly reduced compared to site-specific or unspecific models, confirming the essentiality of motif-specificity (Figure 6.3 A), and ii) there are approximately 100 motif-specific and motif-and-site-specific models with very similar BIC scores to the best model (ΔBIC < 6) (Figure 6.3 B). However, a ΔBIC of less than 6 is insufficient to reject one model in favor of another (see Section 4.1), therefore we obtain rather an ensemble of candidate models for further consideration. All models in the ensemble contain between 6 and 9 motif-specific reaction rates (Figures 6.3 C, E) while 89% of the best 100 models have only one or no site-specific rate (Figures 6.3 D, E).

6.4 Pathway prediction

In order to explore the commonalities among the best models, we grouped them using hierarchical clustering (see Section 4.3) on the reaction rates of the models. We applied an average linking clustering as implemented in the Matlab software (function 'clustergram') using a Minkowski distance with exponent $p = 4$. We clustered the matrix of the

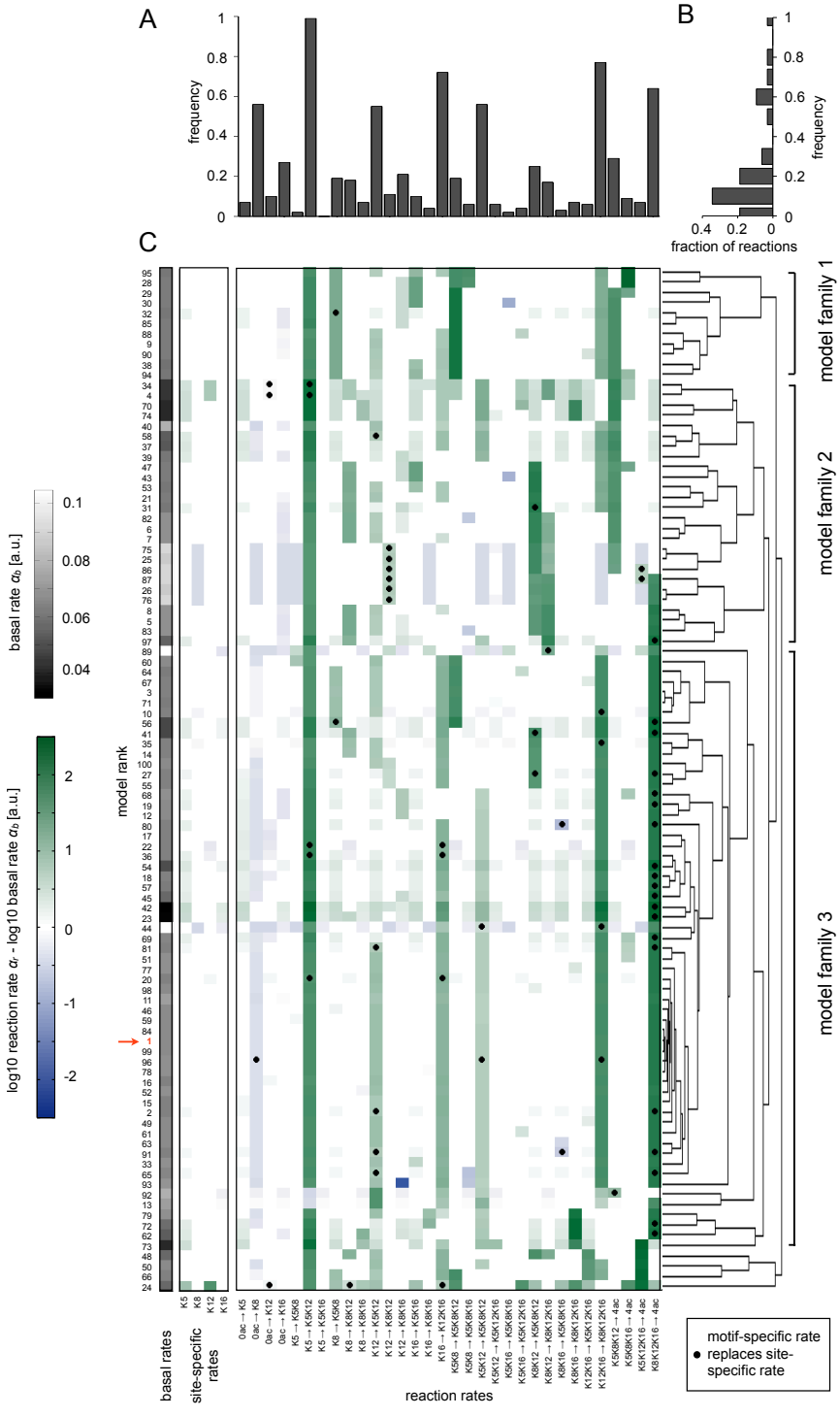


Figure 6.4: Model ensemble analysis (Figure legend on next page).

Figure 6.4: (From previous page). Model ensemble analysis. (A) The histogram shows the abundance distribution for all motif-specific reaction rates among the best 100 models. K5 → K5K12 is supported by 99/100 best models. Seven reaction rates occur with a frequency more than 50%. (B) The motif-specific reactions occur with variable frequency within the best 100 models. (C) We performed hierarchical clustering to group models according to similarity patterns in their estimated acetylation rates. The resulting clusters can be categorized into three model families and a small group of 5 uncategorized models. The black dots in the heatmap indicate cases where the motif-specific reaction rate supersedes the site-specific rate (left columns). The color key shows the magnitude of the log fold-change of reaction rates (colored, right) relative to the basal rate of each model (gray, left). Best model (ranked 1) is highlighted in red (line 75). This Figure and its legend is adopted from Figure 3 of the author's following publication (Blasi et al., 2016a).

32 log-transformed reaction rates of all best 100 models with respect to the models. We partition the hierarchical clustering at the third level from the root of the dendrogram and define clusters with similar patterns among their reaction rates if they are represented by more than 5 models. By doing so, we identified three large model families (at least 5 models per family), distinguished by similar patterns among their reaction rates (Figure 6.4). Moreover, each family contained a different collection of frequently occurring motif-specific rates (Figure 6.5).This procedure allows us to further analyze properties of the model families in terms of their common features.

By examining the structure of motif-specific rates within each family, we are able to predict acetylation pathways that yield specific combinatorial motifs. In particular, we examine the frequency distribution of specific motif-specific reactions within the entire ensemble (Figures 6.4 A, B), and discover a subset of overrepresented motif-specific reactions, which occur in more than 50% of the candidate models. We conservatively threshold the motif-specific reactions at 60% for each model family to yield a set of essential acetylation segments, and connect adjacent segments from this set to construct hypothetical acetylation pathways (Figures 6.6 A-D).

Applying this analysis, we identify four pathways, which are distinct between the three model families (Figure 6.6 A-D). Pathway 1 (supported by 45.8% of the models in family 3) suggests that K5K12 is not only generated via K5 → K5K12 (100% of the models in family) but also via K12 → K5K12 (45.8% of the models in family 3). Pathway 1 further suggests that K5K12 is subsequently acetylated at K8 to generate K5K8K12 (78.0% of the models in family 3), which is consistent with previous biochemical experiments (Makowski et al., 2001). Furthermore, most models (84.8%) in family 3 contain pathway

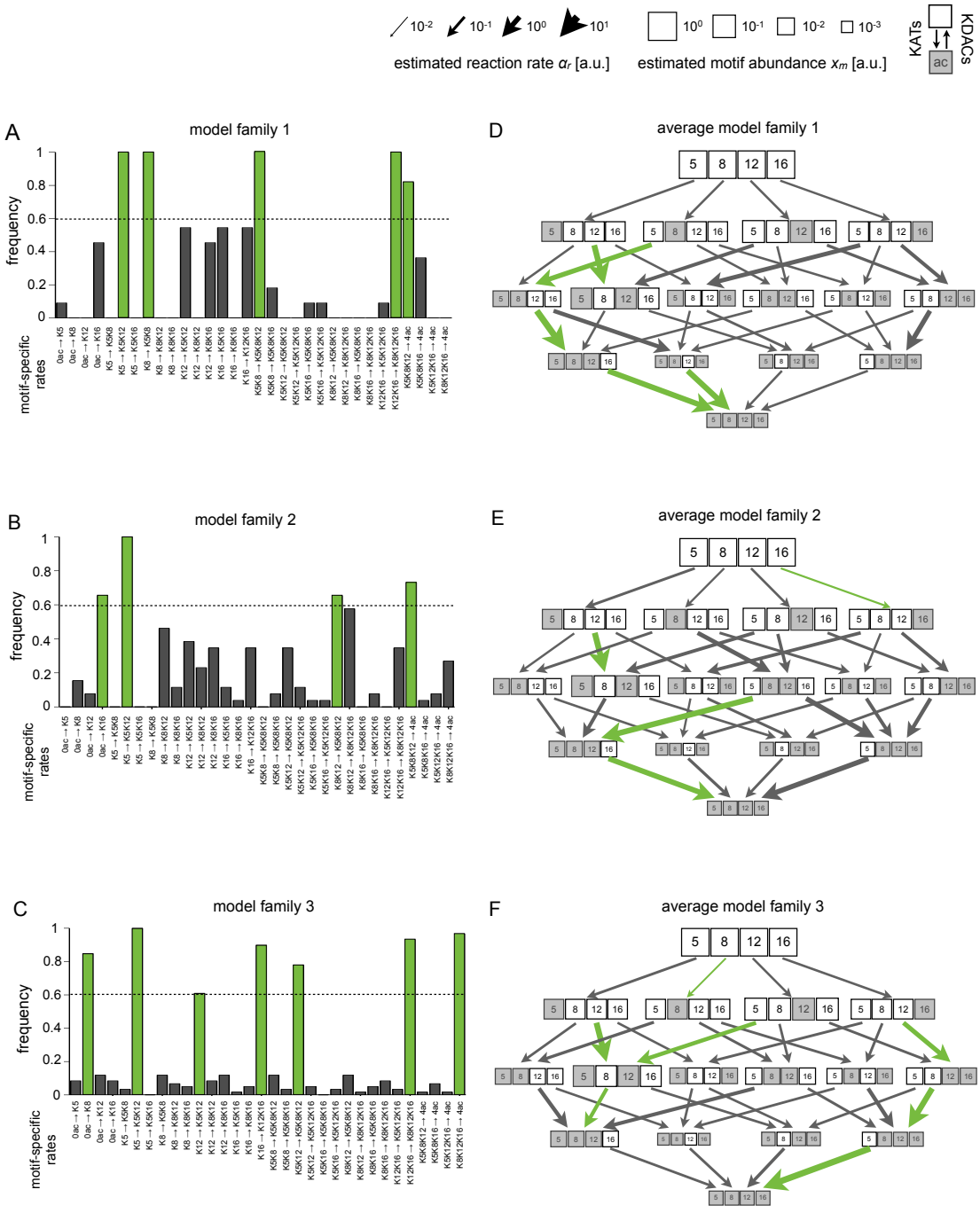


Figure 6.5: Model families allow for the prediction of acetylation pathways (Figure legend on next page).

Figure 6.5: (From previous page). Prediction of acetylation pathways. (A-C) Distribution of motif-specific rates in the three model families that were identified in Figure 3C Reaction rates that occur more frequently than 60% within one family are highlighted in green. (D-F) Network representation of average models for model families 1-3. Box size indicates model-predicted abundances (compare similarity to Figure 6.1 B). Motif-specific reaction rates with 60% support (see A-C) are highlighted in green whereas reaction rates with less support are displayed in black. Each arrow represents the average reaction rate of the respective model family. We define segments connecting more than three motifs with motif-specific reaction rates to be acetylation pathways (see Figure 6.6 A-D). This Figure and its legend is adopted from Supplementary Figure 3 of the author's following publication (Blasi et al., 2016a).

4, which yields the fully acetylated isoform 4ac via the serial acetylation K16 → K12K16 → K8K12K16 → 4ac. We also discover alternative pathways to the fully acetylated 4ac state in model families 2 and 3. Pathway 2 (contained in 81.8% of the models in family 1) suggests K8 → K5K8 → K5K8K12 → 4ac while pathway 3 (contained in 42.3% of the models in family 1) suggests K8K12 → K5K8K12 → 4ac .

6.5 Qualitative validation of computationally predicted pathways

After having trained our model ensemble on the dataset derived from unperturbed Drosophila cells, we now want to qualitatively validate the predicted pathways. We thus analyze an independent dataset where individual KATs have been depleted by RNA interference (Feller et al., 2015). Specifically, we annotate enzymes to the predicted pathways if their ablation leads to a reduction of the product and an increase of the unused substrate. For a details we refer to Blasi et al. (2016a).

In summary, our analysis identifies three acetylation pathways that can be linked to known and recently reported enzyme activity (Figure 6.6 E). The combinatorial acetylation motif K5K12 is generated by HAT1 via K5 (red pathway in Figure 6.6 E). The fully acetylated 4ac state appears to originate from two main routes: The 'inverse K5K8 zipper pathway' (blue pathway in Figure 6.6 E) relies on the enzymes CBP, NAA10, NAT10, MGEA5, and the 'K16 zipper pathway' (orange pathway in Figure 6.6 E) is brought about by the coordinated activity of MOF, TIP60 and CBP. The 'K16 zipper pathway' was previously proposed based on the skewed abundance distribution

in many human cell types as well as in other selected species (Garcia et al., 2007; Zhang et al., 2002). In those species the abundance of K16 exceeds by far that of other mono-acetylated motifs, a trend that continues also for di- and tri-acetylated motifs that contain acetylated K16. While established for other species, such a pathway has not been previously described for Drosophila, nor is it evident from the measured abundances alone (Figure 6.1 B).

6.6 Discussion

The objective of this study was to advance our understanding of how complex acetylation patterns arise. The simplest explanation is that cooccurring acetylations on the histone H4 N-terminus arise due to the uncoordinated action of individual KATs independent of prior modifications close to the substrate lysine. Our mathematical modeling approach supports an alternative scenario, whereby dedicated enzymatic pathways generate combinatorial motifs via motif-sensitive enzymes. We validated the putative acetylation pathways inferred by our models using an independent dataset from a cellular perturbation study.

In order to dissect the distribution of combinatorial motifs, and their respective independent or 'coordinated' enzymatic origins, it is necessary to systematically characterize the abundance of all possible acetylation permutations. While the dataset of Feller et al. (2015) is uniquely suited to this task as it provides comprehensive quantitative measurements for all possible 16 histone H4 acetylation motifs, it also poses several difficulties for mathematical modeling which we addressed by developing a tailored computational framework.

Our framework can be easily adapted in order to test for the existence of (further) general design principles that are responsible for modification regulated biological networks in different cellular contexts and disease scenarios. For example, while our assumption of a universal, constant deacetylation rate is justified by our current knowledge on KDAC activities in Drosophila cells (Feller et al., 2015), it is conceivable that the higher complexity of the human KDAC network (Joshi et al., 2013; Yang and Seto, 2007) may suggest an extension to the modeling approach which relaxes this assumption.

The modeling framework presented here has several benefits. It is exhaustive, in that we compare all model topologies. It utilizes an established information-theoretic technique to rigorously evaluate candidate models. It is directly interpretable, revealing

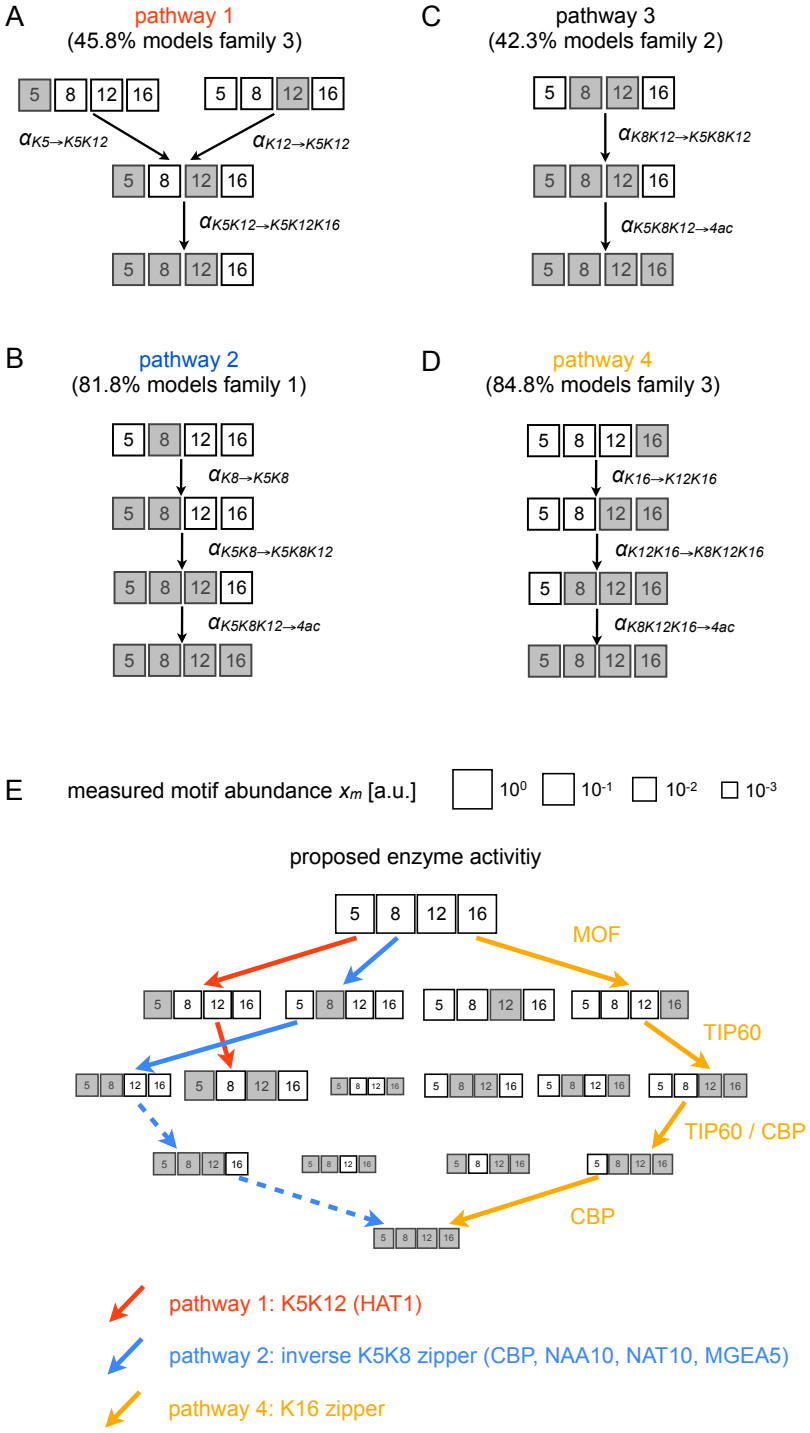


Figure 6.6: Predicted acetylation pathways (Figure legend on next page).

