Experimental and theoretical investigation of estrogen dependent transcription in single cells

Dissertation

zur Erlangung des Grades

„Doktor

der Naturwissenschaften“

am Fachbereich Physik, Mathematik und Informatik

der Johannes Gutenberg-Universität

in Mainz

Stephan Baumgärtner

geb. in Karl-Marx-Stadt
Mainz, den 11.07.2016

Summary

Estrogen is the primary female sex hormone and plays an important role in the regulation of the female reproductive system but also in breast cancer growth. Estrogen diffuses through the cell membrane, binds to intracellular estrogen receptors (ER) which in turn control the activity of estrogen sensitive genes. On the molecular level estrogen signalling is well characterised and serves as a model system for the regulation of transcription of the genomic code into messenger RNA. Biochemical measurements averaging over millions of cells suggested that the control of ER-dependent genes is highly regular and deterministic. Transcription at the single cell level is, however, generally thought to be a stochastic process that occurs in random bursts. In this work, we addressed whether estrogen-dependent transcription is a stochastic process at the single-cell level and how it is modulated by changing estrogen concentrations.

We used a time lapse imaging approach to capture the dynamics of estrogen induced transcription in MCF-7 breast cancer cells. A short reporter construct was integrated into the estrogen sensitive GREB1 gene. Upon transcription into RNA this construct folds into loops that are recognised by a fluorescently labelled protein which allows for visualising transcription under the microscope. We found that GREB1 transcription is stochastic, suggesting that the GREB1 gene randomly switches between phases of transcriptional activity and inactivity in a burst-like manner. We asked how increasing estrogen concentrations lead to higher transcriptional activity, and hypothesised that the stimulus may increase the frequency of the bursts, their duration or their intensity.

To test the formulated hypotheses we fitted stochastic models of varying complexity to transcription time course data recorded at different estrogen concentrations using Sequential Monte Carlo Approximate Bayesian Computation. We calibrated and validated our modelling approach using benchmark problems. For the real experimental data, we found that model selection favours small models of only two states in which the gene stochastically switches between active and inactive transcription ('random telegraph model'). More complex models consisting of more than two states were neglected by model selection, indicating that two rate-limiting steps dominate the many molecular steps that had been described for estrogen-dependent gene activation. By fitting, we showed that estrogen modulates transcription in a dose-dependent manner by modulating the frequency of the stochastic bursts. This suggests that the estrogen stimulus mainly controls the time it takes to reactivate the gene, and to a lesser extent its phases of activity. The corresponding model reflected that a fraction of the cell population does not respond to stimulation at low estrogen concentrations during the observation period ('non-responding subpopulation'), thereby providing insights into cellular heterogeneity.

Cells need to quickly respond to changing environmental conditions by tuning the transcription levels of important genes. To better understand the kinetics of estrogen-dependent gene induction, we developed an integrated model that combines the stochastic transcription model with a quantitative description of the estrogen signalling pathway. This integrated model successfully predicted response times of the GREB1 gene in synchronised cell populations at different estrogen concentrations. We concluded that gene induction by estrogen is slow compared to other hormones, with signalling and stochastic gene switching being the rate-limiting step at low and high estrogen doses, respectively.

Our results provide an integrated mathematical description of estrogen signalling and transcription that is capable to explain experimental data of transcription over a wide range of estrogen concentrations. This work may contribute to a better understanding of heterogeneous growth of estrogen-dependent breast cancers.

Contents

List of Figures

List of Tables

1 Introduction

Gene transcription is a core biological process. Information stored in the DNA is transcribed into RNA which itself serves as a template to build a protein. This flow of information from DNA to RNA to protein often is referred to as the *central dogma of molecular biology*. Proteins as life's building blocks and machines are necessary in every cell but only in certain amounts and at specific times depending on the internal state of the cell and the cellular environment. The process of gene transcription therefore has to be tightly controlled to ensure reliability.

Transcriptional regulation is based upon biochemical interactions between certain proteins and DNA. In human cells genes are only present in two copies or alleles that have to be reached by external signals. In addition, copy numbers of some of the proteins responsible for transcriptional regulation can be low and thus interactions occur at stochastic time points. Therefore, transcription itself is a stochastic process [Thattai and van Oudenaarden, 2001; Raj et al., 2006]. This stochasticity leads to unreliable gene transcription, i.e. noise in the expression levels of proteins. This noise can be amplified when noise afflicted proteins are responsible for the regulation of the expression of other proteins. Thus, genetically identical cells growing under the same environmental conditions can exhibit very different behaviours [Geva-Zatorsky et al., 2006; Cohen et al., 2008; Rausenberger and Kollmann, 2008; Hilfinger and Paulsson, 2011; Jeschke et al., 2013].

To create reliable responses to environmental changes a cell has to cope with uncertainty as one of its most important processes is inevitably stochastic. Mechanistic understanding of such processes provides a deeper understanding of the functioning of a cell. Moreover, it would provide means to predict the behaviour of cells under changing conditions. For example it would allow to predict the result of drug treatments. Therefore, the first question that this work is concerned with is:

Which mechanisms can explain the stochastic nature of gene transcription?

Understanding the mechanism of gene transcription is only the first step. Cells respond to external cues by changing the expression level of proteins, i.e. the abundance of molecules in the cell. Such external cues could be a signalling molecule like a hormone or the availability of nutrients. The cells response often has to be rapid and more importantly it has to be reliable. The next step will be to understand how this mechanism can explain the cells response to varying environmental conditions, i.e. what parameters of the process change upon changes in the environment. This is the second question of this work:

How does stochastic gene transcription change upon changing environmental conditions?

This work applies both stochastic and deterministic approaches to mathematically describe estrogen induced transcription of the GREB1 gene in MCF7 breast cancer cells. The concentration of the hormone estrogen here is used as environmental stimulus. Estrogen has effects on the transcriptional activity of many genes and this work focusses on the GREB1 gene as an important factor in the growth of estrogen dependent breast cancer [Rae et al., 2005; Laviolette et al., 2014]. The estrogen concentration provides an input that modulates transcriptional activity of GREB1.

Nascent transcripts of GREB1 are fluorescently labelled allowing to study transcription in live cells over time by light microscopy. Nascent transcripts are RNA molecules under production. Tracking this production enables to investigate the stochastic dynamics of transcription and the effects of external stimuli on transcription directly.

The theory of Markov Processes provides a powerful mathematical tool to model stochastic processes and hence develop a mechanistic understanding of the process of transcription and its regulation [Ko, 1991; Peccoud and Ycart, 1995; Kepler and Elston, 2001; Paulsson, 2005]. We applied model calibration and selection to find the most likely model and its corresponding parameters. Calibration of stochastic models is a nontrivial task and requires fitting multi dimensional probability distributions. We used a likelihood free simulation based Bayesian approach to perform this task.

To connect the stochastic model of transcription with the environment it was coupled with a deterministic model of the estrogen signalling pathway. A cell detects signals like hormones from its environment by various receptor molecules which transmit information to the cellular interior. Modelling of cellular signalling by means of ordinary differential equations has a fruitful history [Kholodenko, 2006]. We developed a two level model of ordinary differential equations to include the nonlinear dynamics of estrogen signalling into the model transcription.

The next sections will give a brief overview of the current state of knowledge towards the stated questions. The key molecules and processes of gene transcription will be introduced. Subsequently the biological model system that was used for this study will be described followed by an review of different modelling approaches.

1.1 Biology of gene transcription

1.1.1 Gene transcription

Genomic information in the cell is stored by the sequence of the bases adenine, cytosine, thymine and guanine (ACTG) in the DNA. This information is read and transcribed into RNAs which in turn are translated into proteins. A gene is a section of the DNA sequence that encodes a protein. The key protein of reading and transcribing DNA into RNA is the enzyme RNA polymerase. In eukaryotes three different types of RNA polymerase with distinct functions exist. Interest of this work lies in the regulation of transcription of protein encoding genes which are transcribed by RNA polymerase II [Alberts et al., 2014].

The stages of transcription are usually described as initiation, elongation and termination. Initiation defines the process when RNA polymerase located at the transcriptional start site of the gene starts its work transcribing the gene. After transcription is initiated RNA polymerase is elongating the transcript by sequentially adding bases complementary to the DNA template. When transcription of the whole RNA is complete the process terminates. RNA polymerase falls of the DNA and the transcript is released into the cytoplasm where it is further processed and translated into protein [Munk, 2010]. Regulation of transcription by external cues mainly occurs at the level of initiation [Alberts et al., 2014]. Therefore, this work concentrates on the process of transcriptional initiation which is directly connected to the activity of the gene, i.e. being active or not.

Gene activity is controlled by so called transcription factors. DNA not only stores information for the cell on how to build proteins but carries many so called regulatory elements. These sequence elements function as templates for the binding of regulatory factors or transcription factors. Each protein encoding gene is preceded by a promoter region that can carry multiple different binding elements for transcription factors. Binding of a transcription factor can have different effects. If the transcription factor blocks recruitment of RNA polymerase it prevents transcription. In the contrary case when the transcription factor promotes recruitment of RNA polymerase it promotes transcription. Thus, to transcribe a gene usually multiple transcription factors and other coproteins work together to recruit RNA polymerase [Alberts et al., 2014].

1.1.2 Chromatin & Epigenetics

As pointed out in the previous section availability of RNA polymerase is essential but not sufficient for gene transcription. Before RNA polymerase can transcribe a gene the gene itself has to be in an active or transcriptional permissive state. What characterises an active or inactive state of a gene is defined by its chromatin environment. Chromatin is the packed form of DNA in the cell nucleus of eukaryotic cells. This packaging provides an environment that generally is transcriptionally not permissive. To initiate transcription DNA has to be actively unpacked to make it accessible to RNA polymerase. Transcription factors directly bind to chromatin and can induce unpacking and thus bring chromatin in an transcriptionally permissive state.

The first and here most important layer of packaging is wrapping of the DNA around histone complexes forming so called nucleosomes. Histones are octameres of the four histone core proteins H2A, H2B, H3 and H4. All core histones possess high fractions of the alkaline amino acids arginine and lysine (together more than 20%) [Watson et al., 2007]. Thus, their positive charges are neutralised by the negative charges of the DNA backbone which ensures tight packaging. The alkaline nature of the histones can be neutralised by adding acetyl groups to lysines in the histone tails which in turn relaxes DNA packaging and thus promotes transcription. This process of histone acetylation is realised by enzymes called histone acetyl transferases (HATs). The reverse process is promoted by histone deacetylases (HDACs) that remove the actyl groups and therefore cause tighter DNA packaging which in turn shuts down transcription [Stasevich et al., 2014; Kirmes et al., 2015]. A second important histone modification is the addition of methyl groups to lysine residues at the N-termini. Histone methylation often is related to transcriptional inactivity [Greer and Shi, 2012]. Methyl groups in contrast to acetyl groups cause a tighter packaging of the chromatin. Placing of methyl groups is promoted by another group of enzymes the histone methyl transferases (HMTs) [Li et al., 2007]. In general, changes on the chromatin such as addition or removal of acetyl or methyl groups have an impact on gene transcription and are referred to as epigenetics.

Acetylation and methylation of lysine residues in histone tails are counter acting processes that could be imagined as a switching operation between chromatin states that are transcriptionally permissive or prohibitive. In mathematical models these multi step processes often are simplified by assuming on and off states of transcriptional activity. Switching of a gene between such transcriptional active or inactive states can biologically be interpreted as a change in the chromatin making DNA more or less accessible to the transcriptional machinery.

1.2 Modelling of transcription as a stochastic process

Traditional experimental methods in molecular biology are biochemical approaches and are usually not able to study single cells but the average of large cell populations. First quantitative models focussed on the behaviour of cell populations and were deterministic in nature. The first adaptive model of gene regulation was the operon model of Jacob and Monod. This model describes how the availability of sugar as an external signal is translated into the transcription of genes that adjust the cells metabolism towards the energetically most favourable for the current conditions [Jacob and Monod, 1961].

Invention of fluorescence in-situ hybridisation allowed to measure the number of transcripts of specific genes in single cells. Fluorescently labelled RNA probes complement to the RNA transcripts of interest allow to count the number of transcripts per cell [Femino, 1998]. Distributions of RNA numbers between cells were found to show substantial variability [Ko et al., 1990] leading to the conclusion that gene transcription is a stochastic process and thus stochastic simulation techniques are necessary for a mechanistic understanding [Ko, 1991; McAdams and Arkin, 1997]. Formulation of continuous time markov models of transcription where a gene switches between states of active and inactive transcription provide a fruitful approach for a general understanding of the process of transcription [Paulsson, 2005].

The next sections will review results of modelling gene transcription by means of stochastic models. First the one state or always on model and its applications will be reviewed before turning to more complex models. Both types of models are very different in their qualities and both were successfully applied to explain stochastic gene transcription. It is, however, not obvious if either type of model is more general or whether applicability is highly gene specific.

1.2.1 The one state or always on model

The simplest stochastic model of gene transcription that can be imagined is that of a gene that is constantly on. This corresponds to the case where a transcription factor is permanently bound and the chromating is open. New transcripts are produced with a fixed rate k_m and decay with a rate proportional to the amount of existing RNA. This is described as a birth-death or poisson process where the number of transcripts per cell follows a poisson distribution. Thus, the mean number of RNA molecules per cell would be equal to the variance, i.e. the Fano-factor as ratio between the variance and mean is equal to one. The top left panel of figure 1.2.1 depicts the topology of such an always on model. The plots in the upper row of the figure show a simulated time course of RNA copy number starting with no transcripts (middle) and the distribution of RNA molecules per cell over a population of cells (right). The simulated time course reaches a steady state which is defined by the ratio of RNA production and decay rates. The steady state at the same time is the mean number of RNA molecules per cell and, thus, due to the poisson nature of the process the variance, too. So called house keeping genes that are required by the cell throughout the cell cycle and often encode for structural proteins or metabolic enzymes can be described by such a poissonian model but most other genes can not [Zenklusen et al., 2008].

1.2.2 The two state or random telegraph model

Single cell microscopy allowed to investigate transcription in single cells and it was observed that transcription often occurs in so called bursts for large variety of genes [McAdams and Arkin, 1997; Golding et al., 2005; Raj et al., 2006; Chubb et al., 2006]. That is, during a short period of time multiple transcripts are produced. In between bursts no transcription takes place. Transcriptional bursts lead to distributions of RNA molecules per cell that show a variance that can be much larger than the mean and hence the simple poisson process introduced above is not applicable. Such behaviour suggests that transcription follows a certain regularity of switching in transcriptional activity. When the gene is active transcription occurs as a poisson process with rate k_m. When switched to an inactive state no transcription occurs. This switching corresponds to epigentic changes as described earlier. RNA decay is proportional to the amount of RNA present as for the always on model. The middle row of figure 1.2.1 displays simulation results of such a random telegraph model. Model topology is depicted on the left. The second panel shows a simulated time course of promoter states (orange) and RNA copy number (blue). When the gene is on the number of mRNA molecules increases rapidly exhibiting typical bursts. Such a switching process coupled with a poisson process of transcription is able to generate a highly fluctuating number of transcripts over time [Paulsson, 2005]. The third panel shows the distribution of the number of RNA molecules over a population of cells. Its variance is higher than for the always on model. Adjusting transition rates between active and inactive states [Munsky et al., 2012] or transcription initiation rate in the active state [Molina et al., 2013] such a system is able to react to external signals and thus modulating the distribution of RNA molecules per cell. Both modes of transcription always on and random telegraph were observed for different genes in yeast [Zenklusen et al., 2008]. The right panel shows the distribution of the genes off times. The waiting times of a markovian system to stay in one state are exponentially distributed. This leads to a substantial amount of variation in the time courses. More regularity of the switching process can be achieved by introducing more states to create a cycle. Such models will be discussed in the next section.

1.2.3 Cyclic models with more than two states

Complex promoters of eukaryotes carrying multiple binding sites for different transcription and cofactors suggest the existence of more than two states. For instance, first a transcription factor binds to the gene promoter, then epigenetic modifications take place and eventually RNA polymerase is recruited. When transcription is completed these processes are reverted. All these processes require multiple steps in total. Indeed, gene transcription compatible with a three state model was observed for multiple mammalian genes [Suter et al., 2011]. Moreover, Zoller et al found that depending on the promoter structure typically up to five internal states make up a so called promoter cycle [Zoller et al., 2015].

Introducing more than two states of transcriptional activity has consequences on the qualities of the model. An additional state could be a second inactive state that the gene has to pass before it can be activated again. Such an refractory phase causes the waiting

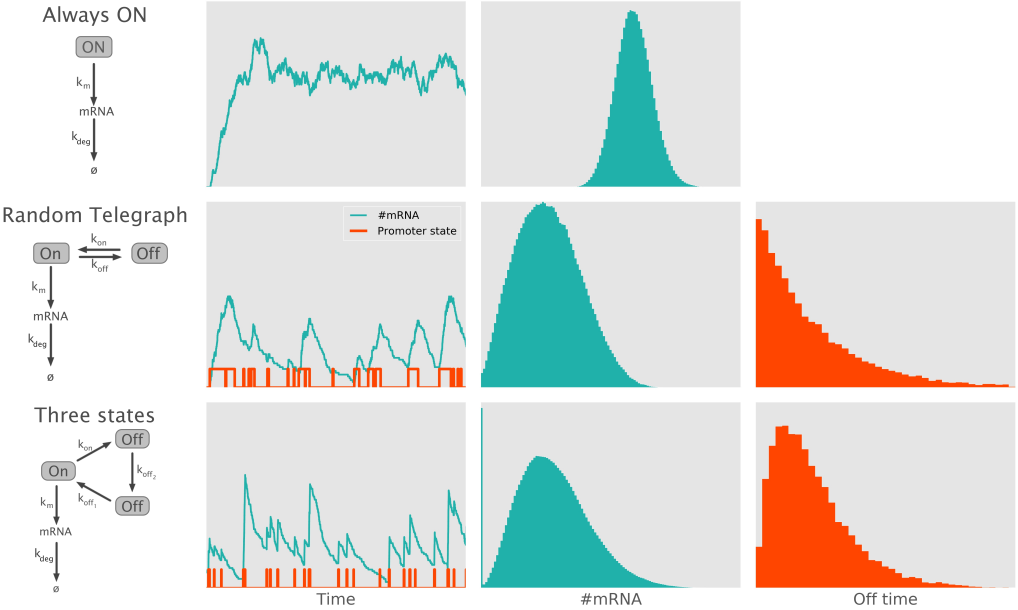

Figure 1.2.1: Qualitative differences in stochastic models with one, two and three internal states. The panels show simulations of transcription for three different models, the always on model (top), the two state or random telegraph model (middle) and a three state model (bottom). Model topologies are depicted on the left. Switching between active and inactive states occurs with rates k_{on} and k_{off} respectively. When the gene is in an active state the gene transcript is produced with initiation rate k_m and degraded with rate k_{deg}. The second column shows time course of the mRNA copy number (blue) and the switching of the gene between active and inactive states (orange). The third column shows the distribution of mRNA molecules per cell over a population of cells. The right column shows the distribution of the time the gene spends in the inactive or off phase.

times that the gene spends in the inactive phase to show a peaked distribution. Both internal off states have their own exponentially distributed waiting time but the total waiting time is the convolution of both individual distributions. This leads to a higher regularity in the switching between the active and inactive phases of the gene since the off time is more accurately defined and thus reducing variability. Introduction of even more states pronounces this effect. The bottom row of figure 1.2.1 shows an example simulation for a three state model. The appearance of gene on phases is more regular than for the two state model and the gene off times follow a peaked distribution.

1.2.4 Importance of heterogeneity in cellular decision processes

Stochastic transcription can lead to substantial variability between genetically identical cells [Elowitz et al., 2002] which was found to be an inherent and important effect. Recent studies revealed that this can have important functional consequences. For instance, among haematopoietic progenitor cells in mice the abundance of certain transcription factors can show substancial variation. So called outlier cells with either very high or very low transcription levels can follow either the erythroid or the myeloid lineage in cellular differentiation[Chang et al., 2008]. This example highlights the biological importance and neccessecity of cell to cell variability for cell fate decisions during development. Cellular heterogeneity on the contrary can be very problematic in case of cancer treatment when a fraction of cells does not respond to drug treatment [Cohen et al., 2008; Paek et al., 2016]. Here different responses to treatment are not desired. Both examples mark the need for an understanding of the processes that drive cellular heterogeneity.

1.3 The role of estrogen signalling for transcription

Estrogen is a steroid hormone and the most important female sex hormone primarily responsible for development and regulation of the female reproductive system. Three major forms of estrogen exist as estrone (E1), estradiol (E2) and estriol (E3). This work concentrates on the effects of $17\beta-$estradiol on the transcription of estrogen sensitive genes in MCF-7 cells. Estradiol or E2 is a small molecule that can diffuse across the cell membrane. E2 is a ligand to the cytoplasmic estrogen receptor α (ER_α, a protein) [Dahlman-Wright et al., 2006]. Ligand bound receptors form homodimers, translocate to the cell nucleus and directly bind to specific DNA binding sites, so called estrogen responsive elements (EREs) [Kumar and Chambon, 1988]. There they act as transcription factor recruiting chromatin remodelling factors and thus eventually promoting transcription. Stimulation of MCF-7 cells with estradiol leads to the expression of hundreds of genes some of which are important for cancer growth. GREB1 (short for: growth regulation by estrogen in breast cancer 1) is an example of a gene whose protein product can lead to cancer growth upon estrogen stimulation [Rae et al., 2005; Laviolette et al., 2014] and provides a possible target for breast cancer therapies [Hodgkinson and Vanderhyden, 2014]. A second important example is the trefoil factor 1 (TFF1 or ps2) gene [Markićević et al., 2014]. Certain types of breast cancer are estrogen dependent, i.e.

they need estrogen to grow. Such cancers can be treated by estrogen antagonists like tamoxifen to inhibit transcription of estrogen dependent genes [Lebedeva et al., 2012]. Thus, understanding of estrogen dependent transcription in breast cancer cells is essential to develop therapies.

1.3.1 MCF7 cells as a model system of hormone induced transcription

MCF-7 (Michigan Cancer Foundation-7) is a breast cancer cell line first cultivated in 1973 [Soule et al., 1973]. It is a well established model system for estrogen dependent cancer growth and therapy [Lacroix and Leclercq, 2004; Comşa et al., 2015]. The cells possess cytoplasmic estrogen receptors und show pronounced proliferation upon estrogen stimulation. An extensive body of literature exists on the temporal dynamics of changes of the chromatin environments of estrogen responsive genes in MCF-7 cells. Thus, MCF-7 cells provide a promising model system to study estrogen induced gene transcription.