Figure 6.6: (From previous page). Predicted acetylation pathways. (A-D) Predicted acetylation pathways that are composed of connected motif-specific acetylation rates with more than 60% support within a family (see Figure 6.5). Each model family is characterized by distinct acetylation reaction pathways. (E) We validate the predicted pathways by an independent KAT depletion dataset (Feller et al., 2015)) and propose candidate enzymes for the acetylation pathways. We find evidence for pathway 1, 2 and 4 (A, B, D) whereas we exclude pathway 3 (C). In pathway 1 (red), HAT1 catalyzes the two main steps of the K5K12 pathway, and a potential third 'promiscuous' step to yield K5K8K12 (not shown). CBP, NAA10, NAT10 or MGEA5 are putative candidates for the first two reactions of the inverse K5K8 zipper (pathway 2, blue, solid). The remaining links of pathway 2 (blue, dashed) are inferred by the model, but the dataset does not allow the assignment of the associated enzymes. In pathway 4 (orange), the K16 zipper is likely generated by the subsequent actions of MOF (0 $\rightarrow$ K16), TIP60 (K16 $\rightarrow$ K12K16 and K12K16 $\rightarrow$ K8K12K16) and CBP (K12K16 $\rightarrow$ K8K12K16 and K8K12K16 $\rightarrow$ 4ac). This Figure and its legend is adopted from Figure 4 of the author's following publication (Blasi et al., 2016a).

acetylation rates, and site- and motif-specificity. Using the model, one can identify pathways that are directly testable, aiding in hypothesis generation and experimental design. Moreover, the model is flexible enough to be adapted to more extensive datasets including e.g. time-series data following perturbation or other PTMs (e.g. methylation and other peptide substrates. Thus the modeling framework is a valuable tool to the life science community for analyzing and interpreting future datasets.

Chapter 7

Time-series analysis of single-cell protein levels with transfer entropy

An important challenge of computational biology is to discover regulatory links (cf. the yearly held DREAM challenges where this is often the aim; see e.g. Marbach et al. (2012)). Recent examples where the interactions of individual genes during the formation of adult blood cells have successfully been learned from single-cell snapshot data are Ocone et al. (2015) and Moignard et al. (2015). In this Chapter we use transfer entropy (see Chapter 5.2) for time-series analysis in order to infer information transfer between two protein levels that were measured over time in single cells. This information transfer could occur directly via a gene regulatory link or via indirect effects (e.g. one protein could be involved in a gene regulatory network that at a later time point triggers the regulation of the other protein). As we outline in Section 2.1 the regulation of gene expression is one of the major sources responsible for establishing cellular heterogeneity among cells with the same genetic information. We want to stress that this Chapter does not contain a methodological novelty but presents the application of an existing mathematical method to new data to find novel biological insights.

We start by introducing time-lapse microscopy data and explaining why transfer entropy is a suitable method for its analysis (Section 7.1). In Section 7.2 we elaborate on the calculation of the probability distributions necessary for the evaluation of the transfer entropy. As a proof of principle we apply transfer entropy to artificial data

simulated from the stochastic simulation algorithm (SSA; Section 7.3) before we use it to infer the information transfer between the time-series of two proteins that play a key role in hematopoiesis (Section 7.4). We then conclude with a discussion of our findings in Section 7.5.

The novel contributions of this Chapter are that a probabilistic time-series analysis method is used for gene expression data analysis. Before transfer entropy has been used in physics (Schreiber, 2000) and in bio-medicine (Lee et al., 2012) to infer directional links between two time-series. Using this method we find an information transfer from PU.1 to Gata-1 in cells with granulocyte-monocyte progenitor (GMP) fated cells but not in megakaryocyte-erythrocyte progenitor (MEP) fated cells (see 2.6 A).

Some results of this Chapter are already contained in a master thesis (Gumpinger, 2015), which the author of this thesis co-supervised, conducted pioneering studies for and also contributed in conceiving the objectives for (see Section 7.5 for an explicit discussion of the individual contributions). In addition to the results presented there, we here apply transfer entropy to artificial data simulated from the SSA. Moreover, we extend the previous results on hematopoietic stem and progenitor cells by providing p-values for the calculated transfer entropy values and give further details on the biological relevance of our findings.

This Chapter is based on and in part identical with the following manuscript that is currently in preparation:

Gumpinger, A.*, **Blasi, T.***, Hennig, H., Theis, F.J. and Marr, C. Transfer entropy: PU.1 transfers information to Gata-1 in GMPs but not in MEPs. *In preparation.* (* equal contributions)

7.1 Biological background and problem statement

A crucial question in stem cell research is how stem and progenitor cells differentiate into mature cells. To dissect different cellular phenotypes and to understand the gene regulatory network during differentiation it is very important to perform single cell experiments (see Figure 2.3). Additionally, to temporally resolve changes during differentiation it is important to continuously monitor the gene expression levels from the same cell (and its descendants) without perturbing the cells substantially (Hoppe et al., 2014).

One way how time-dependent data from the same cell can be measured is single-cell time-lapse microscopy (Schroeder, 2011) where single cells are imaged in constant intervals of time. Time-lapse microscopy can provide both, bright field and fluorescent images. To this end either fluorescently tagged protein antibodies (Kueh et al., 2013) or genetically modified cells that have fluorophores fused to a protein of interest (Filipczyk et al., 2015) can be used.

A well-studied system for cellular decision making is adult hematopoiesis (see Chapter 2.4). There hematopoietic stem cells (HSCs) undergo a hierarchy of subsequently differentiate into more and more restrictive progenitor cells, eventually giving rise to all mature blood cells (Orkin and Zon, 2008). A particular transition within hematopoiesis is the differentiation of common myeloid progenitors (CMPs) into either myeloid-erythrocyte progenitors (MEPs; giving rise to all red blood cells) or granulocyte-monocyte progenitors (GMPs; giving rise to a substantial part of all white blood cells). Biochemical studies showed evidence that this transition is controlled by the proteins PU.1 and Gata-1 Galloway et al. (2005); Rhodes et al. (2005). The details of the interplay between PU.1 and Gata-1 have been investigated by numerous biochemical studies that reported auto-regulatory activity for both PU.1 (Okuno et al., 2005) and GATA-1 (Nishimura et al., 2005) and a mutually antagonistic regulation between them (Arinobu et al., 2007; Liew et al., 2006).

In the mean time several mathematical models were proposed that exhibit dynamical properties allowing to describe the transition from one progenitor into two distinct progenitors that are more restricted (Duff et al., 2012). The biochemical evidence for the mutually antagonistic and auto-regulatory activity motivated the development of so called toggle-switch models that incorporate these interactions. Toggle-switch models both based on deterministic (Chickarmane et al., 2009; Huang et al., 2007; Roeder and Glauche, 2006) and stochastic dynamics (Strasser et al., 2011) were studied and provided further insights into gene regulatory networks suitable to describe cellular decision making by investigating possible model topologies and their parameter spaces.

Recent results by Hoppe et al. (2016), however, indicate that our current perspective on the interaction of PU.1 and Gata-1 during murine hematopoiesis might need to be revised, as they did not find evidence for mutually antagonistic regulation in their time-lapse microscopy data (see Section 7.4). Instead of proposing new models that may give rise to their recent data, we here take an alternative approach and investigate the information transfer between the two protein time-series. Transfer entropy is an ideal candidate to analyze the interdependency between two protein time series since

it does not rely on a particular definition of the underlying gene regulatory network, but it quantifies the net information transferred from one protein to the other in an entirely data-driven way.

7.2 Transfer entropy: a method to measure directional relations

We use transfer entropy to analyze the information transfer between two time-series $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ of protein abundances measured with time-lapse microscopy. It is important to note that, here, we use transfer entropy to generally analyze the information transfer between the two measured protein time-series directly from the data without any assumptions on the particular underlying gene regulatory network.

To evaluate the net transfer of information from one time-series to the other we consider the difference between the transfer entropy (see Section 4.2) from the time-series X to Y and from Y to X:

$$\Delta \mathrm{TE}_{X \to Y}(\tau) = \mathrm{TE}_{X \to Y}(\tau) - \mathrm{TE}_{Y \to X}(\tau). \tag{7.1}$$

The remaining task for the calculation of the transfer entropy is to estimate the joint probability distribution $p(y_i, y_{i-\tau}, x_{i-\tau})$ in Eq. 4.14 from the data. The other two conditional probability distributions necessary for the evaluation of the transfer entropy, Eq. 4.14, can subsequently be obtained from the joint probability distribution via $p(y_i|y_{i-\tau}, x_{i-\tau}) = p(y_i, y_{i-\tau}, x_{i-\tau})/p(y_{i-\tau}, x_{i-\tau})$ and $p(y_i|y_{i-\tau}) = p(y_i, y_{i-\tau})/p(y_{i-\tau})$. Here, we construct the joint probability distribution with kernel density estimation (KDE). Although KDE scales badly with the dimension of the probability distribution that is to be estimated, we here only focus on a three dimensional probability distribution, where KDE delivers suitable results as we verified by visually inspecting the estimated probability distributions (data not shown).

Besides KDE there are different methods to calculate the joint probability distribution from the measured time series. Lee et al. (2012) presented different methods to estimate such probabilities, one of them is adaptive partitioning using the Darbellay-Vajda algorithm (Darbellay and Vajda, 1999). We chose to use KDE for the construction of the joint probability distribution as it practically proved to be most efficient for the task at hand.

Note that for data that is obtained from a biological process, which can be described with the chemical master equation (CME; i.e. where X and Y are the abundances of some molecular species and correspond to individual realizations of the CME) the joint probability distribution $p(y_i, y_{i-\tau}, x_{i-\tau})$ needed in Eq. 4.14 could in principal also be calculated from an analytical solution of the CME.

Kernel density estimation

In this subsection we explain how we obtain the joint probability densities $p(y_i, y_{i-\tau}, x_{i-\tau})$ needed to evaluate the transfer entropy. The joint probability at a given point $(\tilde{y}_i, \tilde{y}_{i-\tau}, \tilde{x}_{i-\tau})$ can be estimated by

$$\begin{aligned} p_{\mathrm{KDE}}(\tilde{y}_i, \tilde{y}_{i-\tau}, \tilde{x}_{i-\tau}) \approx \\ \frac{1}{N}\sum_{j=1}^{N} K\left(\frac{\tilde{y}_i - y_{i+j}}{h_{y_i}}\right) \times K\left(\frac{\tilde{y}_i - y_{i-\tau+j}}{h_{y_{i-\tau}}}\right) \\ \times K\left(\frac{\tilde{x}_{i-\tau} - x_{i-\tau+j}}{h_{x_{i-\tau}}}\right) \end{aligned} \tag{7.2}$$

where for each of the N data points we apply a Gaussian kernel

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-0.5u^2). \tag{7.3}$$

To fix the scaling factor $h_{(.)}$ that defines the bandwidth of the estimated kernels we follow Lee et al. (2012) and use the rule of thumb motivated by Silverman (1986)

$$h_{(.)} = 1.06 b \hat{\sigma} N^{1/5} \tag{7.4}$$

where $\hat{\sigma}$ is the empirical standard deviation of the time-series.

7.3 Application of transfer entropy to data simulated with the stochastic simulation algorithm

We use the artificial data set simulated in Section 3.4 with the stochastic simulation algorithm (SSA) in order to evaluate the capability of transfer entropy to detect the correct information transfer. We take 100 steady-state realizations of a simple gene regulatory network where a protein A that is constantly expressed represses the expression of a gene B where we use the same parameters as denoted in Section 3.4 (see Figure 7.1 A and B).

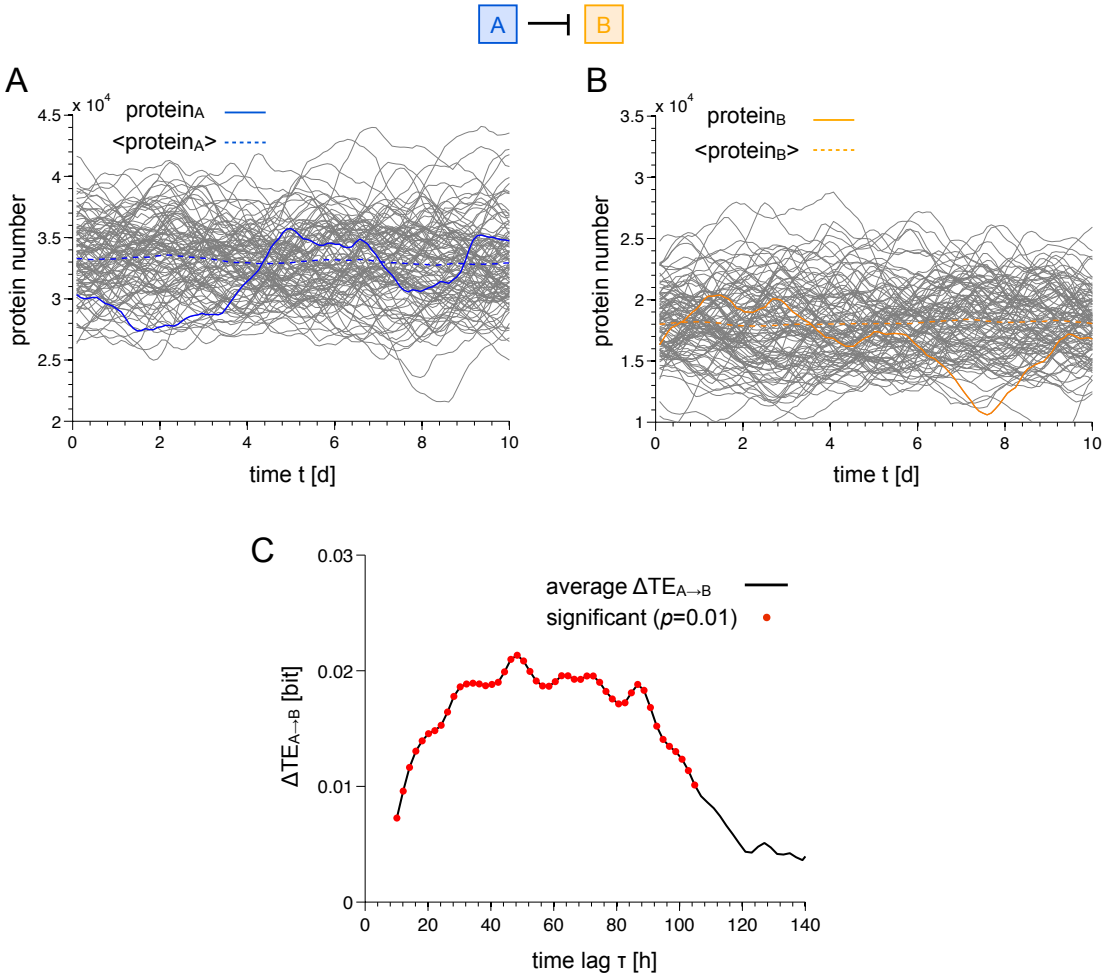


Figure 7.1: Transfer entropy for data simulated with the stochastic simulation algorithm. (A) and (B) Excerpt of 100 steady-state time-series realizations of the simple gene regulatory network displayed in Figure 3.3 where the proteins of a gene A regulate the expression of a gene B. (C) Average difference $\Delta\mathrm{TE}_{A\to B}$ of 100 time-series where gene A regulates gene B. Time lags where $\Delta\mathrm{TE}_{A\to B}(\tau)$ is significantly different from zero with a p-value of $p = 0.01$ are marked with red dots.