1.3.2 Transcriptional cycling in synchronised cell populations

Previous studies of estrogen induced transcription and transcriptional initiation focussed on the ps2 and GREB1 genes on the population level. Experimentally populations of synchronised cells were studied. Synchronisation of the transcriptional activity of estrogen dependent genes was achieved by growing cells in estrogen free medium for three days leading to strongly reduced transcription. The advantage of this procedure is the strong transcriptional response upon estrogen stimulation. Temporal changes in the chromatin environment of the promoter region were studied by chromatin immunoprecipitation (ChIP). This biochemical method utilises specific antibodies against transcription factors or chromatin modifications to isolate DNA fragments bound by the factor of interest. Detection and quantification of the precipitated DNA fragments is done by quantitative PCR [Dahl and Collas, 2008].

A first study applying time course ChIP experiments in synchronised cells revealed cyclic engagement of the estrogen receptor ER_α and various other factors to the promoter of the ps2 gene [Shang et al., 2000]. Raphaël Métivier et al further investigated the binding of transcription factors, chromatin remodelling factors and histone marks to the promoter by the same method. They observed persistent cycles of binding and release

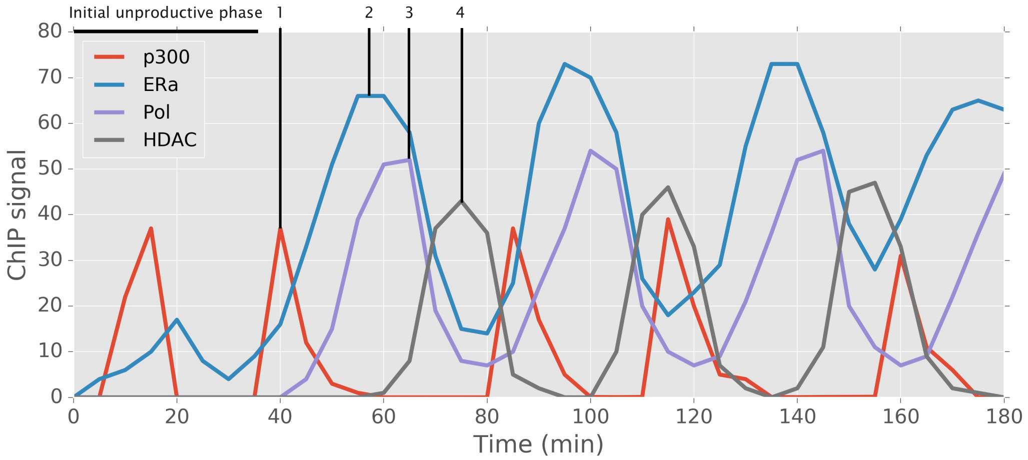

Figure 1.3.1: Transcription and chromatin remodelling factors show cyclic engagement with target promoters. Métivier et al observed cyclical occupancy of multiple factors to gene promoter regions in synchronised cell populations. Here four factors are shown. p300 is a histone acetyl transferase, ERa is the estrogen receptor α, Pol is RNA polymerase and HDAC is a histone deacetylase. The initial unproductive phase (marked by the vertical black line) is due to the treatment of the cells with the RNA inhibitor α-amanitin during synchronisation. Vertical black lines indicate the maxima of the different factors during the first oscillation. Data taken from [Métivier et al., 2003].

of the individual factors to the promoter with a period of approximately 40 minutes [Métivier et al., 2003, 2004, 2008]. In total 46 factors were investigated which all showed sustained oscillations. Figure 1.3.1 shows example time courses of four different factors, namely p300 a histone acetyl transferase, the estrogen receptor ER_α, RNA polymerase (Pol) and histone deacetylase (HDAC). The initial phase of approximately 35 minutes did not show oscillations and was due to the synchronisation procedure. The biological interpretation of the observed oscillations was that the factors assemble at the promoter in a cyclic and ratchet like process [Métivier et al., 2003, 2006; Carlberg, 2010]. This cycle consists of many steps and each step is characterised by either binding or release of one factor to or from the gene promoter. The observed sustained oscillations indicate a high degree of regularity in this process. The position of the peaks of the different oscillations mark the temporal order of events as indicated in figure 1.3.1. For the given examples this would mean first p300 acetylates histones in the promoter region to open the chromatin which is followed by binding of the estrogen receptor ER_α. ER_α binding in turn leads to the recruitment of the transcription machinery including RNA polymerase.

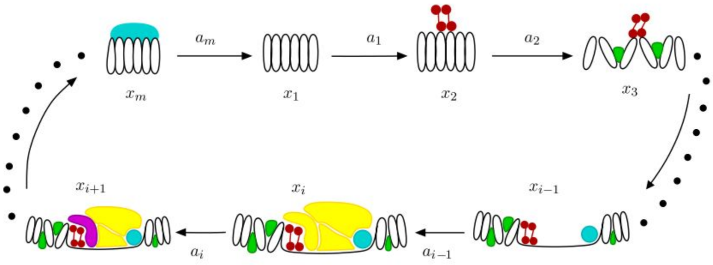

Figure 1.3.2: Cyclic network of epigenetic promoter states. Each state is represented by the set of proteins that is bound to the promoter in one instant of time. Figure taken from [Lemaire et al., 2006].

At the end of the cycle acetyl groups are removed from histones by HDAC. This order of events is consistent with the picture of opening and closing of chromatin to switch between transcriptionally active and inactive states.

Lemaire et al. translated this ratchet like cycle of chromatin states into a deterministic mathematical model. Figure 1.3.2 depicts a graphical representation of the process. Different states x_i are passed through in a cyclic and irreversible manner. An individual state is defined by the set of factors that is bound to the promoter. State x_1 represents the empty promoter. Consecutive states are connected by reactions with rate a_i. These reactions represent the binding or release of one factor to or from the promoter. Different factors here are depicted by different colors. Mathematically the reactions in the cycle can be described by a set of ordinary coupled differential equations [Lemaire et al., 2006].

$$\begin{aligned} \frac{dx_1}{dt} &= a_m x_m - a_1 x_1 \\ \frac{dx_i}{dt} &= a_{i-1} x_{i-1} - a_i x_i \end{aligned} \tag{1.3.1}$$

Where the x_i denote the individual states and the a_i the transition rates between consecutive states. m represents the total number of states in the cycle. Experimentally a single state is not accessible. A single transcription factor can bind to DNA over multiple states and thus via ChIP experiments only the integrated signal of these states can be measured. Mathematically this is stated by:

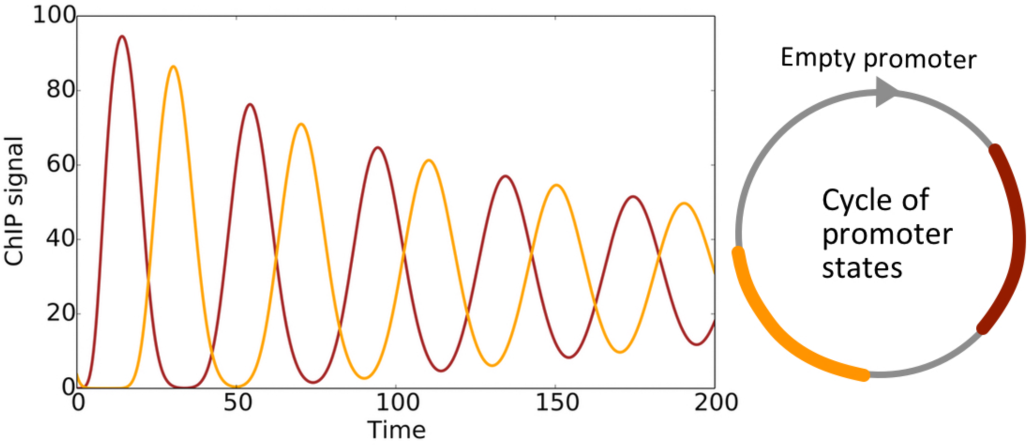

Figure 1.3.3: The lemaire model qualitatively can explain cyclic promoter occupancy. The left panel shows simulated oscillations as would have been created by the two reaction blocks on the cycle on the right. The empty promoter state corresponds to the state x_1 in figure 1.3.2. The marked colored sections of the cycle correspond to a reaction block, i.e. the sum of the states within the block forms the oscillatory signal.

$$C_f(t) = \sum_k S_{kf} x_k(t) \tag{1.3.2}$$

Where the observed ChiP signal $C_f(t)$ of a single factor f is a sum over the individual states $\{k\}$ where that factor is bound. This set $\{k\}$ of states is called a reaction block.

Figure 1.3.3 shows simulations for two hypothetical reaction blocks one early (brown) and one later (orange) in the cycle. In general such a system is able to explain the observed oscillations. The position within the cycle where a certain factor is bound defines the phase of the oscillations and the number of states where the factor belongs to the shape of the oscillations. (see figure 1.3.3).

The interpretation, however, of such a multi state system is difficult. The observed cycles of promoter occupancy are hard to reconcile with the stochastic results of gene transcription in single cells. Observed epigenetic transcription cycles show a high regularity and the only weak damping. The damping of the oscillation described by the ODE system above depends on the number of states m and scales with e^{-1/mT_0} where T_0 is the period of the oscillations [Lemaire et al., 2016]. This makes it necessary to assume a high number of states to explain the observations. As explained earlier, analysis of time course measurements of gene transcription in single cells revealed only a small number of internal promoter states. The high level of regularity at the epigenetic level appears to have only weak consequences on transcriptional activity which shows substantial variability.

Summary

This work focusses on the following two main questions:

1. Which mechanisms can explain the stochastic nature of gene transcription?

2. How does stochastic gene transcription change upon changing environmental conditions?

These two questions were investigated on the estrogen sensitive gene GREB1. GREB1 is an important factor in tumor growth in breast cancer which motivates a detailed study of its transcriptional activity. Transcription of GREB1 was experimentally investigated by live cell microscopy in single cells. Calibration and selection of multi state stochastic models to experimental data revealed the nature of the mechanism that is responsible for the observed transcriptional activity and its parameters. Experimental data obtained at different experimental conditions allowed to investigate the influence of the estrogen stimulus on GREB1 transcription and thus on model parameters.

The cyclic ODE model (equations 1.3.1 and 1.3.2) was fitted to published experimental data of transcriptional cycling. This showed that the results of transcription in single cells are inconsistent with the population measurements of transcription factor binding. For the cyclic model a high number of states had to be assumed whereas for the transcription model only two states were sufficient. Thus, there exists a so far not understood incompatibility between both types of data. Both models, however, are predictive. The cyclic model allowed to predict the combinatorial complexity of transcription factor binding to the gene promoter. That means, it predicts which pairs of factors are bound to the promoter concurrently. The stochastic transcription model together with a model of estrogen signalling to the GREB1 promoter allowed to predict the accumulation of RNA in synchronised populations of cells. Both predictions were tested with independent data sets.

2 Results

This chapter will describe the results of data analysis and model calibration to two types of data. First the cyclic model of transcription factor binding (equations 1.3.1 and 1.3.2) will be fitted to measured time courses of multiple factors. Subsequent sections describe the data analysis and modelling results of transcription measured in single cells.