We calculate the transfer entropy $\mathrm{TE}_{A\to B}(\tau)$ and $\mathrm{TE}_{B\to A}(\tau)$ (Eq. 4.14) for all the 100 realizations and all time lags τ. The average difference between transfer entropies $\Delta\mathrm{TE}_{A\to B}(\tau)$ is displayed for all time lags $\tau < 140$h in Figure 7.1 C. To obtain p-values for the calculated transfer entropy values to be significant, we performed a one sided t-test with a p-value $p = 0.01$ at every time lag.

We find that for time lags 10h $\lesssim \tau \lesssim$ 105h there is a significant information transfer

from the time-series of protein A to protein B. This finding is consistent with the underlying gene regulatory network where protein A is upstream of protein B.

7.4 Application of transfer entropy to time-lapse microscopy data from differentiating hematopoietic stem cells

In this Section we analyze the time-series data of Hoppe et al. (2016). There, PU.1 and Gata-1, two proteins that have been considered to play a key role for the differentiation into the granulocyte-monocyte lineage or the megakaryocyte-erythrocyte lineage (see Section 7.1), were fused with fluorophores. Then murine hematopoietic stem cells (HSCs) were extracted and the intensity levels of the fluorophores were continuously monitored using time-lapse microscopy. Once the cells were differentiated into either GMPs or MEPs they were annotated with an additional surface markers and/or manual annotation based on morphological properties.

The obtained sequence of images were further processed along Buggenthin et al. (2013) where the cells were tracked, segmented and both their fluorescence intensities and morphological features (such as the area of the cell) were extracted. Since stem and progenitor cells proliferate, a single progenitor cell gives rise to a multitude of descendant cells. This defines a relationship between the cells: all cells that stem from the same progenitor cell are grouped into one 'trees'. Moreover, the set of a single cell's direct progenitors are called the cell's 'branch'.

Our analysis starts with normalizing the fluorescence intensity levels by dividing through the cell's size in order to correct for cell cycle effects (see Chapter 5 for a detailed discussion of cell cycle normalization). We select trees based on a filtering strategy as outlined in Gumpinger (2015) where trees that are not annotated to either the GMP or MEP lineage at the end of the experiment are rejected as well as trees that contain cells with strong outliers in the fluorescent intensities. Using this approach we obtain 10 trees where the majority of cells differentiates to the GMP lineage (see Figure 7.2 for a typical example) and 14 trees where the majority differentiates to the MEP lineage (see Figure 7.3 for a typical example).

In order to exploit the time-series data that is contained in one tree, transfer entropy was extended to be capable to incorporate tree-structured data by Gumpinger (2015). In brief, the calculation of the joint probability distribution using KDE, Eq. 7.2, was modified such that it can be estimated for an entire tree (as compared to only an

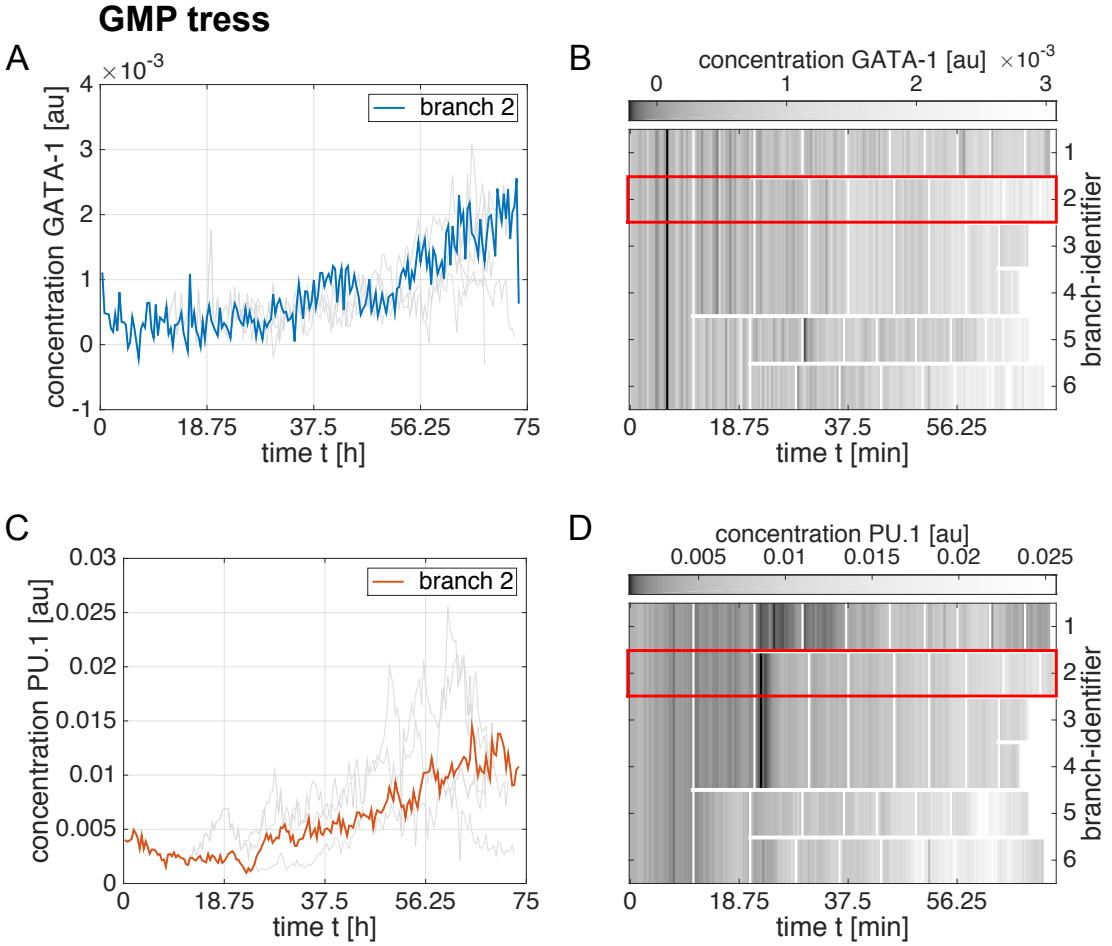


Figure 7.2: Fluorescence intensity of a typical tree with cells differentiating into the granulocyte-monocyte lineage. (A) Gata-1 and (C) PU.1 fluorescence intensity time-series of individual branches that commit to the GMP lineage (branch 2 of the Gata-1 time-series is highlighted in blue and branch 10 of the PU.1 time-series is highlighted in red). (B) and (D) Heatmap representation of the whole trees containing the branches depicted in (A) and (C). Cell divisions are indicated with white vertical lines. Figure adopted from Gumpinger (2015).

individual branch).

We apply transfer entropy for tree structured data to both the GMP and MEP fated trees and find an information transfer from PU.1 to Gata-1 in GMP fated cells but not in MEP fated cells (see Figure 7.4). We obtain values that are significantly different from zero (based on a p-value of $p = 0.01$ calculated via a one-sided t-test) of the transfer entropy for very small time lags, but more importantly many subsequent values in for

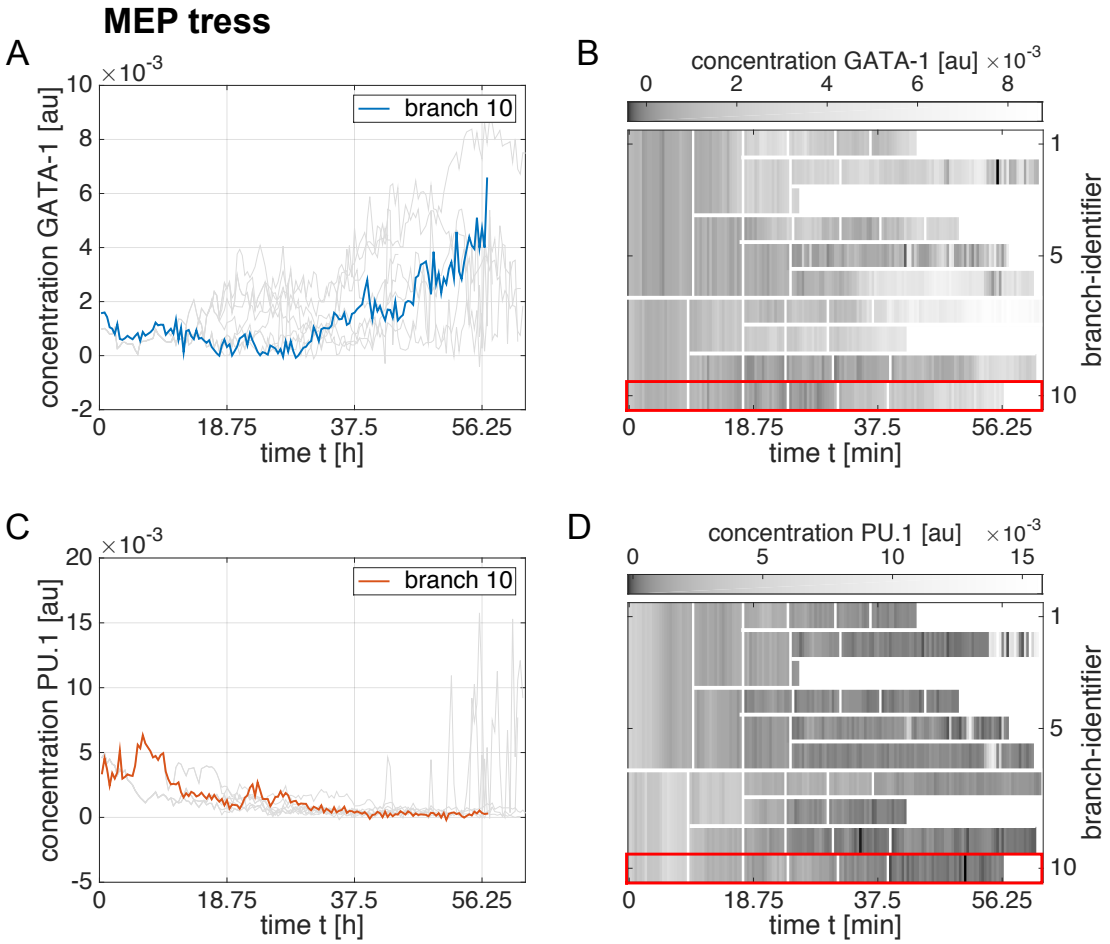


Figure 7.3: Fluorescence intensity of a typical tree with cells differentiating into the megakaryocyte-erythrocyte lineage.(A) Gata-1 and (C) PU.1 fluorescence intensity time-series of individual branches that commit to the MEP lineage (branches 10 of the Gata-1 and PU.1 time series are highlighted in blue and red, respectively). (B) and (D) Heatmap representation of the whole trees containing the branches depicted in (A) and (C). Cell divisions are indicated with white vertical lines. Figure adopted from Gumpinger (2015).

time lags 24h $\lesssim \tau \lesssim$ 44h.

It is important to note that this is a finding that is averaged over all stem and progenitor cells that are passed through during differentiation from HSCs to MEPs/ GMPs. However, under the assumption that the interdependency (e.g. a possible the gene regulatory link) between the two genes PU.1 and Gata-1 does not change during differentiation this does not affect the result.

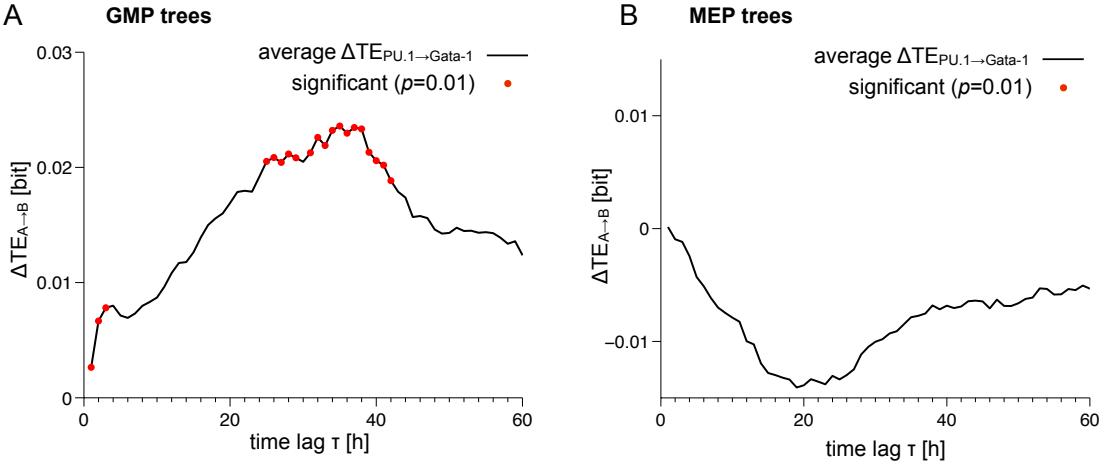


Figure 7.4: Transfer entropy for protein time-series from differentiating hematopoietic stem cells measured with time-lapse microscopy. (A) Average transfer entropy between PU.1 and Gata-1 for GMP fated cells. The difference $\Delta\text{TE}_{\text{PU.1}\to\text{Gata}-1}$ was averaged over all 10 analyzed trees. Values that are significantly different from zero (based on a p-value $p = 0.01$ obtained by a one-sided t-test) are highlighted with red dots. (B) Average transfer entropy between PU.1 and Gata-1 for MEP fated cells (averaged over the 14 analyzed MEP trees). While we find significant values of the transfer entropy indicating information transfer from PU.1 to Gata-1 for GMP fated cells we do not find significant values for MEP fated cells. This Figure is a modified version of Figures 3.18 (C) and (D) of Gumpinger (2015).

7.5 Discussion

In this Chapter we applied transfer entropy to analyze both protein time-series from artificial data and data from differentiating hematopoietic stem and progenitor cells. We could show that transfer entropy successfully detects the underlying information transfer from one protein time-series to another for a simple gene regulatory link for data simulated with the SSA. Moreover, we could show that PU.1 transfers information to Gata-1 in cells that commit to the GMP lineage, whereas we do not detect an information transfer between the two protein time-series in cells that commit to the MEP lineage.

In a master thesis (Gumpinger, 2015) that was co-supervised by the author of this thesis transfer entropy was more rigorously evaluated for artificial data simulated from a simple gene regulatory network using the Ornstein-Uhlenbeck process to simulate intrinsic gene expression noise (see Dunlop et al. (2008) for the used toy system). Moreover,

transfer entropy was compared against cross-correlation (a time-series analysis method used by Dunlop et al. (2008)). Next, transfer entropy was generalized to be applicable to tree-structured data sets. Furthermore, there it was shown that transfer entropy is capable to infer the right gene regulatory link when it is applied to synthetic gene regulatory links in *Escheria coli* investigated by Dunlop et al. (2008). Lastly, transfer entropy was applied to the data set from hematopoietic stem and progenitor cells that we also present here. We want to conclude by noting that it may be possible to use transfer entropy also on pseudo-time series that are constructed from single cell snapshot data (Bendall et al., 2014; Ocone et al., 2015; Trapnell et al., 2014).

Chapter 8

Label-free prediction of cell phenotypes based on imaging flow cytometry data

While we focussed on the analysis of transcription initiation in the previous three Chapters, we now turn to the dissection of cellular heterogeneity among single cells that were extracted from a mixed population. To this end we use supervised machine learning methods (see Section 4.4) to find differences among cellular phenotypes. More specifically, we seek to find a cell's position within its cell cycle (see Section 2.5). In contrast to Chapter 5 where we point out that cell cycle may have confounding effects when analyzing models of gene expression that are not associated with the cell cycle itself, we here specifically look for differences that are due to the cell cycle.

In this Chapter, we work with data from imaging flow cytometry (IFC) that offers information rich, morphological properties of single cells (Section 8.1). In Section 8.2 we present a novel approach to predict cell cycle phases based on bright field and dark field images alone, without the need for any further chemical stains. Subsequently (Section 8.3), we apply the proposed work-flow to different cell types and also to cells under different conditions. Eventually we conclude with a discussion where we note that our workflow is likely to be broadly applicable for the identification of many more cellular phenotypes (Section 8.4).

The novel contributions of this Chapter are the design of a user-friendly workflow to do label-free cell analysis, which involves the extraction of information rich features from

bright field and dark field images and the application of supervised machine learning on these features. The designed workflow is open-source and freely available online (CellProfiler, 2016) and accompanied by step-by-step tutorials and example data sets. By applying our worklfow to data from imaging flow cytometry we find that cell cycle phases can be determined without the need for any additional markers.

This Chapter is based on and in parts identical with the following article:

Blasi, T., Hennig, H., Summers, H.D., Theis, F.J., Cerveira, J., Patterson, J.O., Davies, D., Filby, A., Carpenter, A.E. and Rees, P. (2016). Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nature Communications* 7:10256.