2.1 Cyclic ODE models can explain transcription cycles in cell populations

It was found that multiple factors engage with the promoter of the estrogen sensitive ps2 gene in sustained cycles with a period of approximately 40 minutes as shown in figure 1.3.1 [Métivier et al., 2003]. This observation was interpreted as a set of states that a promoter traverses through in a cyclic and irreversible fashion as depicted in figure 1.3.2. Lemaire et al. proposed a cyclic ODE model to describe the observed oscillations [Lemaire et al., 2006]. Here this model was applied to multiple different factors to elucidate the temporal order of binding and release to the gene promoter. Moreover, the model was fitted with different total numbers of promoter states to estimate the minimal number of states. Moreover, overlap of the reaction blocks of pairs of factors reveals the combinatorial complexity of different factors at the gene promoter. A reaction block is the set of consecutive promoter states at which a factor is bound. Overlap between pairs of blocks indicates that both factors are bound at the same time. Testing the overlap between all pairs of fitted factors defines the combinatorial complexity of the factors. This section describes an approach to fit the proposed model and make predictions about the combinatorial complexity of factors. These predictions will be tested with independent experimental data.

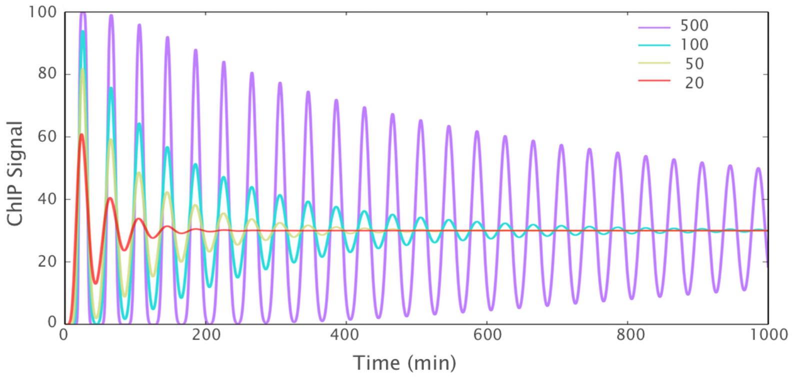

Figure 2.1.1: Systems with more states show slower damping of the oscillations. The plot shows simulations of oscillations for different numbers of states in the cycle. Systems with 20, 50, 100 and 500 states are shown as indicated. Reaction rates for the different simulations were set equal and adjusted so that that the oscillation period was 40 minutes.

2.1.1 Model fitting reveals the temporal order of epigenetic events

Model fitting was done in the following way. First the ODE system of equation 1.3.1 was simulated forward with a predefined number m of states. The reaction rates a_i were assumed to be equal, as this leads to the least pronounced damping of the oscillations (see figure 2.1.1). The value of the rates was set so that the cycling period equaled 40 minutes as found in the data. After simulation of the ODE system the optimal reaction block was fitted independently for each factor by comparison of the model to the data using equation 1.3.2.

The reaction block of one factor is defined by three parameters: the start and the end of the block and its amplitude. The amplitude value integrates contribution of a single promoter to the population signal and the efficiency of the experiment. For optimisation it was restricted to be in the range between zero and one, i.e. in an optimal experiment each promoter could maximally contribute to the population signal only once. The experimental efficiency strongly depends on the specificity of the antibody in use and makes it difficult to compare the amplitude between factors.

Start positions for optimisation were created by latin hypercube sampling [McKay et al., 1979] to ensure dense initial sampling of the parameter space. Each start position was fed into a local optimisation algorithm to eventually find the global optimum. Details of the implementation of the optimisation are given in the methods section. This fitting of reaction blocks was repeated for different numbers of states to find the minimal number

Parameter		**Lower bound**	**upper bound**
start	Start of the reaction block	0	none
end	End of the reaction block	0	none
amp	Amplitude of the reaction block	0	1
a	Relative experimental error	0.1	none
b	Absolute experimental error	10	none

Table 2.1.1: Parameters of a reaction block and bounds set for optimisation. The first three parameter define the reaction block and the last two define the experimental error for the ith data point as $\sigma_i = a y_i^D + b$

necessary to explain the experimental data. The number of states m ranged from 75 to 1000.

The published data used for model fitting did not include any experimental error. Therefore, an error model was included into model fitting. The experimental error was assumed to consist of an absolute and a relative contribution increasing the number of free parameters to five per block. This error model extended the usual χ^2 cost function. χ^2 means the sum of the least squared differences between model and data weighted by the experimental error. The weights were not known and thus were included in fitting leading to an extra term in the cost function in addition to χ^2. This term balances between a small deviation between model and data and large weights. See the methods section for details.

Lower bounds for the parameters describing the absolute and relative error were defined based on the assumed lower detection limit and the sensitivity of the applied experimental method. Table 2.1.1 summarises the free model parameters and corresponding bounds. Table 2.1.2 shows all 17 factors that were included into the model fitting. This is a representative selection out of all 46 observed factors from the original publication.

Figure 2.1.2 shows model fits of systems with different numbers of promoter states to pPol (left) and BRG1 (right) time courses. pPol represents the phosphorylated and thus active form of RNA polymerase II. pPol data can be explained well by most model sizes. The fit obtained with a system of 75 states (purple line), however, shows pronounced damping. To fit the step like oscillations of BRG1 at least 300 states are necessary. Smaller systems were not able to explain the sharp peaks and wider minima. Thus, the minimal required number of states is mostly defined by the shape of the observed oscillations.

Factor	**Funtion**
p300	Histone acetyl transferase
CARM1	Histone methyl transferase
ER_α	Estrogen receptor alpha
p68	DNA helicase
H3K4me2	Histone methylation, activating mark
TFIIB	General transcription factor
AcH3	Histone acetylation, activating mark
GCN5	Histone acetyl transferase
pol	RNA polymerase
TAF130	Subunit of general trascription factor TFIID
TRAP220	Mediator, coactivator of RNA pol II
CDK7	Cyclin dependent kinsase, cell cycle progression
pPol	phosphorylised RNA pol II, i.e. active pol II
TRIP1	Protease regulatory protein
ELP3	HAT, subunit of elongator complex
BRG1	Subunit of the SWI/SNF chromatin modifying complex
HDAC	Histone deacetylase, deactivating

Table 2.1.2: Epigenetic factors included into fitting the population model.

The shape of the oscillations differed among the selected factors although the period was the same. For example the estrogen receptor ER_α shows rather sinus like oscillations but with a significant offset, i.e. during the minimum the signal still exhibits values significantly larger than zero (see figure 1.3.1, blue line). The phosphorylated form of RNA polymerase II (pPol) showed similar behaviour but with only little offset (see figure 2.1.2, left panel, black squares). Other factors like BRG1 showed rather step like oscillations with narrow but pronounced peaks interspersed with intervals where basically no signal could be detected (see figure 2.1.2, right panel, black squares). This second form is particularly difficult to fit with systems consisting only of few promoter states.

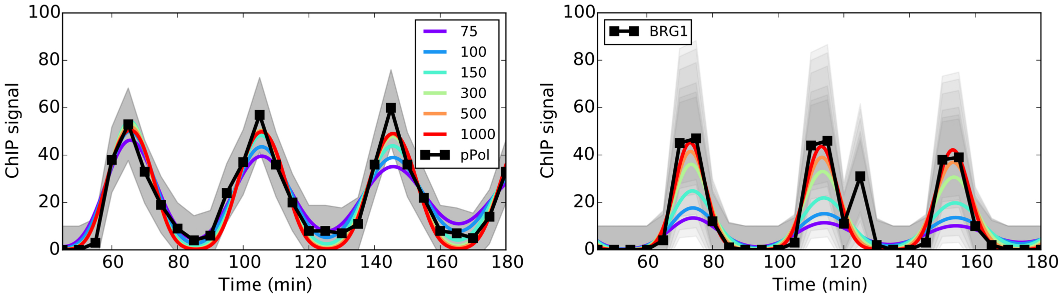

Figure 2.1.2: Cyclic ODE model can explain oscillating ChIP time courses. Model fits for different model sizes incorporating the error model. The left panel shows data (black) and fits for various model sizes (colored lines) to the pPol data. The right panel shows the same for the BRG1 data. Grey shaded regions denote the fitted error region.

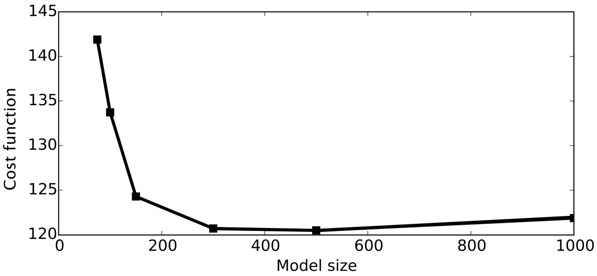

Figure 2.1.3: Several hundred promoter states are necessary to explain regular transcription cycles. Value of the cost function used to fit the model including the experimental error as function of the model size.

To assess the necessary minimal number of promoter states figure 2.1.3 shows the cost function value of the best fit for different model sizes. The curve plateaus for systems larger than 300 states. Thus, to explain the experimental data at least 300 states are necessary. This is an enormous number compared to the few states that are necessary to explain transcription in single cells and reveals a major discrepancy in interpretation of the results. The weak damping of the oscillations, however, can explain this required high number. Small systems show rapid decay of the oscillations not compatible with the data as depicted in figure 2.1.1. The purple lines for a 75 states system in figure 2.1.2 obviously exhibit too rapid damping to explain the observed oscillations.

The best fitting model allows to reveal the temporal order of the binding events at the promoter. Binding events are associated with with the start of the block and release with its end. Figure 2.1.4 shows all 17 fitted reaction blocks to a system of 300 states sorted by the start of the block from early to late. Temporal order of events is in general accordance with biological knowledge. For example first p300 adds acetyl groups to histones (AcH3) which are removed later in the cycle by HDAC. Interestingly the blocks of several factors span over a substantial fraction of the cycle. ER_α for example remains bound for more than three quarters of the cycle and thus seems to interact with most other factors.

2.1.2 Fitting reaction blocks allows to predict combinatorial complexity of factors

Overlap of the fitted reaction blocks allows to predict the combinatorial complexity of different factors and thus to test the model on different data. Combinatorial complexity means the co-occurence of pairs of factors at the promoter at the same time. A feature that is not intuitively clear from data alone. Assessing all possible pairwise interactions between the 46 observed factors experimentally would not be feasible and model fitting could suggest the most interesting candidates pairs.

An overlap of the reaction blocks of two factors within the cycle suggests that both factors are present at the promoter at the same time. Hence the model provides information about co-occurence and possible interdependency. The blocks displayed in figure 2.1.4 for example show a full overlap between ER_α and CARM1 and only partial overlap between ELP3 and HDAC. Quantifying the overlap of pairs of blocks allows to predict the likelihood to find both factors at the promoter at the same time.

Experimentally coocurrence of factors is assessed by performing Re-ChIP experiments, i.e. after a first ChIP against one factor the resulting material is subsequently used for a second ChIP against another factor [Métivier et al., 2003]. A resulting signal indicates that both factors occupied the promoter at the same time. Such Re-ChIP data was published together with the ChIP time courses and provides an independent data set for model verification [Métivier et al., 2003].