The possible future applications of this Chapter are in parts subjected to the following patent application:

Hennig, H.*, **Blasi, T.***, Rees, P.* and Carpenter, A.E.* (2014). Method for Label-Free Image Cytometry. US 61/985,236. (* equal contributions)

8.1 Biological background and problem statement

A major challenge in many modern biological laboratories is obtaining information rich measurements of cells in high-throughput and at single cell resolution. Conventional flow cytometry is a widespread and powerful technique for the measurement of cell phenotype and function using targeted fluorescent stains (Brown and Wittwer, 2000). It is highly suited to the study of cell populations and rare subset identification due to its high-throughput, multi-parameter nature. The fluorescent stains can be used to label cellular components or processes, revealing specific cell phenotypes in the population and quantifying the particular state of each cell. For example, quantifying the proportion of cells in each phase of the cell cycle, including mitotic phases is very useful in the modern biological laboratory. It can be achieved with conventional flow cytometry using multiple stains: Typically, a stoichiometric fluorescent stain for DNA reports the cells' position within the G1, S and G2 phases of the cell cycle (Darzynkiewicz and Huang, 2004), and additional stains are needed to sort mitotic cells into phases. Often these stains are incompatible with live cell analysis (e.g. antibodies against histone modifications; Hans and Dimitrov (2001)) and even if live cell reporters are available (Sakaue-Sawano et al., 2008) these may have confounding effects on the cells. For example the commonly used Hoechst 33342 stain, which binds to the minor groove of the double-stranded DNA can induce single-strand DNA breaks (Chen et al., 1993), or

DRAQ5 (deep red fluorescing bisalkylaminoanthraquinone) the nuclear stain which intercalates with the cell's DNA can influence chromatin organization and lead to histone dissociation (Wojcik and Dobrucki, 2008). Also several different markers are usually required to unambiguously identify all cell cycle phases (Miltenburger et al., 1987). Therefore an assay that reduces or even eliminates the number of stains required to identify phenotypes such as the position in the cell cycle is particularly attractive.

In recent years, the two technologies of fluorescence microscopy and flow cytometry have been integrated to create imaging flow cytometry (IFC, see also Section 2.4 and Basiji et al. (2007)), where an image is captured of each cell as it flows past an excitation source and a CCD detector. It combines conventional flow cytometry's high-throughput speed and easy identification of each individual cell with the fluorescence microscopy's spatial image acquisition. Therefore imaging flow cytometry measures not only fluorescence intensities but also the spatial image of the fluorescence together with brightfield and darkfield images of each cell in a population. The rich information captured using imaging flow cytometry makes it an ideal candidate for the use of high content approaches to identify complex cell phenotypes such as the cell cycle phase of an individual cell. Filby et al. (2011) have previously demonstrated that measuring the shape of the nucleus from cells stained with a nuclear marker using imaging flow cytometry drastically improves the classification of mitotic phases. However, the even richer morphological information that can be extracted using imaging software tools (Eliceiri et al., 2012) offers the prospect of using more advanced multivariate analysis techniques to mine the data and to identify various cell phenotypes, as has been successfully done for traditional microscopy images (Jones et al., 2009; Kamentsky et al., 2011; Perlman et al., 2004; Rajaram et al., 2012). This type of analysis is also usually more accurate and less subjective than any manual analysis of the acquired images (Jones et al., 2009) as well as more robust than typical gating strategies that rely on only few features of the cells.

Here, we report that quantitative image analysis of two largely overlooked channels – brightfield and darkfield, both readily collected by imaging flow cytometers – enables cell cycle-related assays without needing any fluorescence biomarkers. We use the image analysis software CellProfiler (Kamentsky et al., 2011) to extract numerical measurements of cell morphology from the brightfield and darkfield images, then we apply supervised machine learning algorithms to identify cellular phenotypes of interest, in the present case, cell cycle phases. Avoiding fluorescent stains provides several benefits: it reduces effort and cost, avoids potentially confounding side effects of live

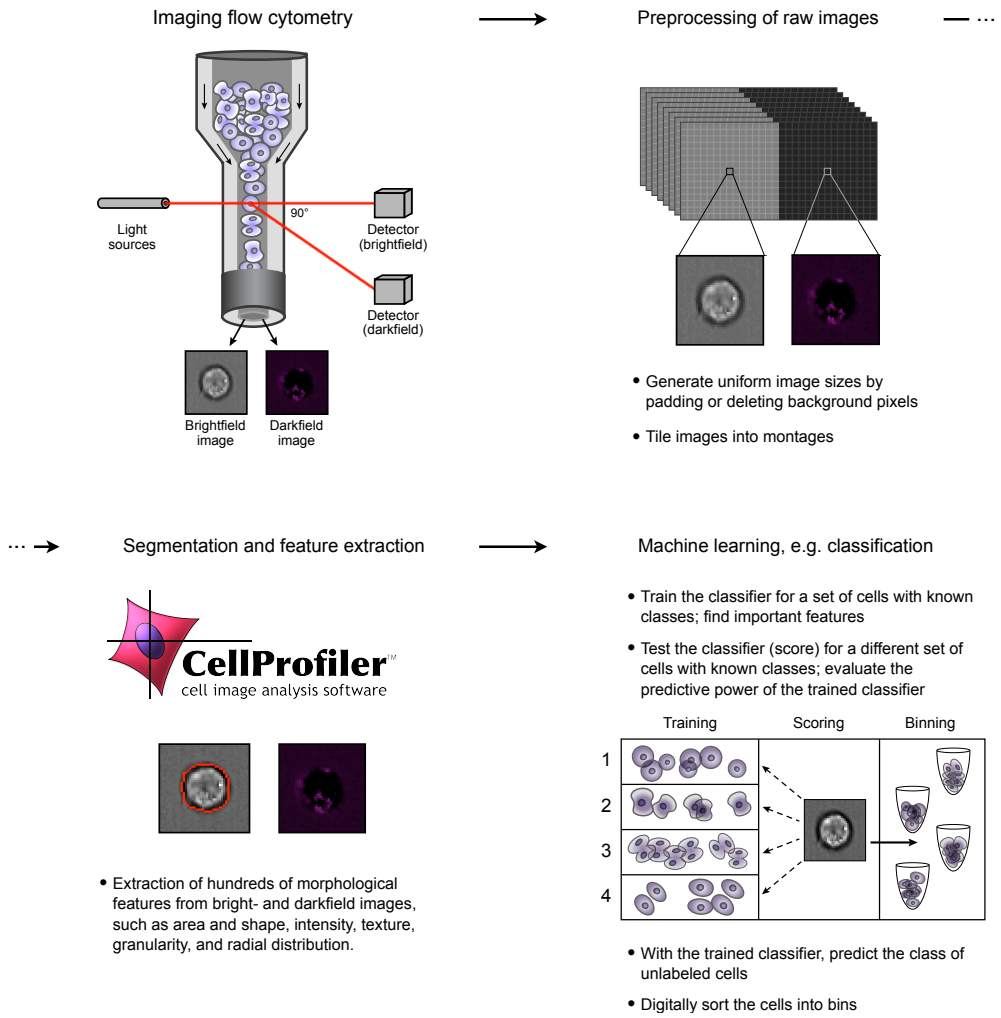


Figure 8.1: First the brightfield and darkfield images of the cells are measured by an imaging flow cytometer. The brightfield and darkfield images depict the light transmitted through the cell and light scattered from the cells within a cone centered at a 90^o angle, respectively. Then the images are preprocessed, where we reshape the images to have their sizes coincide and tile them to montages of 15×15 images. The montages are loaded into the open source image software CellProfiler that we use to segment the cells' brightfield images and to extract morphological features from the images. Lastly, we apply supervised machine learning such as classification. For this purpose we need an annotated set of cells where the actual cell state is known to train the classifier and to test its predictive power. Once the classifier is trained it is used to predict the state of unlabeled cells and to digitally sort the cells into bins. This Figure and its legend is identical with Figure 1 of the author's following publication Blasi et al. (2016b)

cell markers, and frees up the remaining available fluorescence channels of the imaging flow cytometer that can be used to investigate other biological questions.

8.2 Label-free analysis workflow

Image-processing to extract informative features

The first step in the workflow of label-free cell-cycle classification (Figure 8.1) is to acquire brightfield and darkfield images from the cells using an imaging flow cytometer. To allow visual inspection and to optimize the file size for processing, we tile individual cells' brightfield and the darkfield images into 15×15 montages, with up to 225 cells per montage. Then, we load the montages into the open-source imaging software CellProfiler (Kamentsky et al., 2011) for processing. There is sufficient contrast between the cells and the flow media to robustly segment the cells in the brightfield images without the need for any stains. We extract 213 features from the segmented brightfield and the full darkfield image.

The features can be summarized into five categories: size and shape, granularity, intensity, radial distribution, and texture. These image features are the input for supervised machine learning, namely classification and regression (see Section 4.4), which we use to predict each cell's DNA content and the mitotic phases in the cell cycle without the need for any stains. The machine learning algorithms have to be trained on an annotated subset of the investigated cells where the true cell state, i.e. the 'ground truth' is known. The ground truth can be obtained either by manual identification (by a trained biologist or using software tools) or from labeling a subset of the investigated cells with fluorescent stains.

Supervised machine learning to predict cellular phenotypes

We use classification and regression with classification and regression trees (CARTs) in conjunction with a boosting strategy to predict the cell's DNA content and its cell cycle phase (see Section 4.4 for details on boosting with CARTs). In brief, in boosting many 'weak learners' (here short trees with $R = 5$ regions) are subsequently fit to the data when after each step the data becomes re-weighted in order to penalize data points that are not yet well explained by the ensemble of weak learners. Eventually a 'majority vote' of all weak learners leads to the prediction (see e.g. Murphy (2012)).

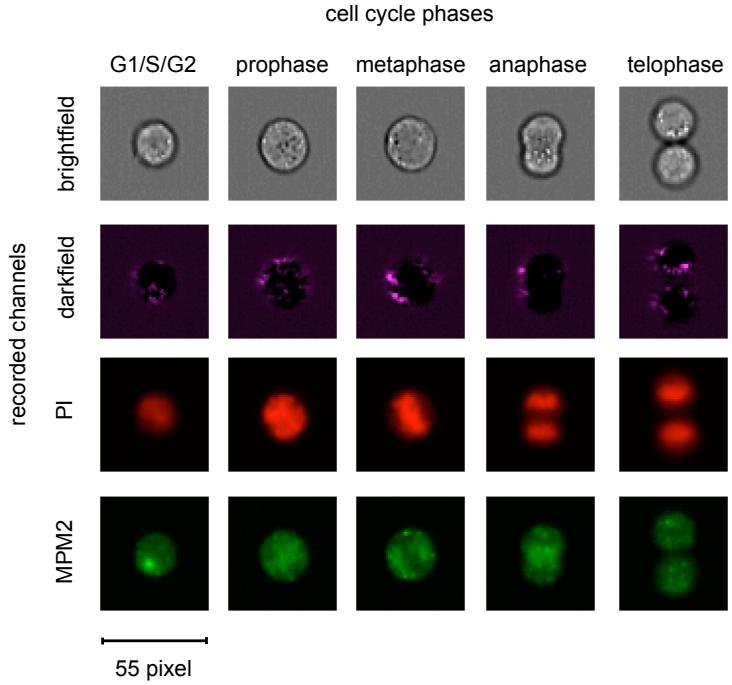


Figure 8.2: Images of Jurkat cells captured by imaging flow cytometry. Typical brightfield, darkfield, PI and MPM2 images of cells in the G1/S/G2 phases, prophase, metaphase, anaphase and telophase of the cell cycle. The size of the images is 55×55 pixels. This Figure and its legend is identical with Supplementary Figure 1 of the author's following publication Blasi et al. (2016b).

For the prediction of the DNA content we use LSboosting (least-squares boosting; Hastie et al. (2009)) as implemented in Matlab's fitensemble routine and for the assignment of the mitotic cell cycle phases we use RUSboosting (boosting with random undersampling; Seiffert et al. (2010)) as also implemented in Matlab's fitensemble routine. In both cases we partition the cells into a training and a testing set. The brightfield and darkfield features of the training set as well as the ground truth of these cells are used to train the ensemble. Once the ensemble is trained we evaluate its predictive power on the testing set. To demonstrate the generalizability of this approach and to obtain error bars for our results the procedure is ten-fold cross-validated.

To prevent overfitting the data and to fix the stopping criterion for the applied boosting algorithms, we performed a five-fold internal cross-validation. To this end, we split up the training set into an internal-training (consisting of 80% of the cells in the training set) and an internal-validation (20% of the cells in the training set) set. We trained the algorithm on the internal-training set with up to 6,000 decision trees. We then

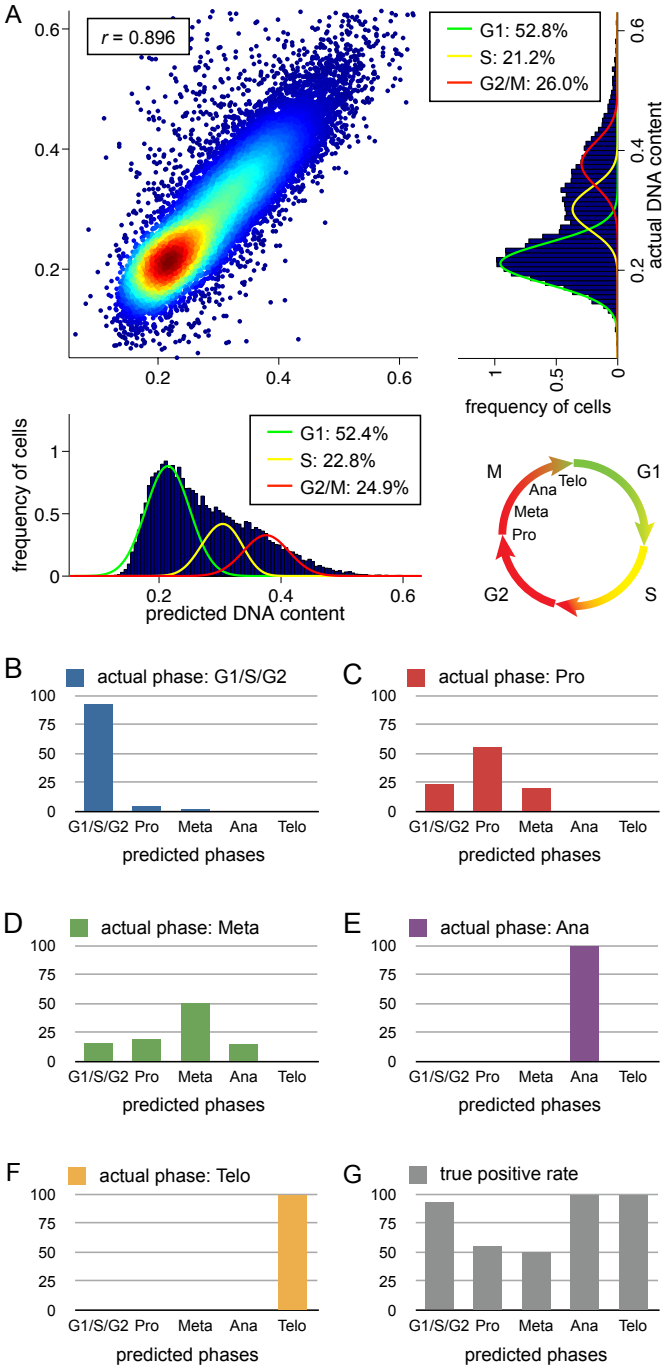


Figure 8.3: Supervised machine learning allows for robust label-free prediction of DNA content and cell cycle phases of Jurkat cells (Figure legend on next page).

Figure 8.3: (Figure on previous page). Supervised machine learning allows for robust label-free prediction of DNA content and cell cycle phases of Jurkat cells. (A) We find a Pearson-correlation of $r = 0.896 \pm 0.007$ between actual DNA content and predicted DNA content (based on regression using brightfield and darkfield morphological features only). We used the Watson pragmatic curve fitting algorithm to specify the fraction of cells in the G1, S and G2 phases. (B-F) For cells that are actually in a particular phase (e.g. (B) shows cells in G1/S/G2), the bar plots show the classification results based on brightfield and darkfield morphological features only (e.g. (B) shows that the few cells in prophase (Pro), metaphase (Meta), anaphase (Ana), and telophase (Telo) are errors). (G) Bar plot of the true positive rates of the cell cycle classification. This Figure and its legend is identical with Figure 2 of the author's following publication Blasi et al. (2016b).

predicted the DNA content/cell cycle phase of the inner-validation set and evaluated the quality of the prediction as a function of the used amount of decision trees. The optimal amount of decision trees is chosen as the one for which the quality of the prediction is best. We repeat this procedure five times and determine the stopping criterion for the whole training set as the average of the five values for the stopping criterion obtained in the internal cross-validation.

8.3 Application of the workflow to biological data sets

Cell-cycle analysis of fixed Jurkat cells

As an initial demonstration of our technique we sought a label-free way to measure important cell cycle phenotypes including a continuous property (a cell's DNA content, from which G1, S and G2 phases can be estimated) and discrete phenotypes (the mitotic phase of a cell: prophase, anaphase, metaphase, and telophase). We used the ImageStream platform to capture images of 32,255 asynchronously growing Jurkat cells (Figure 8.2). As controls, the cells were fixed and stained with Propidium Iodide to quantify DNA content and an MPM2 antibody to identify mitotic cells. These fluorescent markers were used to annotate a subset of the cells with the ground truth needed to train the machine learning algorithms and to evaluate the predictive accuracy of our label-free approach. Since it is infeasible to accurately identify individual cells in the G1, S and G2 phase based only on one nuclear marker (Miltenburger et al., 1987) we do not aim to predict those phases individually but to predict each cell's DNA content. Subsequently, we use the Watson pragmatic curve fitting algorithm (Watson

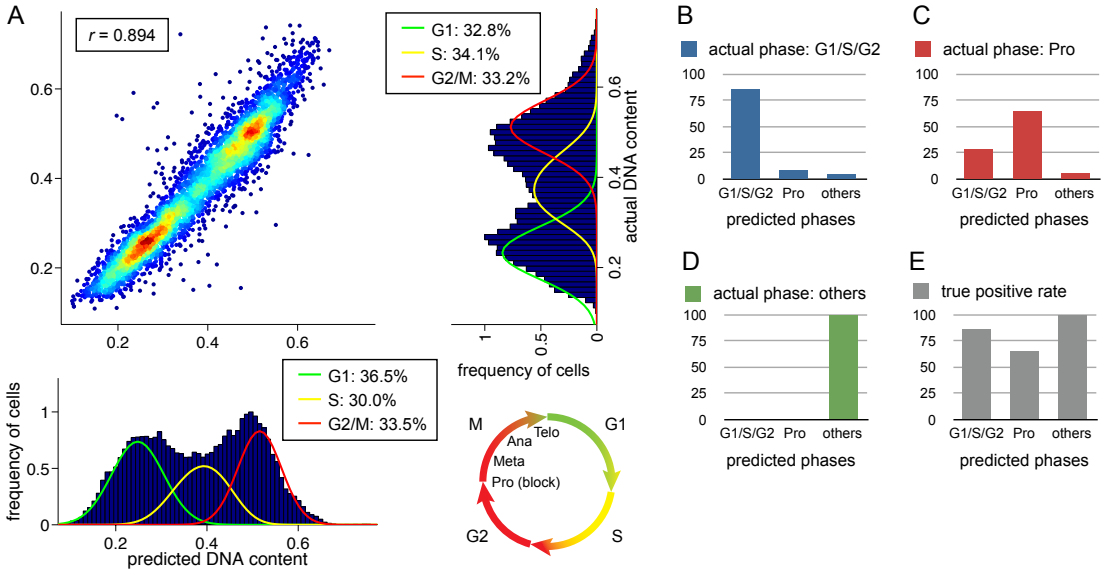


Figure 8.4: Label-free prediction of DNA content and cell cycle phases for fixed Jurkat cells treated with a prophase-blocking agent. (A) Based only on brightfield and darkfield features, we find a Pearson-correlation of $r = 0.894 \pm 0.032$ between actual DNA content and predicted DNA content using regression. We applied the Watson pragmatic algorithm to determine the G1, S and G2/M phases in the DNA histograms. (B-D) For cells that are actually in a particular phase (e.g. (B) shows cells in G1/S/G2), the bar plots show the classification results (e.g. (B) shows that the few cells in prophase (Pro) and the other mitotic phases (others) are errors). Note that we grouped the cells in metaphase, anaphase and telophase into one class since we only detected very little cells in those phases after treatment with the prophase blocking agent. (E) Bar plot of the true positive rates of the cell cycle classification. Using boosting with random undersampling to compensate for class imbalances, we obtain true positive rates of 87.6±2.2% (P), 87.6±2.2% (G1/S/G2) and 100% (others). This Figure and its legend is identical with Figure 3 of the author's following publication Blasi et al. (2016b).

et al., 1987) to estimate the percentage of cells in each of the G1/S/G2M phases based on the predicted DNA content.

Using only cell features measured from brightfield and darkfield images we were able to devise a regression ensemble (using least squares boosting that accurately predicts each cell's DNA content, obtaining a Pearson correlation of $r = 0.896 \pm 0.007$ between predicted and actual nuclear stain intensity (Figure 8.3A). This highly accurate prediction of the DNA content can be used to further categorize G1, S and G2/M cells or to allocate each cell a time position within the cell cycle via the ergodic rate analysis, where cells are sorted according to their DNA content (Kafri et al., 2013). Moreover, we were

able to classify mitotic phases (using random undersampling (Seiffert et al., 2010)) to compensate for the high class imbalance) with true positive rates of $55.4 \pm 6.6\%$ (for prophase), $50.2 \pm 16.3\%$ (for metaphase), 100% (for anaphase and telophase) and $93.1 \pm 0.5\%$ for the non-mitotic phases (Figures 8.3B-G.

We analyzed which features have the most significant contributions for the prediction of both the nuclear stain and the mitotic phases by 'leave one out' cross-validation (data not shown). We find that leaving one feature out has only a minor effect on the results of the supervised machine learning algorithms we used, likely because many features are highly correlated to others. The most important features are intensity, area, shape, and radial distribution of the brightfield images.

Detection of mitotic phase block

The assessment of the therapeutic blocking of the cell cycle (in a particular phase) is of particular importance. We tested the method's ability to predict the DNA content of Jurkat cells treated with 50μM Nocodazole, a mitotic blocking agent. To confirm the magnitude of the block of cells in mitosis we performed three additional replicates demonstrating an average increase of cells in the G2/M phase of $19.0 \pm 11.1\%$ compared with the control. The label-free prediction of the DNA content has a Pearson correlation of $r = 0.894 \pm 0.032$ with the true DNA content (PI is used as a fixed cell nuclear stain to provide the ground truth for the machine learning algorithms) and the percentage of cells in the G1, S and G2/M phases are in excellent agreement (Figure 8.4A).

Therefore the technique is successfully detecting the expected increase in the G2/M cells due to the blocking agent based on the predicted DNA content. Again, we were able to classify mitotic phases and found true positive rates of $67.6 \pm 7.4\%$ (for prophase), 100% (for the other mitotic phases) and $87.6 \pm 2.2\%$ for the non-mitotic phases (Figures 8.4B-E). Treatment of the cells with the mitotic blocking agent lead to an increase in the percentage of prophase cells from 1.88 to 11.07, which is confirmed by comparison with the ground truth (data not shown) and in agreement with the identified magnitude of the block of cells in mitosis.

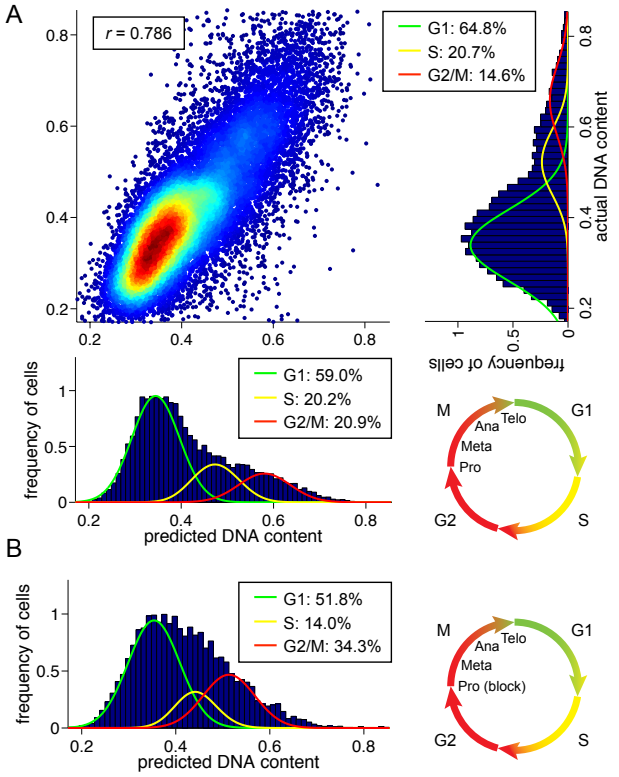


Figure 8.5: Label-free prediction of DNA content for live Jurkat cells and detection of a phase blockage. (A) Supervised machine learning (trained using live cells stained with DRAQ5 to determine the DNA content) allows for robust label-free prediction of the DNA content of live cells based only on brightfield and darkfield images. We find a Pearson-correlation of $r = 0.786 \pm 0.010$ between actual DNA content and predicted DNA content using regression. We believe this reduction in correlation from the value of 0.896 obtained for fixed cells to be a consequence of the greater variability of the uptake of the live DNA dye compared with the staining achieved with fixed cells. Despite the reduction in correlation a value of 0.786 is still high enough to make this a viable method for the cell cycle analysis of live cells. As previously we determine the fraction of cells in the G1, S and G2/M phases using the Watson pragmatic curve fitting algorithm. (B) We predict an increase of 13.4% in the G2/M phase after the cells were treated with 50μM Nocodazole, which is in good agreement with the average increase of $19.0 \pm 11.1\%$ in G2/M as was found for three independent cell populations under the same treatment. The phase-blocked data set was not labeled with any marker. Instead, we trained our machine learning algorithm on the untreated data set, which was labeled with a DRAQ5 DNA stain (see A) and used the trained machine learning algorithm to predict the DNA stain of the blocked cells. This Figure and its legend is identical with Figure 4 of the author's following publication Blasi et al. (2016b).

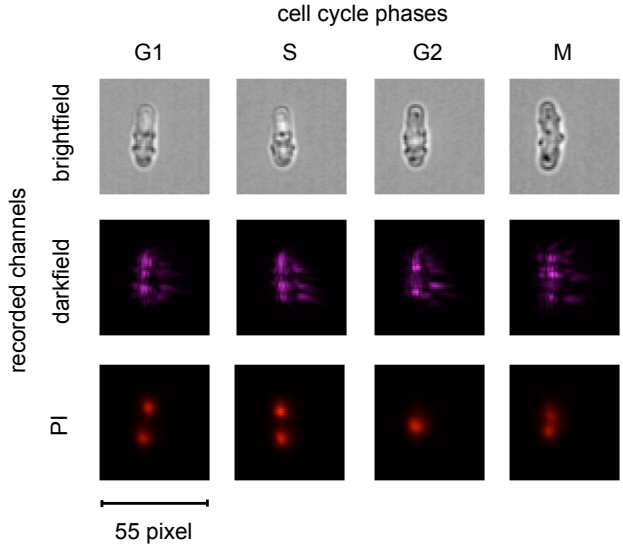


Figure 8.6: Images of fission yeast cells captured by imaging flow cytometry. Typical brightfield, darkfield and PI images of cells in the G1, S, G2 and M phases of the cell cycle. This Figure and its legend is identical with Supplementary Figure 4 of the author's following publication Blasi et al. (2016b)

Cell-cycle analysis of live Jurkat cells and detection of mitotic phase block

Many experimental protocols require live cells rather than fixed. We tested the ability of the technique to detect cell cycle changes in live Jurkat cells. To provide ground truth (that is, the expected cell cycle distribution), the cells were stained with DRAQ5, a live cell DNA stain (Figure 8.5A). Like most live-cell-compatible DNA stains, DRAQ5 is not an ideal marker because of the variability of uptake of the dye in live cells (Yuan et al., 2004), nonetheless, we obtain a Pearson correlation of $r = 0.786 \pm 0.010$ for predicting the DNA content of untreated cells.

With a regression ensemble trained on the stained live cells we are also able to predict the effect of treatment with a phase blocking agent on an entirely unstained data set (Figure 8.5B). We detect an increase of cells in the G2/M phase from 20.9% to 34.3% when the cells are treated with 50μM Nocodazole; this is consistent with the average increase of $19.0 \pm 11.1\%$ obtained from repeating the phase block experiments with stained cells.

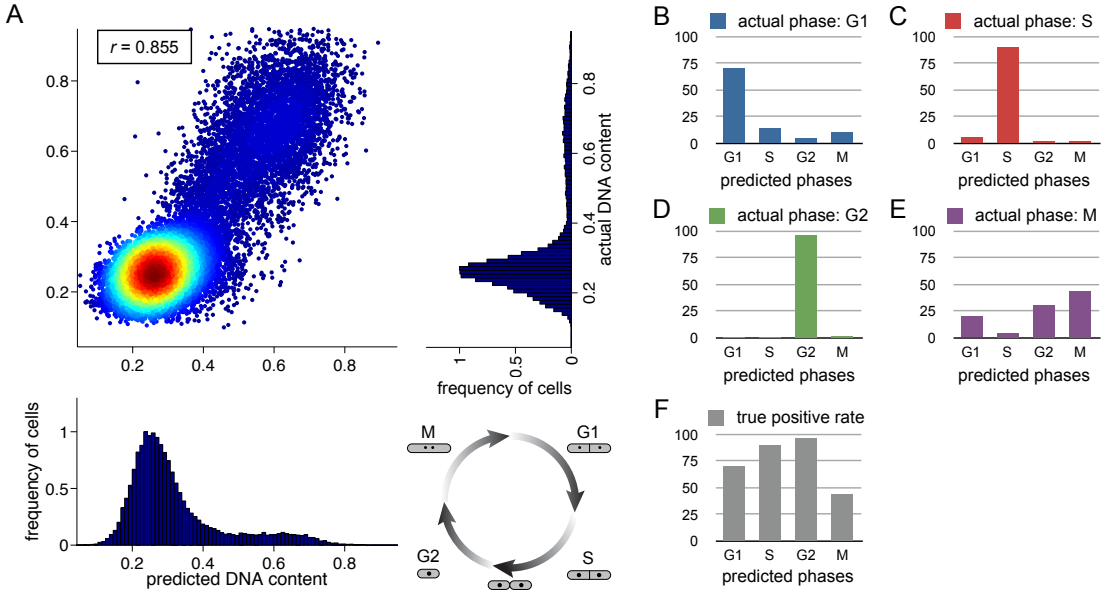


Figure 8.7: Label-free prediction of DNA content and cell cycle phases for fission yeast cells. (A) Based only on brightfield and darkfield features, we find a Pearson-correlation of $r = 0.855 \pm 0.006$ between actual DNA content and predicted DNA content using regression. Note that the fission yeast cell cycle is different from the Jurkat cell cycle since the two daughter cells divide between the S and G2 phases (and not at the end of M phase as is the case for Jurkat cells). (B-E) For cells that are actually in a particular phase (e.g. (B) shows cells in G1), the bar plots show the classification results (e.g. (B) shows that the cells in S, G2 and M are errors). (F) Bar plot of the true positive rates of the cell cycle classification. Using boosting with random undersampling to compensate for class imbalances, we obtain true positive rates of $70.3 \pm 4.1\%$ (G1), $90.13 \pm 1.3\%$ (S), $96.8 \pm 0.1\%$ (G2) and $43.9 \pm 4.7\%$ (M). This Figure and its legend is identical with Figure 5 of the author's following publication Blasi et al. (2016b)

Cell-cycle analysis of fission yeast

To explore the generality of our method for other cell types, we tested it on another species, fission yeast (Figure 8.6). The yeast cells were fixed and stained with PI to measure the DNA content of each cell; subsequently the cells were assigned to the G1, S, G2 or M phase by manually gating on image based metrics from the PI channel of the Image-stream data (Patterson et al., 2015), which provided the ground truth. The label free regression predicts a DNA content with a Pearson correlation of $r = 0.855 \pm 0.006$ (Figure 8.7A) and a classification accuracy of $70.2 \pm 2.2\%$ (G1), $90.1 \pm 1.1\%$ (S), $96.8 \pm 0.3\%$ (G2) and 44.0 ± 8.4 (M) (Figures 8.7B-F).

8.4 Discussion

We demonstrate here that it is possible to determine a cell population's DNA content and mitotic phases based entirely on features extracted from cells' brightfield and darkfield images, as obtained in high-throughput via imaging flow cytometry. The method requires an annotated dataset to train the machine learning algorithms, either by staining a subset of the investigated cells with markers, or by visual inspection and assignment of cell classes of interest. Once the machine-learning algorithm is trained for a particular cellular phenotype, the consistency of imaging flow cytometry allows high-throughput scoring of unlabeled cells for discrete and well-defined phenotypes (e.g. mitotic cell cycle phases) and continuous properties (e.g. DNA content).

The same basic strategy can be readily adapted to measure other phenotypes, making this a generally useful approach for label-free, single-cell phenotyping in the modern biological laboratory. The method can also be used retrospectively on datasets that do not have the necessary stains for phenotype identification, providing an annotated dataset is available to train the algorithms. While current imaging flow cytometers do not have physical cell-sorting capabilities and for now our approach is suited to experimental contexts where samples are analyzed only, this approach may offer the possibility to entirely avoid any fluorescent stain and opens up the perspective for a new generation of image flow cytometers that could operate without fluorescence channels.

Label-free identification of phenotypes enables continuous, non-destructive monitoring of cell samples, minimizes potentially confounding influences of the stains on the cells, and maximizes available fluorescence channels to investigate biological questions such as the search for novel hallmarks in cell cycle (Zuleta et al., 2014), the identification of stem and progenitor cells (Xia and Wong, 2012), or the proliferation of cancer cells (Chan et al., 2012).

Chapter 9

Summary and Outlook

Men say they know many things;
But lo! they have taken wings, –
The arts and sciences,
And a thousand appliances;
The wind that blows
Is all that any body knows.

Henry David Thoreau [VII]

9.1 Summary

In this thesis we developed new and applied existing mathematical methods to analyze cellular heterogeneity in a data-driven way. We started by introducing the relevant biological background where we explained how transcriptional regulation leads to heterogeneity among cells with the same DNA and by giving an overview over currently available biological data (Chapter 2). We then derived mathematical frameworks that are capable to describe biological processes (Chapter 3). In Chapter 4 we pointed out the importance of conjoining the mathematical models with biological data and presented several probabilistic approaches that are suitable for this task.