Before making concrete predictions it is necessary to assess parameter identifiability in more detail. This means how well the found parameters are defined based on the shape of the cost function. Model fits showed until here were the best fits obtained by multi start local optimisation yielding a point estimation of the global optimum. Profile likelihood provides a way to locally explore the cost function landscape along the individual parameter directions by step wise changing one parameter and optimise the remaining parameters [Raue et al., 2009; Kreutz et al., 2013]. By defining a threshold of the cost function it is possible to determine confidence intervals for the obtained

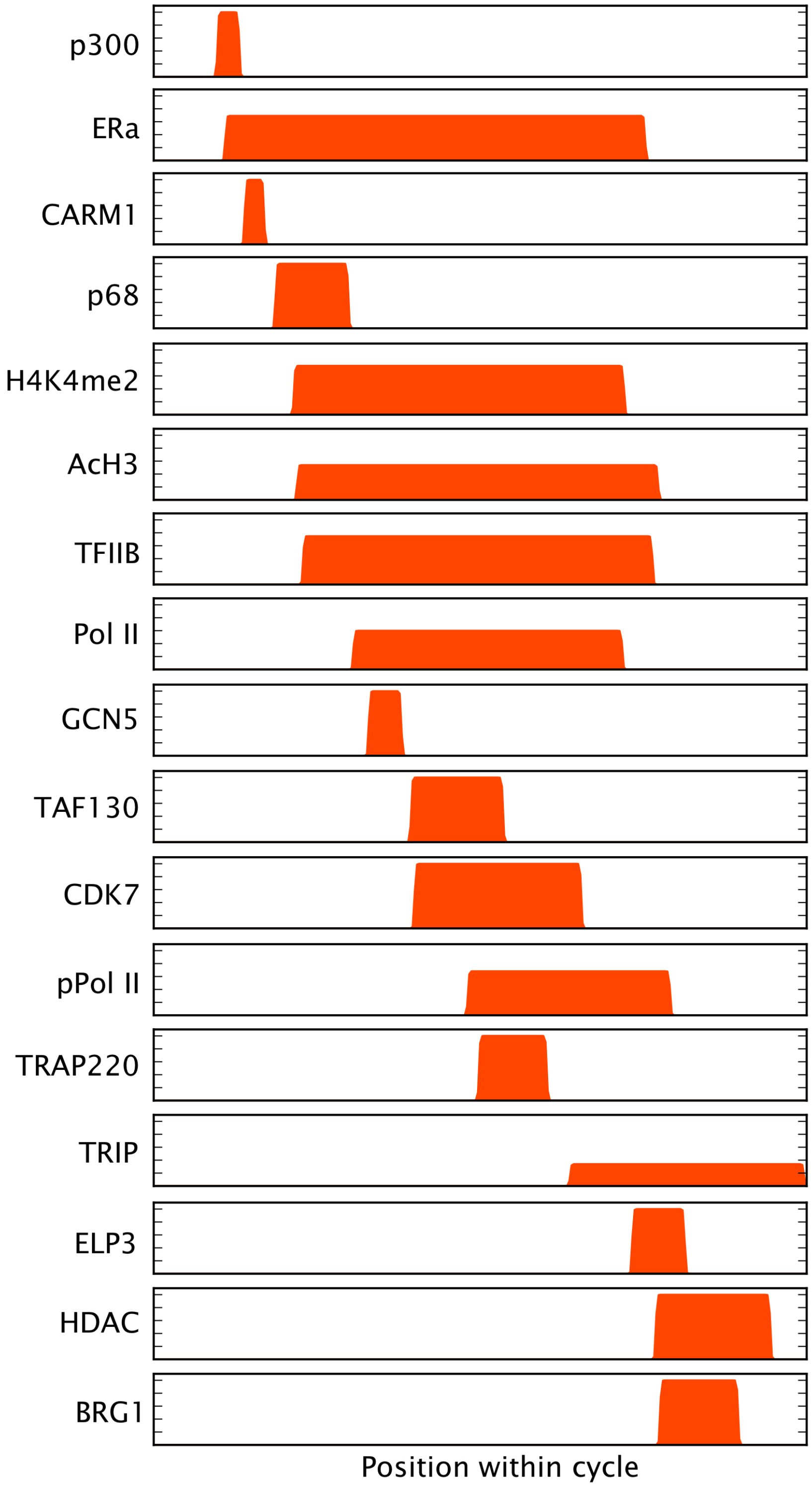

Figure 2.1.4: Fitting of reaction blocks reveals the temporal order of epigenetic events. Resulting positions and amplitudes of reaction blocks fitted to a 300 state system. Blocks were sorted from early to late in the cycle. The x-axis denotes the position within the cycle and the y-axis the amplitude of the block.

parameters. Such likelihood based confidence intervals are defined by a confidence region around the fitted optimum $\hat{\theta}$.

$$\begin{aligned} \{\theta|\chi^2(\theta) - \chi^2(\hat{\theta}) < \Delta_\alpha\} \\ \Delta_\alpha = \chi^2(\alpha, df) \end{aligned} \tag{2.1.1}$$

The threshold Δ_α is calculated using the χ^2 distribution with confidence level α and $df = 1$ for point wise confidence intervals [Neale and Miller, 1997]. If the confidence interval is narrow the corresponding parameter is well identified.

Parameters of each fitted block were profiled by changing one parameter in small steps and optimising the other parameters leading to collections of parameter sets for reaction blocks yielding fits with cost functions below the threshold Δ_α and, thus, within the confidence interval.

To predict combinatorial complexity the overlap between reaction blocks between pairs of factors were calculated. To be significant the overlap between two pairs of blocks had to be at least 50%, i.e. the overlap needed to be at least as long as half the length of the shorter of the two blocks. The predicted likelihood then was the percentage of reaction blocks with parameters within the confidence intervals that showed at least such an minimal overlap. Figure 2.1.5 displays the prediction as a color map with results obtained from fitting a system with 300 states as this provided the best compromise between a good fit and a small system size (see figure 2.1.3). Lighter blue colors indicate high likelihood of pairwise occurrence of the given factors. The displayed matrix is symmetric and in the upper half the accordance between predictions and published Re-ChIP data is indicated. If the predicted value was larger than 50% and the experiment shows a signal the prediction was assumed to be correct. Predictions were incorrect either when cooccurence was predicted but not observed or when it was observed but not predicted. Green indicates a match between prediction and data and red a mismatch. 17 fitted factors allow to predict the occurrence of 272 possible pairs 29 of which where available in the published data set. For 26 of the 29 pairs the prediction was found to be correct and incorrect only in three cases. This large majority of correct predictions indicates the strength of the model to correctly describe the biological process of epigenetic binding and release of many different factors to chromatin.

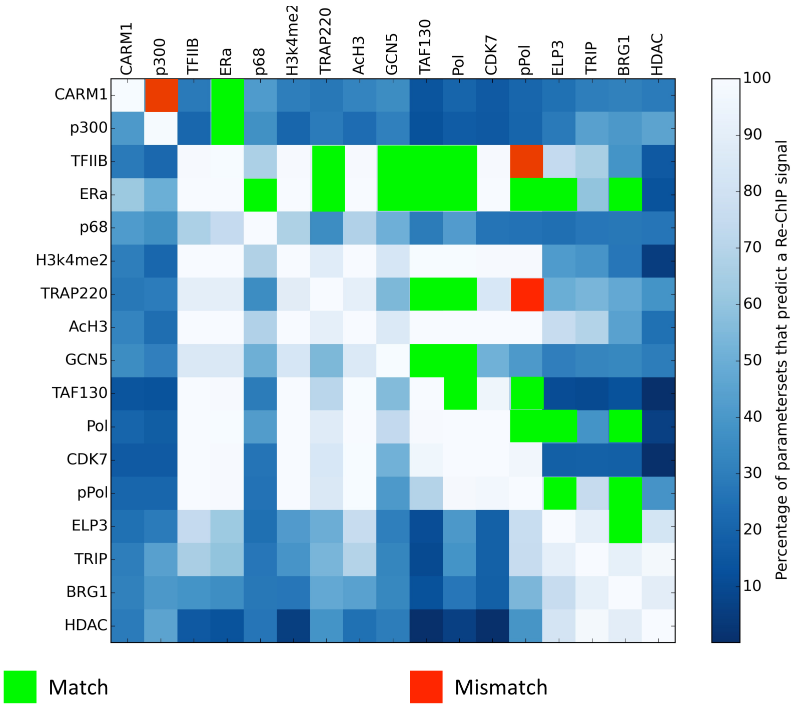

Figure 2.1.5: Fitted reaction blocks allow to predict the combinatorial complexity of promoter binding factors. Color map showing the combinatorial complexity of pairwise occurring factors. All parameter sets found by profile likelihood to deliver model fits with a cost function below a threshold were tested for pairwise overlap of the reaction blocks of at least 50%. The colormap shows the percentage of parameter sets that showed such an overlap with blue indicating low and white high values. Green and red squares indicate matches and mismatches between the model prediction and Re-ChIP experiments from Metivier et al [Métivier et al., 2003].

Summary

Mathematical modelling of observed transcription cycles by a large set of coupled ODEs allows to explain the data very well. Moreover, it enables to predict the combinatorial complexity of individual factors binding to the gene promoter. Interpretation of the results of fitting epigenetic transcription cycles in the light of recent results of transcription in single cells, however, is intricate. Model fitting of the transcription cycles led to the conclusion that several hundred epigentic states are necessary to explain the existing data. Single cell transcription, however, revealed that a gene is switching only between up to five or six states of transcriptional activity as explained earlier.

In experiments studying transcription cycles gene transcription itself was not analysed but rather changes of the chromatin environment at the locus. Those changes, i.e. binding and release of proteins or epigentic marks at the chromatin, occurred in a cyclic fashion. The method of choice to study such epigenetic interactions experimentally is chromatin immunoprecipitation, a population based biochemical method. This has several shortcomings. First the method can only detect those cells that show a signal with a certain efficiency and, thus, provide no precise way to detect the fractions of responding and non responding cells and, thus, permitting investigation of cell to cell variability. Second it includes a huge number of cells and the read out is the average of the detected cells. Third, as a biochemical method it is experimentally very elaborate which prevents a good temporal resolution in time course measurements. Live microscopy of single cells carrying a fluorescently labelled artificial transcript circumvent the shortcomings of population methods. The next section will introduce a system of an artificial reporter construct that was coupled with an endogenous promoter to observe transcription directly under the microscope.

2.2 Fluorescent labelling allows to visualise gene transcription in single cells

Experimental work including creation and characterisation of the cell line, all cell culture work, microscopy and image analysis was done by Christoph Fritzsch in cooperation with Monika Kuban from the work group of George Reid at the IMB Mainz.

The method of choice to label single transcripts in living cells to investigate transcription over time is by integrating an artificial construct into the DNA sequence of the gene of interest. Such a construct carries repeated sequences that fold into such a stem loop structure once transcribed into RNA. A stem loop is a structure where a short stretch of RNA folds back onto itself forming a loop as depicted in the inset of figure 2.2.1. These structures are recognised by a protein named PP7 which carries a GFP tag and thus place fluorescent marks on the transcript [Raj and van Oudenaarden, 2009]. Figure 2.2.1 schematically depicts the process of transcription of a gene carrying a stem loop forming section. Once the Pol II starts to transcribe the integrated section the transcribed RNA folds into stem loops that are recognised by the fluorescently labelled PP7 proteins making it visible under the microscope. While RNA polymerase II transcribes the gene the transcript remains attached to the transcription site leading to localisation the PP7-GFP.