In the original contributions of this thesis, we analyzed cellular heterogeneity on two levels (see Figure 9.1). On the one side, we found a new workflow to better identify cellular phenotypes from a heterogeneous mixture of cells by applying mathematical methods from supervised machine learning (Chapter 8; Figure 9.1 A), on the other

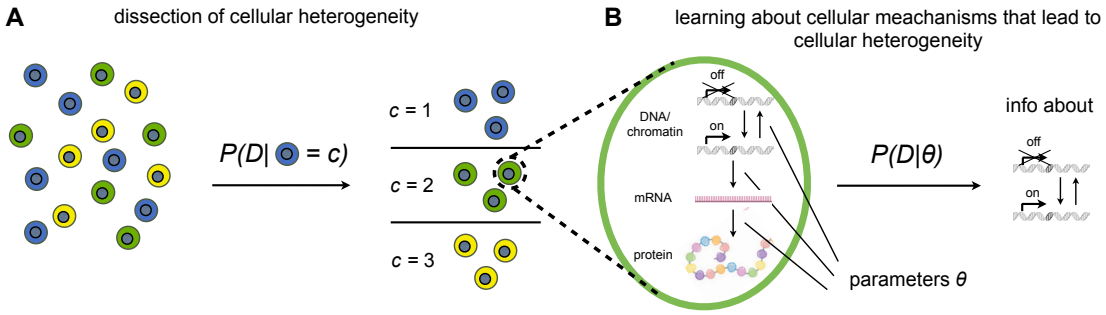


Figure 9.1: Analyzing mechanisms that lead to different cellular phenotypes and dissecting cellular heterogeneity. (A) Three different cellular phenotypes (green, blue and orange) form a heterogeneous cell population. Methods from machine learning (see Chapters 4.3 and 4.4) such as classification can be used to dissect the heterogeneity and to identify the different cellular phenotypes given data D from the cells. (B) Within a cell genes are transcribed into mRNA and translated into proteins (see Section 2.1). We can formulate mathematical models that describe this biological process in a parametrized way (see Chapter 3). Given biological data D and the parameters of the model θ we can infer information about the regulation of gene expression by formulating and optimizing the likelihood $P(D|\theta)$ (see Chapter 4.1). We argue that differences in gene expression are the main source of cellular heterogeneity (dashed line; see Section 2.1). Parts of the Figure are adopted from O'Connor and Adams (2010).

side, we derived and applied mathematical models to investigate biological processes in the cell that lead to the formation of different cellular phenotypes, such as chromatin modifications (see Chapter 6) and gene regulation (Chapters 5 and 7; Figure 9.1 B).

More specifically:

- In Chapter 5 we presented a novel mathematical framework, cgcorrect, that takes confounding sources of variability for both dimension reduction and steady-state distribution analysis into account. By applying our method to single cell qPCR data, we found that many genes that are important during hematopoiesis are rather continuously expressed than bursty when correcting for this variability.

- In Chapter 6 we found evidence for motif-specific histone H4 acetylation by comparing a large set of mathematical models to data from LC-MS. Our approach allowed us predict acetylation pathways and could be used to find candidate enzymes that may be involved in the pathways.

- In Chapter 7 the application of transfer entropy to time-series data of the proteins PU.1 and Gata-1 that were measured during hematopoiesis revealed that PU.1

transfers information Gata-1 in cells that eventually commit to the granulocyte-monocyte lineage.

- In Chapter 8 we provided a new workflow to identify the cell's position in the cell cycle for imaging flow cytometry data. We could show that it is possible to avoid fluorescent stains and that morphological features of the brightfield and darkfield images are sufficient to obtain a precise identification.

9.2 Outlook

An iterative loop between the identification of cellular phenotypes and understanding their underlying mechanisms

In this thesis we outlined how differences in gene expression eventually lead to different cellular phenotypes from a bottom-up perspective. For diagnostic purposes, however, it turns out to be fruitful to reverse this perspective:
The recent advent of high-throughput single-cell technologies has made the screening of large numbers of single cells possible. This new data enables a convenient identification of cellular phenotypes (see e.g. Grün et al. (2015) for RNA-seq data or Chapter 8 for imaging flow cytometry data) and is anticipated to lead to a new era in diagnostic research, where malignant cells can be identified in tissue samples (Sandberg (2014); see also Figure 9.1 B) without any a priori knowledge about the underlying mechanisms.
It is the iterative loop between the identification of (malignant) cellular phenotypes and the quest for an improved understanding of cellular processes that may contribute to better understand the formation of diseases and serve as a starting point for the development of therapies.

On the way to single cell epigenetics

In Chapter 6 we analyzed the abundances of histone acetylation states that were measured by untargeted mass spectrometry from bulk samples (Feller et al., 2015). Single cell analysis has already become possible in many areas of epigenetics such as DNA methylation, e.g. via bisulfite sequencing, and chromatin modifications via targeted antibody assays (see Bheda and Schneider (2014) for a review). However, the quality assessment of antibody binding properties is still heavily debated in literature where questions concerning the specificity of targeted approaches are raised (see Kungulovski

et al. (2014) for a discussion). While an untargeted approach without the need for antibodies avoids these issues, current mass spectrometry methods are only on the way to reach sensitivities that are suitable for the analysis of small sample volumes that are given by single cells (Lombard-Banek et al., 2016). A suitable technological framework for future applications is the integration of microfluidics with mass spectrometry to lab-on-a-chip mass spectrometry (LOC-MS) as proposed by Oedit et al. (2015).
Untargeted mass spectrometry based approaches on the other hand do not contain local information in terms of the genomic loci where the modifications occur. This information, however, is necessary to construct more detailed models of gene expression where additional epigenetic layers can be incorporated as compared to the simplified three stage model of gene expression (as discussed in Section 2.1). Since chromatin modifications can be inherited from one cell to the other they provide a fundamental mechanism to guide linage choices in differentiating cells (see Margueron and Reinberg (2010)). Therefore, a major future goal will be the integration of single cell epigenetics data that are obtained with orthogonal experimental techniques to provide new insights into the biological mechanisms that drive cellular heterogeneity.

Linking the scales: towards data-driven multi-scale models of single cells

In this thesis, we analyzed biological data that stem from several scales in the cell (i.e. chromatin, mRNA, protein and morphology). A long term goal for bio-mathematics and computational biology is to integrate data from multiple scales.
Since current measurements are often invasive to the cells (including killing the cells) it is often infeasible to obtain data from more than one scale per measurement. For a multi-scale model it is therefore necessary to combine data sets from different experimental methods. As previously discussed, however, the data sets have different properties (e.g. time-series or snapshot data; single cell or population based) depending on the experimental technique. Moreover, since the experimental methods are often not broadly established yet (but only conducted by a few highly specialized labs) the investigated cell lines or other experimental conditions may differ rendering a direct comparison between the data sets into a challenging task.
Nevertheless, both future experimental and mathematical advances may enable to integrate the data from some or all of these scales into multi-scale models to improve the predictive power of current mathematical approaches.

Beyond gene expression in single cells

In this thesis, we focussed on the analysis of heterogeneity among isolated single cells. In their natural environment, however, mammalian cells are usually organized in tissue and organs and constantly exchange signals with their surrounding cells (see e.g. Berridge (2016)). This leads to a constant in and outflow of molecular species (such as nutritions or hormones) that may also influence the gene expression in the cell (see e.g. Becker et al. (2010)). In order to gain a more comprehensive understanding of the biological processes that lead to cellular heterogeneity it is crucial to incorporate cell-cell interactions and their natural environment into the current analysis frameworks.

The need for novel mathematical methods

So far already many achievements have been made by the bio-mathematical community that improved our understand of biological processes. With single-cell high-throughput experiments becoming more and more established and the measured number of samples continuously increasing (see e.g. Macosko et al. (2015) and Klein et al. (2015) for single-cell RNA-sequencing) the demand for suitable mathematical methods to analyze this data is still growing.
Since a better understanding of cellular mechanisms and their malfunctioning has direct implications for the development of new therapeutical approaches the impact of new bio-mathematical methods is likely high. In the light of the aforementioned open challenges it becomes clear that the role of bio-mathematics will become more and more important for the data-driven analysis of cellular systems.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Water, P. (2006). *The Molecular Biology of the Cell*. Garland Science, New York, USA, fifth edition. 11

Allahverdi, A., Yang, R., Korolev, N., Fan, Y., Davey, C. A., Liu, C.-F., and Nordenskiöld, L. (2011). The effects of histone H4 tail acetylations on cation-induced chromatin folding and self-association. *Nucleic Acids Research*, 39(5):1680–1691. 87

Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. CRC Press Taylor and Francis Group, Boca Raton, FL, first edition. 17

Amigo2 (2016). *Amigo2* searched for gene ontology of the cell cycle (go:0007049) filtered for homo sapiens. `http://nakama.berkeleybop.org/amigo/term/GO:0007049`. Accessed: 2016-01-19. 17

Angel, A., Song, J., Dean, C., and Howard, M. (2011). A polycomb-based switch underlying quantitative epigenetic memory. *Nature*, 476:105–108. 88

Annunziato, A. (2008). DNA packaging: nucleosomes and chromatin. *Nature Education*, 1:26. 14

Ansorge, W. J. (2009). Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4):195–203. 21

Arinobu, Y., Mizuno, S., Chong, Y., Shigematsu, H., Iino, T., Iwasaki, H., Graf, T., Mayfield, R., Chan, S., Kastner, P., and Akashi, K. (2007). Reciprocal activation of Gata-1 and PU.1 marks initial specification of hematopoietic stem cells into myeloerythroid and myelolymphoid lineages. *Cell Stem Cell*, 1(4):416–427. 26, 109

Avery, O. T., MacLeod, C. M., and McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of *pneumococcal* types. *Journal of Experimental Medicine*, 79:137–158. 12

Bandura, D. R., Baranov, V. I., Ornatsky, O. I., Antonov, A., Kinach, R., Lou, X., Pavlov, S., Vorobiev, S., Dick, J. E., and Tanner, S. D. (2009). Mass cytometry: technique for real time

single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical Chemistry*, 81:6813–6822. 23

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nature Reviews Genetics*, 13:552–564. 53

Basiji, D. A., Ortyn, W. E., Liang, L., Venkatachalam, V., and Morissey, P. (2007). Cellular image analysis and imaging by flow cytometry. *Clinics in Laboratory Medicine*, 27:653–670. 24, 121

Battich, N., Stoeger, T., and Pelkmans, L. (2013). Image-based transcriptomics in thousands of single human cells at single-molecule resolution. *Nature Methods*, 10:1127–1133. 82

Becker, V., Schilling, M., Bachmann, J., Baumann, U., Raue, A., Maiwald, T., Timmer, J., and Klingmüller, U. (2010). Covering a broad dynamic range: information processing at the erythropoietin receptor. *Science*, 328(5984):1404–1408. 137

Bendall, S. C., Davis, K. L., Amir, E. D., Tadmor, M. D., Simonds, E. F., Tiffany, J. C., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*, 157(3):714–725. 117

Bengio, Y., Goodfellow, I. J., and Courville, A. (2015). *Deep Learning*. Book in preparation for MIT Press. 62

Benson, L. J., Phillips, J. A., Gu, Y., Parthun, M. R., Hoffman, C. S., and Annunziato, A. T. (2007). Properties of the type B histone acetyltransferase Hat1: H4 tail interaction, site preference, and involvement in DNA repair. *The Journal of Biological Chemistry*, 282:836–842. 87

Berridge, M. (2016). Cell Signalling Biology. `http://www.cellsignallingbiology.org/csb/`. Accessed: 2016-01-19. 137

Bertoli, C., Skothelm, J. M., and de Bruin, R. A. (2013). Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology*, 14:518–528. 82

Bheda, P. and Schneider, R. (2014). Epigenetics reloaded: the single-cell revolution. *Trends in Cell Biology*, 24(11):712–723. 86, 135

Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., and Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*, 40(6):463–471. 1

Bjornson, Z. B., Nolan, G. P., and Fantl, W. J. (2013). Single-cell mass cytometry for analysis of immune system functional states. *Current Opinion in Immunology*, 25(4):484–494. 86

Blasi, T., Feller, C., Feigelman, J., J., H., Imhof, A., Theis, F. J., Becker, P. B., and Marr, C. (2016a). Combinatorial histone acetylation patterns are generated by motif-specific reactions. *Cell Systems*, 2:49–58. 91, 96, 97, 100, 102, 105

Blasi, T., Hennig, H., Summers, H. D., Theis, F. J., Cerveira, J., Patterson, J. O., Davies, D., Filby, A., Carpenter, A. E., and Rees, P. (2016b). Label-free cell cycle analysis for high-throughput imaging flow cytometry. *Nature Communications*, 6:10256. 122, 124, 126, 127, 129, 130, 131

Breiman, L. (2001). Random Forests. *Machine Learning*, 45:5–32. 62

Brennecke, P., Anders, S., Kim, J. K., Kolodziejczyk, A. A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S. A., Marioni, J. C., and Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, 10(11):1093–1095. 21, 67, 70

Brown, M. and Wittwer, C. (2000). Flow cytometry: principles and clinical applications in hematology. *Clinical Chemistry*, 46:1221–1229. 120

Buettner, F., Moignard, V., Göttgens, B., and Theis, F. J. (2014). Probabilistic PCA of censored data: accounting for uncertainties in the visualization of high-throughput single-cell qPCR data. *Bioinformatics*, 30(13):1867–1875. 67

Buettner, F., Natarajan, K. M., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, 33:155–160. 70

Buettner, F. and Theis, F. J. (2012). A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics*, 28(18):i626–i632. 55, 66

Buggenthin, F., Marr, C., Schwarzfischer, M., Hoppe, P. S., Hilsenbeck, O., Schroeder, T., and Theis, F. J. (2013). An automatic method for robust and fast cell detection in bright field images from high-throughput microscopy. *BMC Bioinformatics*, 14:297. 113

Burda, P., Laslo, P., and Stopka, T. (2013). The role of PU.1 and GATA-1 transcription factors during normal and leukemogenic hematopoiesis. *Leukemia*, 24:1249–1257. 26

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., and Lewis, S. (2009). AmiGO Hub, web presence working group. AmiGO: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289. 17

CellProfiler (2016). Imaging flow cytometry analysis using CellProfiler. `http://www.cellprofiler.org/imagingflowcytometry`. Accessed: 2016-01-19. 120

Chan, K. S., Koh, C. G., and Li, H. Y. (2012). Mitosis-targeted anti-cancer therapies: where they stand. *Cell Death & Disease*, 3:e411. 27, 132

Chen, A. Y., Yu, C., Gatto, B., and Liu, L. F. (1993). DNA minor groove-binding ligands: a different class of mammalian DNA topoisomerase inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 90(17):8131–8135. 120

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, 96:270–281. 51

Chickarmane, V., Enver, T., and Peterson, C. (2009). Computational modeling of the hematopoietic erythroid-myeloid switch reveals insights into cooperativity, priming, and irreversibility. *PLoS Computational Biology*, 5:e1000268. 109

Citri, A., Pang, Z. P., Südhof, T. C., Wernig, M., and Malenka, R. C. (2012). Comprehensive qPCR profiling of gene expression in single neuronal cells. *Nature Protocols*, 7(1):118–127. 66

Clancy, S. (2008). DNA transcription. *Nature Education*, 1:41. 14, 22

Coifman, R. R., Lafon, S., Lee, A. B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S. W. (2006). Geometric diffusions as tools for harmonic analysis and structure definition of data: diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431. 57

Conlon, I. and Raff, M. (2003). Differences in the way a mammalian cell and yeast cells coordinate cell growth and cell-cycle progression. *Journal of Biology*, 2:7. 71, 82

Cooper, S. (2004). Control and maintenance of mammalian cell size. *BMC Cell Biology*, 5:35. 82

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297. 60

Crick, F. H. C. (1958). The biological replication of macromolecules. *Symp. Soc. Exp. Biol.*, XII:138. 11

Crick, F. H. C. (1970). Central dogma of molecular biology. *Nature*, 227:561–563. 11

Crick, F. H. C., Leslie, B., Brenner, S., and Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, 192:1227–1232. 15

Darbellay, A. G. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *Transactions on Information Theory*, 45(4):1315–1321. 110

Darzynkiewicz, Z. and Huang, X. (2004). Analysis of cellular DNA content by flow cytometry. *Current Protocols in Immunology*, 5:5.7. 120

Dickson, M. A. and Schwartz, G. K. (2009). Development of cell-cycle inhibitors for cancer therapy. *Current Ontology*, 16(2):36–43. 27

Dion, M. F., Altschuler, S. J., Wu, L. F., and Rando, O. J. (2005). Genomic characterization reveals a simple histone H4 acetylation code. *Proceedings of the National Academy of Sciences of the United States of America*, 102(15):5501–5506. 87

Dodd, B., Micheelsen, M. A., Sneppen, K., and Thon, G. (2007). Theoretical analysis of epigenetic memory by nucleosome modification. *Cell*, 129(4):813–822. 88

Dose, A., Liokatis, S., Theillet, F. X., Selenko, P., and Schwarzer, D. (2011). NMR profiling of histone deacetylase and acetyl-transferase activities in real time. *ACS Chemical Biology*, 6(5):419–424. 87

Duff, C., Smith-Miles, K., Lopes, L., and Tian, T. (2012). Mathematical modelling of stem cell differentiation: the PU.1–Gata-1 interaction. *Journal of Mathematical Biology*, 64:449–468. 26, 109

Dunlop, M. J., Cox, R. S., Levine, J. H., Murray, R. M., and Elowitz, M. B. (2008). Regulatory activity revealed by dynamic correlations in gene expression. *Nature Genetics*, 40(12):1493–1498. 116, 117

Elf, J. and Ehrenberg, M. (2003). Fast evaluation of fluctuations in biochemical networks with the linear noise approximation. *Genome Research*, 13(11):2475–2484. 34

Eliceiri, K. W., Berthold, M. R., Goldberg, I. G., Ibanez, L., Manjunath, B. S., Martone, M. E., Murphy, R. F., Peng, H., Plant, A. L., Roysam, B., Stuurman, N., Swedlow, J. R., Tomancak, P., and Carpenter, A. E. (2012). Biological imaging software tools. *Nature Methods*, 9:697–710. 24, 121

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186. 18

Engblon, S. (2006). Computing the moments of high dimensional solutions of the master equation. *Applied Mathematics and Computation*, 180(2):498–515. 34

Everetts, A. G., Zee, B. M., Dimaggio, P. A., Gonzales-Cope, M., Coller, H. A., and Garcia, B. A. (2013). Quantitative dynamics of the link between cellular metabolism and histone acetylation. *The Journal of Biological Chemistry*, 288(17):12142–12151. 88

Fan, Y. J. and Kamath, C. (2014). On the selection of dimension reduction techniques for scientific applications. *Annals of Information Systems*, 17:91–121. 57

Farlik, M., Sheffield, N. C., Nuzzo, A., Datlinger, P., Schönegger, A., Klughammer, J., and Bock, C. (2015). Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. *Cell*, 10(8):1386–1397. 86

Feller, C., Formé, I., Imhof, A., and Becker, P. B. (2015). Global and specific responses of the histone acetylome to systematic perturbations. *Molecular Cell*, 57:559–571. 85, 86, 87, 88, 89, 91, 93, 102, 103, 105, 135

Filby, A., Perucha, E., Summers, H. D., Rees, P., Chana, P., Heck, S., Lord, G. M., and Davies, D. (2011). An imaging flow cytometric method for measuring cell division history and molecular symmetry during mitosis. *Cytometry Part A*, 79:496–506. 27, 121

Filipczyk, A., Marr, C., Hastreiter, S., Feigelman, J., Schwarzfischer, M., Hoppe, P. S., Loeffler, D., Kokkaliaris, K. D., Endele, M., Schauberger, M., Skylaki, S., Hasenauer, J., Anastassiadis, K., Theis, F. J., and Schroeder, T. (2015). Network plasticity of pluripotency transcription factors in embryonic stem cells. *Nature Cell Biology*, 17:1235–1246. 109

Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156. 63

Galloway, J. L., Wingert, R. A., Thisse, C., Thise, B., and Zon, L. I. (2005). Loss of gata1 but not gata2 converts erythropoiesis to myelopoiesis in zebrafish embryos. *Developmental Cell*, 8:109–116. 109

Garcia, B., Hake, S. B., Diaz, R. L., Kauer, M., Morris, S. A., Recht, J., Shabanowitz, J., N., M., Strahl, B. D., Allis, C. D., and Hunt, D. F. (2007). Organismal differences in post-translational modifications in histones H3 and H4. *Journal of Biological Chemistry*, pages 7641–7655. 103

Gelman, A. and Meng, X. (1998). Simulating normalization constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, 7(4):457–511. 51

Gentle, J. E. (2004). *Random Number Generation and Monte Carlo Methods*. Springer, New York, USA, second edition. 75

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434. 2

Gillespie, D. T. (1992). A rigorous derivation of the chemical master equation. *Physica A*, 188:404–425. 2, 30, 32, 33, 44

Gillespie, D. T. (2000). The chemical Langevin equation. *The Journal of Chemical Physics*, 113:297–306. 36

Gillespie, D. T. (2007). Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:53–55. 30, 34

Glusman, G., Caballero, J., Robinson, M., Kutlu, B., and Hood, L. (2013). Optimal scaling of digital transcriptomes. *PLoS ONE*, 8:e77885. 70

Grün, D., Lyubimova, A., Kester, L., Wiebrans, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525:251–255. 135

Grunstein, M. (1997). Histone acetylation in chromatin structure and transcription. *Nature*, 389:349–353. 85

Gumpinger, A. (2015). *Time series analysis of gene expression data during stem cell decision making using Transfer Entropy.* http://www.helmholtz-muenchen.de/icb/teaching/completed-theses/index.html. 7, 108, 113, 114, 115, 116

Guo, G., Huss, M., Tong, G. Q., Wang, C., Sun, L. L., Clarke, N. D., and Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental Cell*, 18(4):675–685. 70

Haghverdi, L., Buettner, F., and Theis, F. J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*, 31(18):2989–2998. 57

Hans, F. and Dimitrov, S. (2001). Histone H3 phosphorylation and cell division. *Oncogene*, 20:3021–3027. 120

Hasenauer, J. and Theis, F. J. (2013). *Parameter inference for stochastic and deterministic dynamic biological processes.* Lecture Notes. 52, 53

Hastie, T., Tibishirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning.* Springer, New York, USA, second edition. 55, 62, 63, 124

Hatano, A., Chiba, H., Moesa, H. A., Taniguchi, T., Nagaie, S., Yamanegi, K., Takai-Igarashi, T., Tanaka, H., and Fujibuchi, W. (2011). Cellpedia: a repository for human cell information for cell studies and differentiation analyses. *Database*, 2011. 1

Herzenberg, L. A., Parks, D., Sahaf, B., Perez, O., Roederer, M., and Herzenberg, L. (2002). The history and future of the fluorescence activated cell sorter and flow cytometry: a view from stanford. *Clinical Chemistry*, 48(10):1819–1827. 23

Hindson, B. J., Ness, K. D., Masquelier, D. A., Belgrader, P., Heredia, N. J., Makarewicz, A. J., Bright, I. J., Lucero, M. Y., Hiddessen, A. L., Legler, T. C., et al. (2011). High-throughput droplet digital pcr system for absolute quantitation of dna copy number. *Analytical chemistry*, 83(22):8604–8610. 69

Hoppe, P. S., Coutu, D. L., and Schroeder, T. (2014). Single-cell technologies sharpen up mammalian stem cell research. *Nature Cell Biology*, 16:919–927. 3, 19, 108

Hoppe, P. S., Schwarzfischer, M., Loeffler, D., Kokkaliaris, K. D., Hilsenbeck, O., Moritz, N., Endele, M., Filipczyk, A., Rieger, M. A., Marr, C., Strasser, M., Schauberger, B., Burtscher, I., Ermakova, O., Bürger, A., Lickert, H., Nerlov, C., Theis, F. J., and Schroeder, T. (2016). Random pu.1 / gata1 protein ratios do not induce early myeloid lineage choice. *submitted.* 26, 109, 113

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441. 56

Huang, H., Lin, S., Garcia, B. A., and Zhao, Y. (2015). Quantitative proteomic analysis of histone modifications. *Chemical Reviews*, 115(6):2376–2418. 23

Huang, S. (2009). Reprogramming cell fates: reconciling rarity with robustness. *BioEssays*, 31(5):546–560. 37

Huang, S., Guo, Y. P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, 305(2):695–713. 109

Hug, S. (2015). *From low-dimensional model selection to high-dimensional inference: tailoring Bayesian methods to biological dynamical systems*. PhD Thesis, Garching, Germany. 51, 53

International-Human-Genome-Sequencing-Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature*, 431:931–945. 12

Islam, S., Kjallquist, U., Moliner, A., Zajac, P., Fan, J. B., Lonnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, 21(7):1160–1167. 66, 69

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166. 66, 69

Jackson, R. J., Hellen, C. U. T., and Pestova, T. V. (2010). The mechanism of eukaryotic translation and principles of its regulation. *Nature Reviews Molecular Biology*, 11:113–127. 11

Jahnke, T. and Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1):1–26. 34

Jennuwein, T. and Allis, C. D. (2001). Translating the histone code. *Science*, 293:1074–1080. 87

Jensen, O. N. (2004). Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. *Current Opinions in Chemical Biology*, 8:33–41. 2

Jones, T., Carpenter, A. E., Lamprecht, M. R., Moffat, J., Silver, S. J., Grenier, J. K., Castoreno, A. B., Eggert, U. S., Root, D. E., Golland, P., and Sabatini, D. M. (2009). Scoring diverse cellular morphologies in image-based screens with iterative feedback and machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 106:1826–1831. 121

Joshi, P., Greco, T. M., Guise, A. J., Luo, Y., Yu, F., Nesvizhskii, A. I., and Cristea, I. M. (2013). The functional interactome landscape of the human histone deacetylase family. *Molecular Systems Biology*, 9:672. 103

Kafri, R., Levy, J., Ginzberg, M. B., Oh, S., Lahav, G., and Kirschner, M. W. (2013). Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. *Nature*, 494:480–483. 70, 127

Kamentsky, L., Jones, T. R., Fraser, A., Bray, M. A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W., and Carpenter, A. E. (2011). Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics*, 27:1179–1180. 121, 123

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. 52

Katan-Khaykovich, Y. and Struhl, K. (2002). Dynamics of global histone acetylation and deacetylation in vivo: rapid restoration of normal histone acetylation status upon removal of activators and repressors. *Genes & Development*, 16(6):743–752. 91

Keller, P. J., Schmidt, A. D., Wittbrodt, J., and Stelzer, E. H. K. (2008). Reconstruction of zebrafish early embryonic development by scanned light sheet microscopy. *Science*, 322(5904):1065–1069. 24

Kim, J. K. and Marioni, J. C. (2013). Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology*, 14(1):R7. 66, 81, 82

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshikin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, 161:1187–1201. 137

Konry, T., S., S., Sabhachandani, P., and Cohen, N. (2016). Innovative tools and technology for analysis of single cells and cell-cell interactions. *Annual Review of Biomedical Engineering*, 18:259–284. 21

Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science*, 184(4139):868–871. 12

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705. 87

Kreutz, C., Rodriguez, M. M. B., Maiwald, T., Seidl, M., Blum, H. E., Mohr, L., and Timmer, J. (2007). An error model for protein quantification. *Bioinformatics*, 23(20):2747–2753. 93

Krishnapuram, B., Carin, L., Figueiredo, M. A. T., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–968. 60

Kueh, H. Y., Champhekhar, A., Nutt, S. L., Elowitz, M. B., and Rothenburg, E. V. (2013). Positive feedback between PU.1 and the cell cycle controls myeloid differentiation. *Science*, 341(6146):670–673. 109

Kungulovski, G., Kycia, I., Tamas, R., Jurkowska, R. Z., Kudithipudi, S., Henry, C., Reinhardt, R., Labhart, P., and Jeltsch, A. (2014). Application of histone modification-specific interaction domains as an alternative to antibodies. *Genome Research*, 24:1842–1853. 135

Lara-Astiasio, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D. A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N., and Amit, I. (2014). Chromatin state dynamics during blood formation. *Science*, 345(6199):943–949. 26

Larson, D. R. (2011). What do expression dynamics tell us about the mechanism of transcription? *Current Opinion in Genetics & Development*, 21(5):591–599. 66, 81

Lawrence, N. D. (2004). Gaussian process latent variable models for data visualization of high dimensional data. *Advances in Neuronal Information Processing Systems*, 16:329–336. 57

Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783–1816. 57

Lee, J., Nemati, S., Silva, I., Edwards, B. A., Butler, J. P., and Malhotra, A. (2012). Transfer entropy estimation and directional coupling change detection in biomedical time series. *BioMedical Engineering OnLine*, 11:19. 108, 110, 111

Levsky, J. M. and Singer, R. H. (2003). Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science*, 116:2833–2838. 23

Liang, J. and Qian, H. (2010). Computational cellular dynamics based on the chemical master equation: a challenge for understanding complexity. *Journal of Computer Science and Technology*, 25:154–168. 30

Licatalosi, D. D. and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics*, 11:75–87. 11

Liew, C. W., Rand, K. D., Simpson, R. J. Y., Yung, W. W., Mansfield, R. E., Crossley, M., Proetorius-Ibba, M., Nerlov, C., Poulsen, F. M., and Mackay, J. P. (2006). Molecular analysis of the interaction between the hematopoietic master transcription factors Gata-1 and PU.1. *The Journal of Biological Chemistry*, 281:28296–28306. 109

Liviak, K. J., Wills, Q. F., Tipping, A. J., Datta, K., Mittal, R., Goldson, A. J., Sexton, D. W., and Holmes, C. C. (2013). Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods*, 59:71–79. 70

Lombard-Banek, C., Moody, S. A., and Nemes, P. (2016). Single-cell mass spectrometry for discovery proteomics: quantifying translational cell heterogeneity in the 16-cell frog (xenopus) embryo. *Angewandte Chemie*, 55(7):2454–2458. 86, 136

Lucchesi, J. C. and Kuroda, M. I. (2015). Dosage compensation in *Drosophila*. *Cold Spring Harbor Perspectives in Biology*, 7(5):a019398. 87

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A. R., Kamitak, N., Martersteck, E. M., Trombetta, J. J., Weitz, D. A., Sames, J. R., Shalek, A. K., Regev, A., and McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, pages 1202–1214. 137

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297. 57

Makowski, A. M., Dutnall, R. N., and Annunziato, A. T. (2001). Effects of acetylation of histone H4 at lysines 8 and 16 on activity of the Hat1 histone acetyltransferase. *The Journal of Biological Chemistry*, 47:43499–43502. 87, 100

Marbach, D., Costello, J. C., Küeffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Consortium, T. D., Kellis, M., Collins, J. J., and Stolovitzky, G. (2012). Wisdom of crowds for robust gene network inference. *Nature Methods*, 9:796–804. 107

Margueron, R. and Reinberg, D. (2010). Chromatin structure and the inheritance of epigenetic information. *Nature Reviews Genetics*, 11:285–296. 136

Matlin, A. J., Clark, F., and Smith, C. W. J. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6:386–398. 2

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York, NY, first edition. 57

Meyer, C. A. and Liu, X. S. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature Reviews Genetics*, 15:709–721. 21

Milo, R. and Phillips, R. (2015). *Cell Biology by the numbers*. Garland Science, online draft from book.bionumbers.org. 2

Miltenburger, H. G., Sachse, G., and Schliermann, M. (1987). S-phase cell detection with a monoclonal antibody. *Developments in Biological Standardization*, 66:91–99. 121, 126

Mir, M., Wang, Z., Shen, Z., Bednarz, M., Bashir, R., Golding, I., Prasanth, S. G., and Popescu, G. (2011). Optical measurement of cycle-dependent cell growth. *Proceedings of the National Academy of Sciences of the United States of America*, 108(32):13124–13129. 70

Mitchison, J. M. (2003). Growth during the cell cycle. *International Review of Cytology*, 226:165–258. 67, 71

Mitchison, J. M. (2005). Single cell studies of the cell cycle and some models. *Theoretical biology & medical modelling*, 2:4. 71

Moignard, V., Macaulay, I. C., Swiers, G., Buettner, F., Schütte, J., Calero-Nieto, F. J., Kinston, S., Joshi, A., Hannah, R., Theis, F. J., Jacobsen, S. E., de Bruijn, M. F., and Göttgens, B. (2013). Characterization of transcriptional networks in blood stem and progenitor cells using high-throughput single-cell gene expression analysis. *Nature Cell Biology*, 15(4):363–372. 70, 78, 79, 83

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., Nishikawa, I., Piterman, N., Kouskoff,

V., Theis, F. J., Fisher, J., and Göttgens, B. (2015). Decoding the regulatory network of early blood development from single-cell expression measurements. *Nature Biotechnology*, 33:269–276. 107

Mukhopadhyay, S. and Sengupta, A. M. (2013). The role of multiple marks in epigenetic silencing and the emergence of a stable bivalent chromatin state. *PLoS Computational Biology*, 9(7):e1003121. 88

Munsky, B. and Khammsh, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124:044104. 34

Munsky, B., Neuert, G., and van Oudenaarden, A. (2012). Using gene expression noise to understand gene regulation. *Science*, 336(6078):183–187. 80

Murphy, K. P. (2012). *Machine Learning A Probabilistic Perspective*. The MIT Press, Cambridge, Massachussetts. 50, 55, 56, 57, 59, 62, 123

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 56:3–48. 51

Nirenberg, M. W. and Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 47(10):1588–1602. 15

Nishimura, S., Takahashi, S., Kuroha, T., Suwabe, N., Nagasawa, T., Trainor, C., and Yamamoto, M. (2005). A GATA box in the GATA-1 gene hematopoietic enhancer is a critical element in the network of GATA factors and sites that regulate this gene. *Molecular and Cellular Biology*, 20(2):713–723. 109