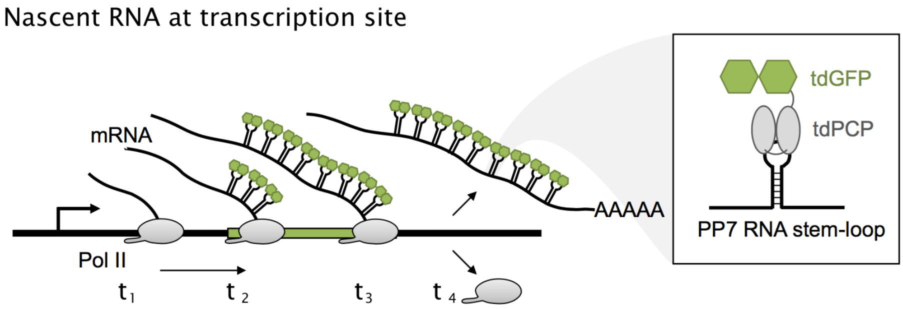

Figure 2.2.1: Fluorescently labelled PP7 visualises transcription in live cells. As RNA polymerase II (grey) transcribes the integrated artificial section of the gene (green rectangle on the black line representing the gene) the nascent transcript (thin black lines) folds into the stem loop structure. Stem loops are bound by tandem dimers of PP7 coat proteins (tdPCP) which themselves are labelled by tandem dimers of the green fluorescent protein (tdGFP) [Chalfie et al., 1994]. Before Pol II reaches the integrated section (t_1) the transcript does not carry any stem loops and thus is not visible under the microscope. Once Pol II started transcribing the integrated section stem loops start to form and are bound by fluorescently labelled PP7 core proteins (t_2). After finishing the stem loop section (t_3) Pol II continues RNA elongation but the transcript remains connected to the transcription site and thus the site remains visible. After termination Pol II falls of the DNA and releases the transcript into the nucleoplasm (t_4) where it does not contribute to the signal of the transcription site any more. Figure kindly provided by Christoph Fritzsch.

A transcription site is the actual position within the cell nucleus where transcription takes place and that is visible under the microscope as a bright spot as shown in figure 2.2.3. Thus, this method allows to follow transcription over time by measuring the emitted fluorescence light intensity at the transcription site. In vivo multiple polymerases are present at the gene each generating its own transcript. Thus the detected signal is the sum over all transcripts at the transcription site at one time point.

In previous studies the relatively simple locus of the estrogen sensitive ps2 gene was used to study the dynamics of epigenetic transitions. Simple meaning in this context that a short promoter with a single estrogen responsive element is in close proximity to the gene itself. An estrogen responsive element is a binding site to the estrogen receptor ER$_\alpha$. In addition, the ps2 gene codes for a single and rather short transcript. The GREB1 locus as a another example of an estrogen sensitive gene appears more complex than ps2 as depicted in figure 2.2.2. Experimentally, however, it turned out to be easier to integrate the desired reporter construct into the GREB1 locus than into the ps2 locus. The GREB1

Reporter length	88000 bp
Number of stem loops	24
Length of single stem loop	60 bp
Start of first stem loop	1300 bp

Table 2.2.1: Structure of the synthetic transcript used to visualize transcription. Stem loops of the transcript are recognised by the GFP labelled PP7 protein. The start of the first stem loop marks the position of the stem loop section within the full transcript.

promoter region carries at least three different estrogen responsive elements. All three were shown to exhibit transcriptional cycles similar to the ps2 locus [Sun et al., 2007]. In contrast to ps2 the GREB1 gene codes for multiple transcripts and the transcript studied here is comparatively long with a length of 88000 base pairs.

The recently developed CRISPR/Cas9 method provides an effective way to integrate artificial constructs into the genome at a well defined position [Jinek et al., 2012]. This has the advantage that the artificial gene is controlled by an endogenous promoter and thus expected to behave in the same way as without the artificial construct.

The final integrated construct into the GREB1 gene consisted of 24 repeats of stem loop forming PP7 binding sites and the necessary homology arms to combine it stably with the genome. After transcription RNA is further processed. Sections called exons of the transcript are cut out and combined to the final mRNA coding for a protein. The remaining sections of the RNA are called introns and are quickly degraded. This process of RNA processing is called splicing [Chow et al., 1977] and it was shown that splicing already can occur during transcript elongation and while the transcript is still connected to the transcription site [Coulon et al., 2014]. As integration site exon two of the full transcript was chosen as indicated in the bottom row of figure 2.2.2. Integrating the reporter construct into an exon avoids the effect of splicing out the stem loop section from the transcript which would have led to a loss of signal. Moreover, placing the reporter construct in an exon avoids to consider splicing in a mathematical description of the process. Table 2.2.1 summarises the structure of the reporter construct. The exact structure of the transcript will be important later and has to be incorporated into the model representation of the system.

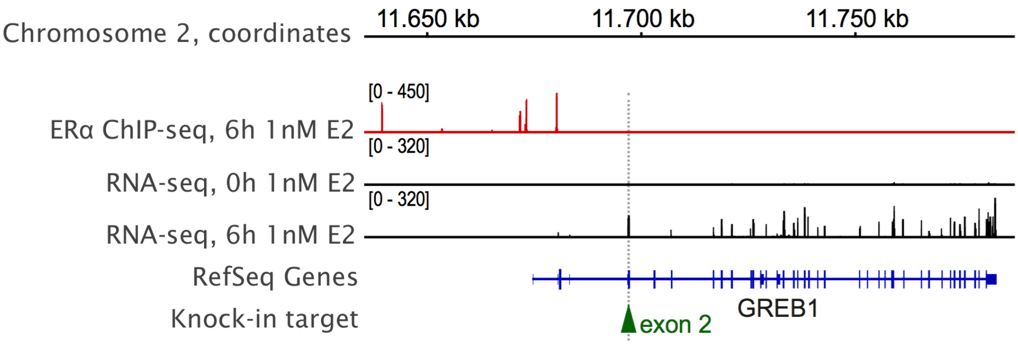

Figure 2.2.2: Structure of the GREB1 locus and position of the PP7 stem loop knock in. The GREB1 locus with coordinates on chromosome 2 (top). The second line shows peaks of ER_α binding in the promoter region as detected by ChIP-seq six hours after estrogen induction. Hight of the peaks indicates the strength of the signal. The next two lines depict the signal from two RNA-seq experiments. At zero and 6 hours after induction of the cells with 1 nM estrogen. Each peak marks the signal detected for one exon of the transcript. Annotated exons from the REfSeq data base are marked in the second last row. The dashed vertical line marks the position of the integrated stem loop forming section of the reporter transcript in exon 2.

2.2.1 The static dose response cannot explain the dynamic behaviour of transcription

Before assessing the dynamic features of estrogen dependent transcription the created cell line carrying the artificial reporter transcript in the endogenous GREB1 locus was tested to show physiological behaviour. Figure 2.2.3 shows example microscopy images of cells carrying the reporter construct. Transcription sites are visible as small bright spots within cell nuclei. Cells were imaged at different estrogen concentrations. Snapshots of cell populations showed a dose dependent fraction of cells exhibiting transcription sites with the more visible sites at higher doses. This observation was a first indication that the biological model system is responsive to estrogen and producing more transcripts at higher doses.

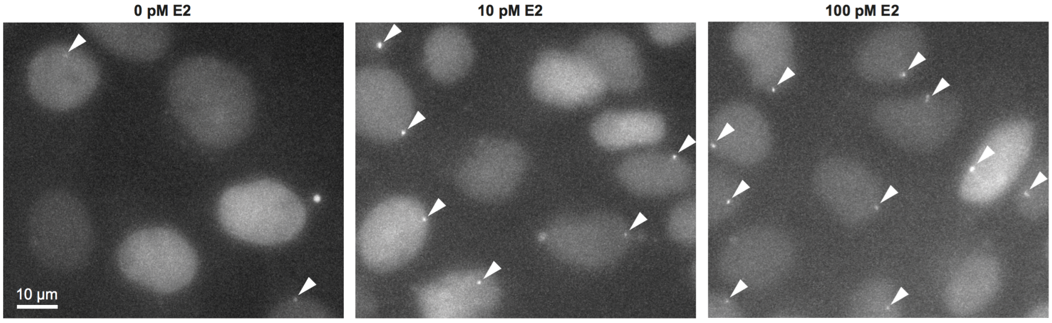

Figure 2.2.3: The number of transcription sites in cell populations is stimulus dependent. Microscopy images of cells showing transcription sites at different estrogen stimuli of 0, 10 and 100 pM (from left to right). Transcription sites are visible as small bright spots marked by arrow heads. Bright areas containing the transcription sites are cell nuclei. Shown images are maximum intensity projections of z-stacks of 12 to 17 images to capture the full volume of the cells. Images were provided by Christoph Fritzsch.

The right panel of figure 2.2.4 shows a quantitative representation of the fraction of cells showing a transcription site as a function of the estrogen concentration. The response here is the fraction of cells showing a transcription within a population at one instance of time. This curve has a typical sigmoidal shape and could be fit by a Hill-Function with a Hill coefficient of $n_H = 0.86$. Interestingly, even without estrogen a fraction of 15% of cells still showed transcription sites.

The left panel of figure 2.2.4 shows the transcriptional does response of cell populations treated with different doses of estrogen. The data was obtained by reverse transcription quantitative PCR (RT-qPCR) to quantify the amount of GREB1 transcripts. RNA was collected from cell populations, reversely transcribed into DNA and DNA fragments of the GREB1 transcripts were quantified by qPCR and normalisation against the GAPDH house keeping gene. GAPDH transcription is assumed to be estrogen independent [Barber et al., 2005]. Grey dots show data from the wild type (WT allele) and green dots from the allele carrying the reporter construct.

Both curves show typical sigmoidal shapes. Fitting of a Hill-Function found Hill-Coefficients of $n_H = 1.19$ and $n_H = 1.12$ for both alleles respectively. Both static dose responses in figure 2.2.4 clearly indicate that transcription of the GREB1 gene is estrogen dependent with higher transcription at higher doses. However, the values shown in theses curves represent static snap shots of cell populations and do not contain any temporal information, hence from such curves it cannot be learnt how the cells achieve the higher

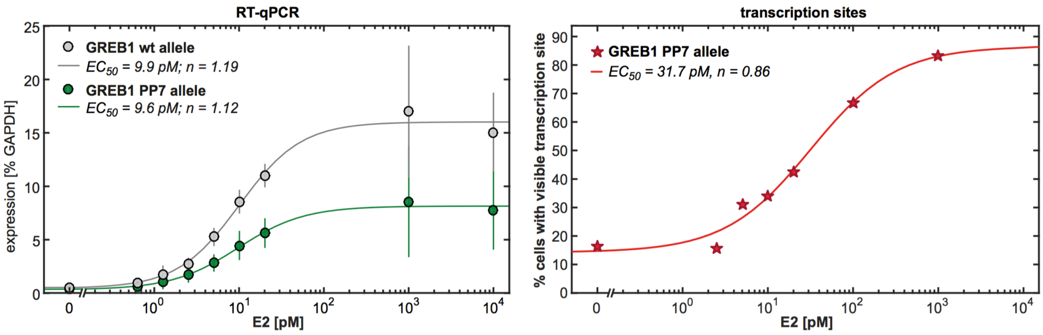

Figure 2.2.4: The transcription level is dose dependent. The left panel shows the static dose response of transcriptional activity as population everage of the wild type alleles (grey dots) and the modified allele carrying the reporter construct (green dots). Transcriptional activity was measured by RT-qPCR and is reported relative to abundance of the house keeping gene GAPDH. Error bars represent the standard deviation of three biological replicates. To both set of data points a sigmoidal hill function was fitted. Resulting parameters of half induction (EC50) and hill coefficient n are shown in the legend. The right panel displays the static dose response of the fraction of cells showing transcriptions sites within a larger population of cells, thus the reported values are based on single cells. Again in the legend the values of EC_{50} and n of a fitted hill function are indicated. Both panels were provided by Christoph Fritzsch.

number of transcripts. Several hypotheses appear plausible. i) The cells can upregulate the transcription rate k_m, ii) estrogen can influence the time the gene is in an active state or iii) a combination of both. To find the most plausible hypothesis requires investigation of the temporal behaviour of transcription.