Nurse, P. (2000). A long twentieth century of the cell cycle and beyond. *Cell*, 100(1):71–78. 27

Ocone, A., Haghverdi, L., Mueller, N. S., and Theis, F. J. (2015). Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data. *Bioinformatics*, 31(12):i89–i96. 107, 117

O'Connor, C. M. and Adams, J. U. (2010). *Essentials of Cell Biology*. NPG Education, Cambridge, MA. 14, 16, 22, 28, 134

Oedit, A., Vulto, P., Ramautar, R., Lindenburg, P. W., and Hankemeier, T. (2015). Lab-on-a-chip hyphenation with mass spectrometry: strategies for bioanalytical applications. *Current Opinion in Biotechnology*, 31:79–85. 136

Okuno, Y., Huang, G., Rosenbauer, F., Evans, E. K., Radomska, H. S., Iwasaki, H., Akashi, K., Moreau-Gachelin, F., Li, Y., Zhang, P., Göttgens, B., and Tenen, D. (2005). Potential autoregulation of transcription factor PU.1 by an upstream regulatory element. *Molecular and Cellular Biology*, 25(7):2832–2845. 109

Orkin, S. H. and Zon, L. I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. *Cell*, 132(4):631–644. 26, 109

Ozsolak, F. and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12:87–98. 23

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. *Nature*, 461(7265):814–818. 69

Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., and Raj, A. (2015). Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. *Molecular Cell*, 58(2):339–352. 67, 69

Parthun, M. R. (2012). Histone acetyltransferase1: more than just an enzyme? *Biochimica et biophysica acta*, 1819:256–263. 87

Parthun, M. R., Widom, J., and Gottschling, D. E. (1996). The major cytoplasmic histone acetyltransferase in yeast: links to chromatin replication and histone metabolism. *Cell*, 87(1):85–94. 87

Patterson, J. O., Swaffer, M., and Filby, A. (2015). An imaging flow cytometry-based approach to analyse the fission yeast cell cycle in fixed cells. *Methods*, 82:74–84. 131

Peccoud, J. and Ycart, B. (1995). Markovian modeling of gene-product synthesis. *Theoretical population biology*, 48(2):222–234. 41, 67

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., and Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews Genetics*, 14:288–295. 2

Perlman, Z. E., Slack, M. D., Feng, Y., Mitchison, T. J., Wu, L. F., and Altschuler, S. J. (2004). Multidimensional drug profiling by automated microscopy. *Science*, 306:1194–1198. 121

Pertea, M. and Salzberg, S. L. (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biology*, 11:206. 2

Petsko, G. A. and Ringe, D. (2004). *Protein Structure and Function*. Sinauer Associates, Sunderland, MA. 11

Phillips, T. and Shaw, K. (2008). Chromatin remodeling in eukaryotes. *Nature Education*, 1(1):209. 11

Plongthongkum, N., Dinh, H. D., and Zhang, K. (2014). Advances in the profiling of DNA modifications: cytosine methylation and beyond. *Nature Reviews Genetics*, 15:647–661. 21

Popescu, G., Park, K., Mir, M., and Bashir, R. (2014). New technologies for measuring single cell mass. *Lab on a Chip*, 14:646–652. 82

Portela, A. and Esteller, M. (2010). Epigenetic modifications and human disease. *Nature Biotechnology*, 28:1057–1068. 15

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):e309. 41, 43, 67, 81, 82

Raj, A. and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216 – 226. 67, 82

Raj, A. and van Oudenaarden, A. (2009). Single-molecule approaches to stochastic gene expression. *Annual Review of Biophysics*, 38:250–270. 41, 69

Rajaram, S., Pavie, B., and J., A. S. (2012). PhenoRipper: software for rapidly profiling microscopy images. *Nature Methods*, 9:635–637. 121

Ramsköld, D., Luo, S., Wang, Y. C., Li, R., Deng, Q., Faridani, O. R., Daniels, G. A., Khrebtukova, I., Laurent, L. C., Schroth, G. P., and Sandberg, R. (2012). Full-length mRNA-seq from single.cell levels of RNA and individual circulating tumor cells. *Nature Biotechnology*, 30:777–782. 67

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929. 51, 52

Rhodes, J., Hagen, A., Hsu, K., Deng, M., Liu, T. X., Look, A. T., and Kanki, J. P. (2005). Interplay of pu.1 and gata1 determines myelo-erythroid progenitor cell fate in zebrafish. *Developmental Cell*, 8:97–108. 109

Richman, R., Chicoine, L. G., Collini, M. P., Cook, R. G., and Allis, C. D. (1988). Micronuclei and the cytoplasm of growing Tetrahymena contain a histone acetylase activity which is highly specific for free histone H4. *The Journal of Cell Biology*, 106(4):1017–1026. 87

Rieger, M. A. and Schroeder, T. (2012). Hematopoiesis. *Cold Spring Harbor Perspectives in Biology*, 4(12):a008250. 26, 28

Roeder, I. and Glauche, I. (2006). 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):852–865. 109

Rubakhin, S. S., Romanova, E. V., Nemes, P., and Sweedler, J. V. (2011). Profiling metabolites and peptides in single cells. *Nature Methods*, 8:S20–S29. 3

Sakaue-Sawano, A., Kurokawa, H., Morimura, T., Hanyu, A., Hama, H., Osawa, H., Kashiwagi, S., Fukami, K., Miyata, T., Miyoshi, H., Imamura, T., Ogawa, M., Masai, H., and Miyawaki, A. (2008). Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, 132:487–498. 120

Sandberg, R. (2014). Entering the era of single-cell transcriptomics in biology and medicine. *Nature Methods*, 11:22–24. 55, 135

Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K., and Petzold, L. R. (2011). StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458. 39

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K. D., Imai, T., and Ueda, H. U. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biology*, 14(4):R31. 70

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5:197–227. 62

Schoenberg, D. R. and Maquat, L. E. (2012). Regulation of cytoplasmic mRNA decay. *Nature Reviews Genetics*, 13:246–259. 11

Schreiber, T. (2000). Measuring Information Transfer. *Physical Review Letters*, 85:461–464. 53, 55, 108

Schroeder, D. V. (2000). *An Introduction to Thermal Physics*. Addison Wesley Longman, San Francisco, CA. 44

Schroeder, T. (2011). Long-term single-cell imaging of mammalian stem cells. *Nature Methods*, 8:S30–S35. 109

Schwannhäusser, B., Busse, D., Dittmar, G., Schuchhardt, J., Wolf, J., W., C., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473:337–342. 18, 39, 41

Sedighi, M. and Sengupta, A. M. (2007). Epigenetic chromatin silencing: bistability and front propagation. *Physical Biology*, 4(4):246–25. 88

Seiffert, C., Khoshgoftaar, T. M., van Hulse, J., and Napolitano, A. (2010). RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man and Cybernetics, Part A, Systems and Humans*, 40:185–197. 124, 128

Seto, E. and Yoshida, M. (2014). Erasers of histone acetylation: the histone deacetylase enzymes. *Cold Spring Harbor Perspectives in Biology*, 6(4):a018713. 89

Shahbazian, M. D. and Grunstein, M. (2007). Functions of site-specific histone acetylation and deacetylation. *Annual Review of Biochemistry*, 76:75–100. 87

Shahrezaei, V. and Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45):17256–17261. 41, 43, 44, 66, 81, 82

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423,623–656. 54

SHOGoiN (2016). *SHOGoiN* human cell taxonomy. `http://shogoin.stemcellinformatics.org/cell`. Accessed: 2016-01-19. 1

Signolet, J. and Hendrich, B. (2015). The function of chromatin modifiers in lineage commitment and cell fate specification. *The FEBS Journal*, 282(9):1692–1702. 27

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London. 111

Smith, E. and Shilatifard, A. (2010). The chromatin signaling pathway: diverse mechanisms of recruitment of histone-modifying enzymes and varied biological outcomes. *Molecular Cell*, 40(5):689–701. 87

Sneppen, K. and Dodd, I. B. (2012). A simple histone code opens many paths to epigenetics. *PLoS Computational Biology*, 8(8):e1002643. 88

Sobel, R. E., Cook, R. G., and Allis, C. D. (1994). Nonrandom acetylation of histone H4 by a cytoplasmic histone acetyltransferase as determined by novel methodology. *The Journal of Biological Chemistry*, 269(28):18576–18582. 87

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluation systematic relationships. *University of Kansas Scientific Bulletin*, 28:1409–1438. 58

Ståhlberg, A. and Martin, B. (2010). Single-cell gene expression profiling using reverse transcription quantitative real-time PCR. *Methods*, 50:282–288. 66

Stephens, D. J. and Allan, V. J. (2003). Light microscopy techniques for live cell imaging. *Science*, 300(5616):82–86. 24

Strahl, B. D. and Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, 403:41–45. 87

Strasser, M., Theis, F. J., and Marr, C. (2011). Stability and multiattractor dynamics of a toggle switch based on a two-stage model of stochastic gene expression. *Current Opinion in Genetics & Development*, 21:147–153. 109

Straub, T. and Becker, P. B. (2011). Transcription modulation chromosome-wide: universal features and principles of dosage compensation in worms and flies. *Current Opinion in Genetics & Development*, 21:147–153. 87

Suganuma, T. and Workman, J. L. (2011). Signals and combinatorial functions of histone modifications. *Annual Review of Biochemistry*, 80:473–499. 87

Tan, M., Luo, H., Lee, S., Jin, F., Yang, J. S., Montellier, E., Buchou, T., Cheng, Z., Rousseaux, S., Rajagopal, N., Lu, Z., Ye, Z., Zhu, Q., Wysocka, Q., Ye, Y., Khochbin, S., Ren, B., and Zhao, Y. (2011). Identification of 67 histone marks and histone lysine crotonylation as a new type of histone modification. *Cell*, 146(6):1016–1028. 87

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K., and Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6(5):377–382. 66

Tang, F., Lao, K., and Surani, M. A. (2011). Development and applications of single-cell transcriptome analysis. *Nature Methods*, 8(4s):S6–S11. 3

Thattei, M. and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8614–8619. 67, 82

The-UniProt-Consortium (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212. 2

The-UniProt-Consortium (2016). *UniProtKB* filtered for homo sapiens. `http://www.uniprot.org/uniprot/`. Accessed: 2016-01-19. 2

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622. 57

Toyama, B. H., Savas, J. N., Park, S. K., Harris, M. S., Ingolia, N. T., Yates, J. R., and Hetzer, M. W. (2013). Identification of long-lived proteins reveals exceptional stability of essential cellular structures. *Cell*, 154(5):971–982. 91

Trapnell, C., Cacchiarelli, D., Grimbsy, J., Pokharei, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32:381–386. 117

Turner, B. M. (2000). Histone acetylation and an epigenetic code. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 22(9):836–845. 87

Vallejos, C. A., Marioni, J. C., and Richardson, S. (2015). BASiCS: Bayesian analysis of single-cell sequencing data. *PLoS Computational Biology*, 24:e1004333. 67, 70

Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 1:1–48. 57

Van Kampen, N. G. (1997). *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam, The Netherlands, second edition. 34

VanGulder, H. D., Vrana, K. E., and Freeman, W. M. (2008). Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*, 44(5):619–626. 23

Venzon, D. J. and Moolgavkar, S. H. (1988). A method for computing profile-likelihood based confidence intervals. *Applied Statistics*, 37(1):87–94. 51

Vermeulen, K., Van Bockstaele, D. R., and Berneman, Z. N. (2003). The cell cycle: a review of regulation, deregulation and therapeutic targets in cancer. *Cell Proliferation*, 36(3):131–149. 27

Vogelstein, B. and Kinzler, K. W. (1999). Digital PCR. *Proceedings of the National Academy of Sciences of the United States of America*, 96(16):9236–9241. 69

Voss, T. C. and Hager, G. L. (2014). Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nature Reviews Genetics*, 15:69–81. 16

Walther, T. C. and Mann, M. (2010). Mass spectrometry-based proteomics in cell biology. *The Journal of Cell Biology*, 190(4):491–500. 23

Wang, D. and Steven, B. (2010). Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology*, 28(6):281–290. 66

Warren, L., Bryder, D., Weissman, I. L., and Quake, S. R. (2006). Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. *Proceedings of the National Academy of Sciences of the United States of America*, 103(47):17807–17812. 78

Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids. A structure for deoxyribose nucleic acids. *Nature*, 171:737–738. 12

Watson, J. V., Chambers, S. H., and Smith, P. J. (1987). A pragmatic approach to the analysis of DNA histograms with a definable G1 peak. *Cytometry*, 8:1–8. 126

Whichard, Z. L., Sarkar, C. A., Kimmel, M., and Corey, S. J. (2010). Hematopoiesis and its disorders: a systems biology approach. *Blood*, 115:12. 26

Wickramasinghe, V. O. and Laskey, R. A. (2015). Control of mammalian gene expression by selective mRNA export. *Nature Reviews Molecular Biology*, 16:431–442. 11

Wikipedia (2016). Genetischer code. `https://de.wikipedia.org/wiki/Genetischer_Code`. Accessed: 2016-01-19. 14

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62. 53

Wojcik, K. and Dobrucki, J. W. (2008). Interaction of a DNA intercalator DRAQ5, and a minor groove binder SYTO17, with chromatin in live cells – influence on chromatin organization and histone-DNA interactions. *Cytometry A*, 73:555–562. 121

Xia, X. and Wong, S. T. (2012). Concise review: a high-content screening approach to stem cell research and drug discovery. *Stem Cells*, 30:1800–1807. 132

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., Huang, J., Li, M., Wu, X., Wen, L., Lao, K., Li, R., Qiao, J., and Tang, F. (2013). Single-cell

RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural & Molecular Biology*, 20:1131–1139. 66

Yang, X.-J. and Seto, E. (2007). HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene*, 26:5310–5318. 103

Yuan, C. M., Douglas-Nikitin, V. K., Ahrens, K. P., Luchetta, G. R., Braylan, R. C., and Yang, L. (2004). DRAQ5-based DNA content analysis of hematolymphoid cell subpopulations discriminated by surface antigens and light scatter properties. *Cytometry Part B: Clinical Cytometry*, 58:47–52. 130

Zerihun, M. B., Vaillant, C., and Jost, D. (2015). Effect of replication on epigenetic memory and consequences on gene transcription. *Physical Biology*, 12(2):026007. 88

Zhang, K., Williams, K. E., Huang, L., Yau, P., Siino, J. S., Bradbury, E. M., Jones, P. R., Minch, M. J., and Burlingame, A. L. (2002). Histone acetylation and deacetylation: identification of acetylation and methylation sites of HeLa histone H4 by mass spectrometry. *Molecular & Cellular Proteomics*, 1:500–508. 103

Zheng, Y., Sweet, S. M. M., Popovic, R., Martinez-Garcia, E., Tipton, J. D., Thomas, P. M., Licht, J. D., and Kelleher, N. L. (2012). Total kinetic analysis reveals how combinatorial methylation patterns are established on lysine 27 and 36 of histone H3. *Proceedings of the National Academy of Sciences of the United States of America*, 109(34):13549–13554. 88

Zopf, C. J., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Computational Biology*, 9(7):e1003161. 82

Zuleta, I. A., Aranda-Diaz, A., Li, H., and El-Samad, H. (2014). Dynamic characterization of growth and gene expression using high-throughput automated flow cytometry. *Nature Methods*, 11:443–448. 132

[I] Laozi. *Tao Te Ching.* Chapter 64. Cited after Ostasieninstiut der Hochschule Ludwigshafen am Rhein: `http://www.oai.de/en/45-publikationen/sprichwort/818-eine-reise-von-tausend-meilen-beginnt-mit-dem-ersten-schritt.html`. Accessed: 2016-01-19.

[II] Goethe, J. W. von. *Chinesisch-Deutsche Jahres- und Tageszeiten.* Chapter 10. Cited after Projekt Gutenberg-DE: `http://gutenberg.spiegel.de/buch/johann-wolfgang-goethe-gedichte-3670/421`. Accessed: 2016-01-19.

[III] Box, G. E. P., and Draper, N. R. (1987). *Empirical Model Building and Response Surfaces.* John Wiley & Sons, New York, NY. p. 424. Cited after wikiquotes.org: `https://en.wikiquote.org/wiki/George_E._P._Box`. Accessed: 2016-01-19.

[IV] Ockham, W. of. *Summa Totius Logicae.* Book I, Chapter 12. Cited after wikiqoutes.org: `https://en.wikiquote.org/wiki/William_of_Ockham`. Accessed: 2016-01-19.

[V] Ovid, P. O. *Metamorphoses*. Book IV, 287. Cited after wikiquotes.org: `https://en.wikiquote.org/wiki/Ovid`. Accessed: 2016-01-19.

[VI] Feuerbach, A. *Ein Vermächtnis von Anselm Feuerbach*. Chapter 36. Cited after Projekt Gutenberg-DE: `http://gutenberg.spiegel.de/buch/ein-vermachtnis-von-anselm-feuerbach-4462/36`. Accessed: 2016-01-19.

[VII] Thoreau, H. D. *Men Say They Know Many Things*. Cited after poetryfoundation.org: `http://www.poetryfoundation.org/poem/175541`. Accessed: 2016-01-19.