2.2.2 Time course experiments allow to investigate the dynamic dose response

We wanted to scan the time course behaviour from low to medium and high levels of estrogen, thus we obtained different time course data sets under varying stimuli. The static dose response showed in figure 2.2.4 exhibited half induction at $EC_{50} \approx 10$ pM and full saturation at 1000 pM. This led to the selection of four different estrogen concentrations, namely 5 pM, 10 pM, 20 pM and 1000 pM. For each concentration multiple cells were observed: 53, 46, 74 and 68 cells, respectively. The observation time was 750 min with an imaging interval of 3 min. Thus, each observed cell was imaged 250 times. At each time point a z-stack of images was captured to ensure to image the whole

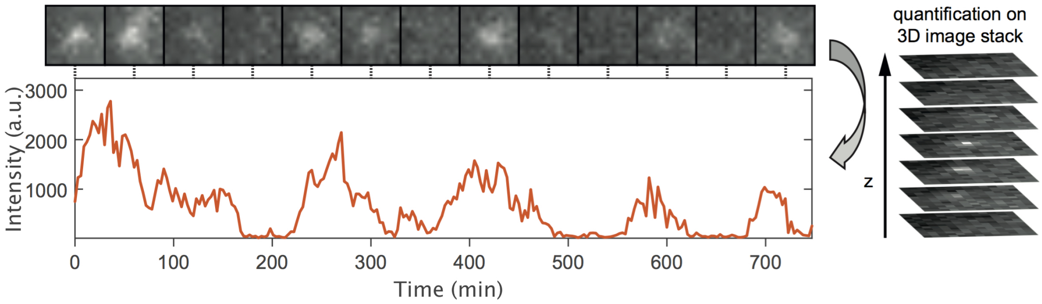

Figure 2.2.5: Image analysis of time laps microscopy images yields time courses of intensity fluctuations of transcription sites. The plotted red line represents an example intensity time profile of a single cell. The panel of images shown above the plot shows the the maximum intensity projection of the region of the transcription site at the marked time points. Images on the right show an example z-stack of an active transcription site. The figure was provided by Christoph Fritzsch.

volume of the cell. Transcription sites in the images were tracked by a software written by Christoph Fritzsch in MatLab (Mathworks, Natick MA, USA). To segment the images in a first step the cell nuclei were found. As can be seen in figure 2.2.3 the cell nucleus is brighter than the rest of the cells. This is due to the presence of free PP7 proteins that carry a GFP tag. Within each nucleus transcription sites were found by searching for the brightest spot. To that spot a three dimensional Gaussian distribution was fitted. The final signal was the integrated light intensity originating from the volume of the fitted gaussian. To ensure to follow the same transcription site over time the position of one transcription site between consecutive images were connected and checked for consistency. This tracking yielded intensity time courses similar to the one shown on figure 2.2.5. As a measure of the background signal a site within the cell nucleus distant from the transcription site was measured in each image.

2.2.3 Transcription in synchronised single cells does not show cycles

Combination of cell synchronisation followed by single cell microscopy was used to test whether transcription in single cells follows an oscillatory pattern. Cells were grown in estrogen free medium for three days to achieve transcriptional synchronisation similar to the chromatin immunoprecipitation experiments analysed earlier. The upper panel in figure 2.2.6 shows measured intensity time courses of synchronised single cells. Estrogen was added after 42 minutes of imaging. Cells imaged in estrogen free medium did not

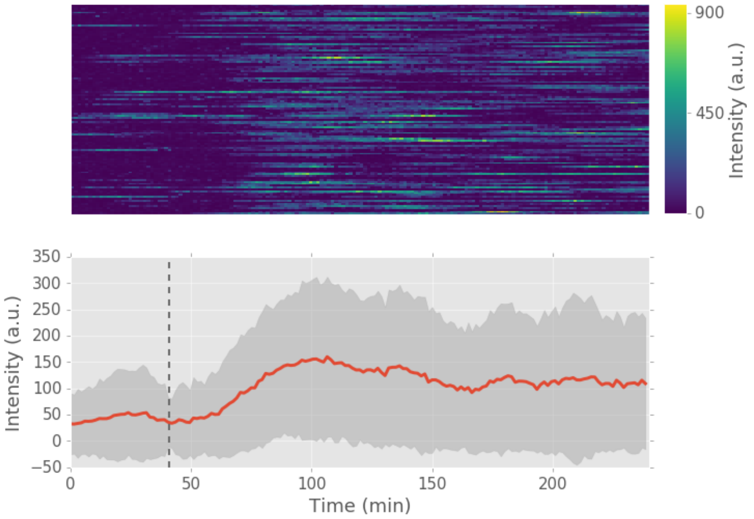

Figure 2.2.6: Transcription in synchronised single cells does not show oscillations. The top panel shows time course measurements of light intensity of single transcription sites. Each row represents on time course. In total 112 time courses are are shown. Cells were synchronised by estrogen starvation for three days. The bottom panel shows the mean light intensity value over all cells. The grey shaded area denotes the ±SD region and the dashed line indicates the time point of estrogen addition to the cells.

show substancial transcription. Soon after the addition of estrogen bright transcription sites appeared. Transcriptional activation synchronosly occurred in all cells within a small time window. Single cells did not show visible oscillations of transcription. Averaging over all cells provides a way to directly compare single cell measurements with ChIP results. The average signal of all imaged cell, however, does not exhibit oscillations as might have been expected from ChIP experiments. This further indicates the difficulty to reconcile data from both types experiments. The remaining part of this chapter concentrates on the analysis and interpretation of single cell transcription data.

2.3 Transcription in single cells is stochastic and shows dose dependent features

Data sets contained single cell time course intensity measurements of multiple cells imaged under the same experimental conditions. Exploratory data analysis revealed several features of the data sets that highlight population and single cell aspects that showed dose dependency. Those features motivated formulation of a mathematical modelling

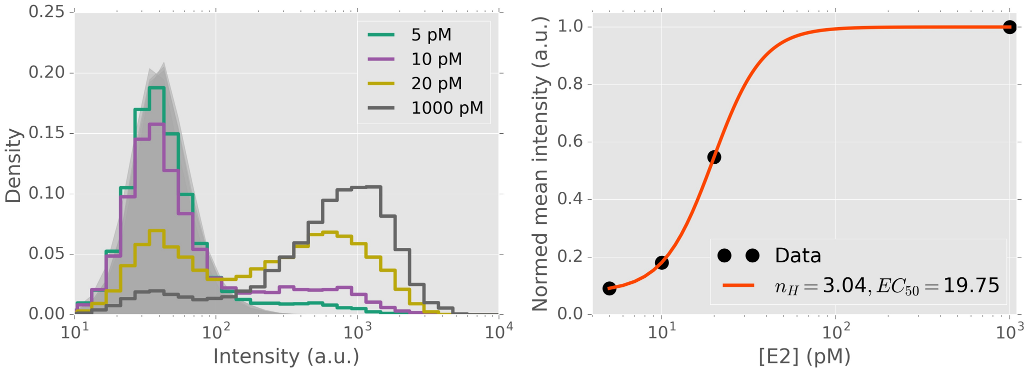

Figure 2.3.1: Global distribution of emitted light intensity is bimodal and dose dependent. The left panel shows global histograms of light intensity emitted from transcription sites (colored lines) observed at different experimental conditions. The grey shaded area denotes the distributions of the background light intensity of the different experiments. Bins are defined logarithmically and for each data set the bin count is divided by the total number of data points within the data set. The right panel shows the mean light intensity as a function of the estrogen concentration (black circles). The mean values are normalized to the mean intensity observed at 1000 pM. The orange line shows the fit of a Hill-Function. The parameter values for the Hill-Coefficient n_H and the EC_{50} are indicated in the figure legend.

frame work and in addition were used during model calibration. This section will give an overview of the statistical analysis that was applied to experimental data. Moreover, it will give some direct hints of important model parameters and their dose dependent regulation.

2.3.1 Transcriptional activity shows bimodal distributions

The distribution of all measured intensity values within one data set gives a broad overview of the global behaviour of the observed cell population irrespective of the time course nature of the data. The left panel in figure 2.3.1 shows histograms of measured fluorescence intensities for all four investigated estrogen stimuli (colored lines). The grey shaded area depicts the background fluorescence intensities for the four different experiments. The background did not vary significantly between experiments (see also table 4.3.1 in the methods section). The distributions change characteristically with increasing stimulus. At the lowest stimulus most of the observed fluorescence falls into the background peak (green line) indicating that cells rarely respond and if they do not exhibit high intensity values. With increasing stimulus a second peak above the background becomes more pronounced leading to a bimodal intensity distribution. At a saturating estrogen concentration of 1000 pM almost all detected signal lies above the background showing that transcription sites appear to be active most of the time (dark grey line). In addition, the right peak of the histogram is shifted to the right with increasing stimulus indicating a transition in behaviour from digital to analog. This characteristic distribution of fluorescence intensities indicates the burst like nature of transcription. A feature that was used to compare experimental data with simulated data sets during model fitting.

On average single cells are expected to show a similar sigmoidal dose response behaviour as cell populations. The right panel of figure 2.3.1 shows the mean intensities from the histograms on the left as a function of the stimulus. Mean intensities were normalised to the mean intensity of the 1000 pM data set. The values show the typical sigmoidal behaviour of a dose response curve. Fitting of a Hill-Function (orange line) revealed a Hill-Coefficient of $n_h = 3$ and an $EC_{50} = 20$. Thus, the dose dependent increase in mean light intensity is much steeper than in the transcriptional dose response shown in figure 2.2.4 where a Hill-Coefficient of $n_H \approx 1$ was observed. A discrepancy that might be due to the different methods of observation and of the different data sets.

2.3.2 Dose dependent features of stochastic single cell transcription

The aim of this work is to study the dynamics of transcription which made it necessary to perform time course experiments. A tool to investigate the temporal behaviour of a dynamic stochastic process is its autocorrelation function (ACF). It measures the temporal interdependence of consecutive time points and hence allows to detect regularity in single cell time course data.

Figure 2.3.2 gives an overview of three features of the different data sets via colormaps, namely the data itself rescaled to z-scores (left), the individual autocorrelation functions (center) and the distribution of z-score values of single cells. Each row of panels shows features of one data set. The estrogen stimulus is indicated on the left. Red colors in all color maps represent high numerical values, white intermediate and blue low values (see the color bars above each column). Pixel rows of panels within one row of panels correspond to the same cell.

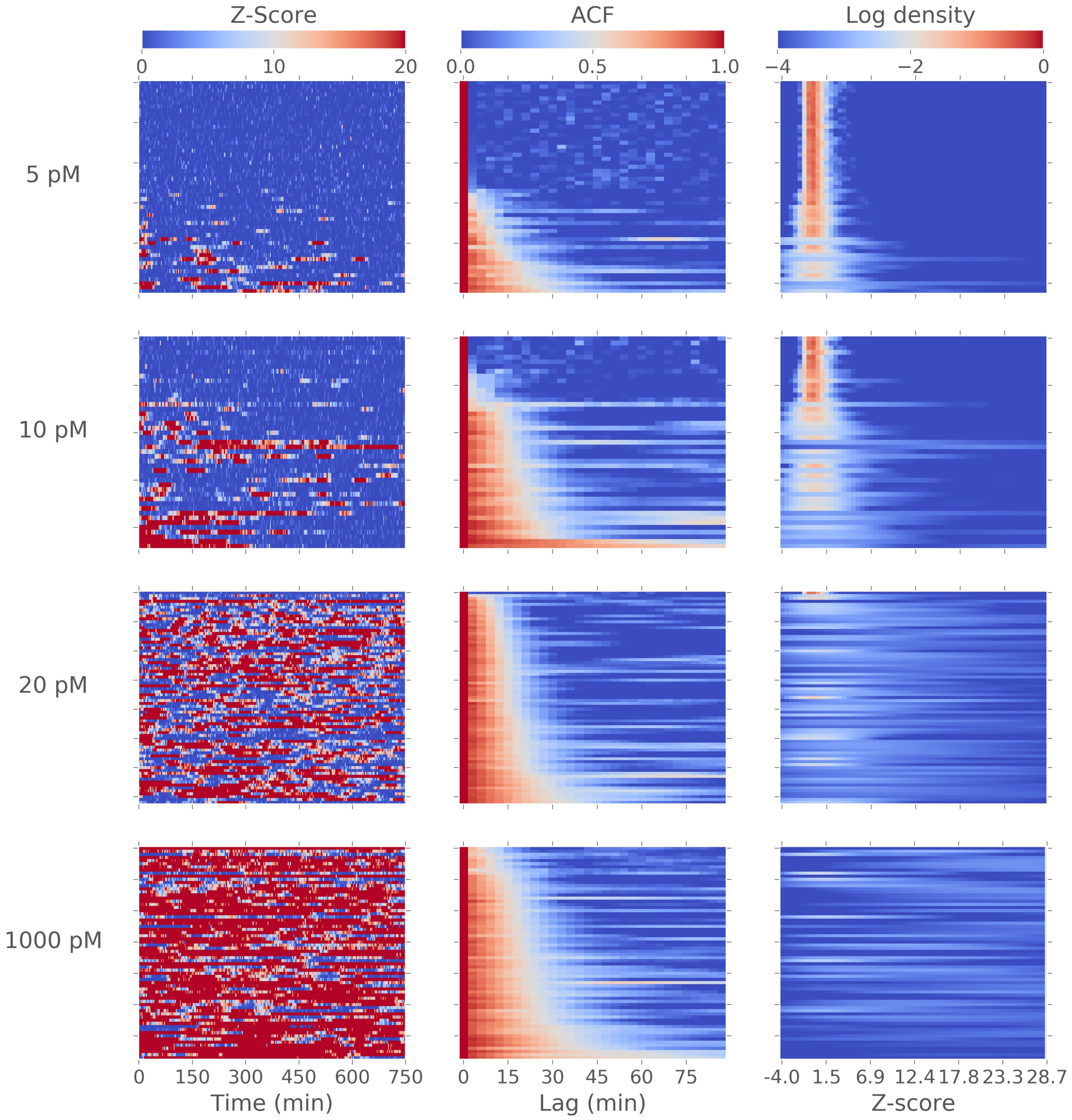

Figure 2.3.2: Single cells show dose dependent dynamic features. Colormaps to visualise features of cell to cell variablity. Columns represent different features and rows represent experimental conditions. Estrogen concentrations are indicated on the left of each row. The data sets consisted of 53, 45, 73 and 68 cells for stimuli of 5 pM, 10 pM, 20 pM and 1000 pM respectively. Each row of pixels within a colormap represents one cell and the cells were sorted according to their autocorrelation half life from short (top) to long (bottom). The left column displays z-score values calculated according to equation 2.3.1. The middle column represents the corresponding autocorrelation functions. For the distribution of z-score values of each individual cell a kernel density estimation (KDE) was performed. The right column shows the log density values of the KDE. Colorbars relating color with numerical values are shown at the top of each column.

The left column shows the light intensity values transformed to z-scores according to:

$$z_{score} = \frac{data - \bar{bg}}{\sigma_{bg}} \tag{2.3.1}$$

Where $\bar{bg}$ denotes the mean and σ_{bg} the standard deviation of the back ground signal. This reduces the dynamic range of the original data from 1 to 10000 to a range from approximately -4 to 30 and allows to better compare data sets side by side. The colormaps clearly indicate that transcription is stochastic and that the level of transcription increases with increasing estrogen stimulus as already shown in the histograms of figure 2.3.1.

The central column depicts Autocorrelation values. ACF was calculated following the usual definition in statistics as:

$$ACF(\tau) = \frac{E[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2} \tag{2.3.2}$$

E represents the expectation value operator, τ the time lag and μ the mean intensity value. A time course of a purely random signal would show an instantaneous drop in the ACF to zero and hence no autocorrelation. Such behaviour would be expected for the background signal. Time courses carrying temporal information and that hence are not completely random would exhibit a slower decay of the autocorrelation function. These two extreme cases easily allow to detect cells that are responding to an external stimulus or not. With the latter showing an instantaneous drop in the ACF representing only a random signal. This indeed was observed for multiple cells in the 5 pm, 10 pM and 20 pM data sets. This fraction of not responding cells decreased with increasing estrogen stimulus.

Cells within one data set where sorted according to their ACF half life from short (top) to long (bottom). ACF half live is the lag at which the ACF drops to a value of one half. The ACF half life of the responding cells show a characteristic spread indicating cell to cell variability. This is visualised by the triangular pattern in the lower left corner of the color maps. The mean autocorrelation of a data set and the distribution of autocorrelation half lives were used during model calibration to compare simulations with data.

In the right column the distribution of the z-score values of each single cell is shown by its log density. The distribution of observed intensity values for a single cell is another interesting feature describing the transcriptional activity. In case of a non responding cell most of the measured intensity values is close to the background signal which can be seen in case of the 5 pM and 10 pM data sets. Responding cells exhibit a much wider spread in the distribution of their exhibited intensity values. A histogram of measured values from a single cell is sparse and difficult to visualise due to the limited number of observations (n = 250). To overcome this and to gain a single cell intensity distribution the kernel density estimation (KDE) is displayed.

Data representation in figure 2.3.2 highlights several features of the data sets. The depiction in the left column highlights that transcription is apparently intrinsically stochastic. Moreover, it shows that with increasing stimulus more cells are responding by showing high fluorescence signals. In the 5 pM data set almost half of the cells are not responding. Showing immediate drop in ACF and a strongly peaked distribution of intensity values around a z-score of 0. This fraction of non responsive cells decreases with increasing stimulus to only one cell in the 20 pM and none in the 1000 pM data set. ACFs sorted by their half life from short to long show a characteristic triangular pattern in the lower left corner of the color maps indicating pronounced variability among responding cells within each data set (middle column). Showing high intensity values (z-scores) strongly correlates with ACF half life with highest numbers at longest half lives. At the same time cells with long ACF half lives show a wide spread in intensity distribution. The first cell in the 1000 pM data set is active during the whole observation time and thus does not show regulation leading to a quick drop in the ACF.

2.3.3 Transcriptional activity may be frequency modulated

As explained earlier transcription for many genes was found to switch stochastically between active and inactive states. An interesting feature is how long either phase is and whether they are influenced by experimental conditions. Finding peaks within the observed time courses and estimation of their width and relative distances allows a rough estimation of active and inactive times of the gene. A peak indicates transcriptional activity. Peak width is a measure of the active time and the distance between peaks is a measure of inactive times. Peak detection was performed by first smoothing the raw data by a median filter with a window size of three data points followed by transformation of the resulting time course to z-scores. Thresholding of the z-scores created a binary mask with values of one indicating high intensity values of a peak corresponding to transcriptional activity and zero values indicate transcriptional inactivity. Gaps in the

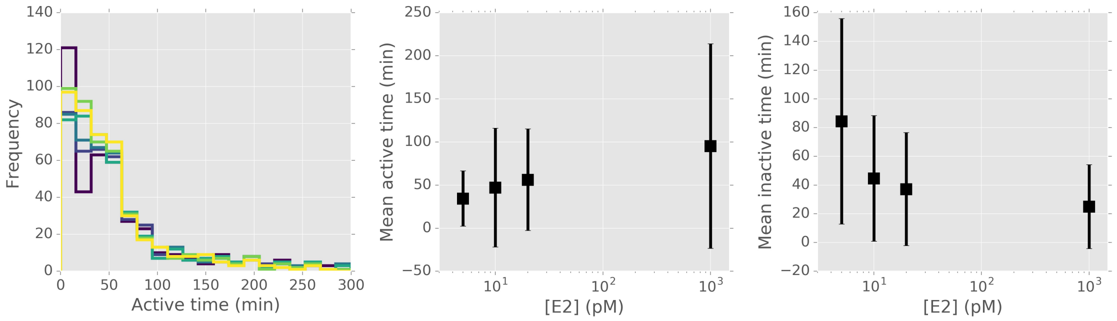

Figure 2.3.3: Active and inactive times extracted directly from data indicate dose dependent regulation. Approximate estimation of the duration of active and inactive times were done by transforming the data to z-scores and finding consecutive data points above (active) or below (inactive) a threshold. The left panel shows histograms of extracted active times from the 10 pM data set for different threshold values ranging from 1 (purple) to 6 (yellow). The middle and right panel show the mean values of the active (middle) and inactive times (right) as functions of the estrogen stimulus for a threshold value of four. Errorbars represent ±SD regions. Mean and standard deviation were calculated for all peaks found in one data set.

mask, i.e. a single zero within a sequence of ones ore vice versa, were closed to avoid finding random peaks or minima. The number of consecutive data points above or below the threshold multiplied by the imaging interval of three minutes are measures of active or inactive times respectively.

The left panel in figure 2.3.3 shows histograms of found actives times within the 10 pM data set. Colored lines represent histograms for different threshold values from one (purple) to six (yellow). Thresholds above a value of four (dark green) do not show significant changes in the distribution and thus four appears to be the lowest threshold providing robust values values for on and off times. A z-score of four indicates data points that are four standard deviations above the background signal. A number that appears significant to consider the gene to be active. The middle and right panel of the figure show mean active (middle) and inactive (right) times extracted from the experimental data as functions of the estrogen stimulus. Both times show weak but sustained trends, namely that the active time is increasing and the inactive time is decreasing with increasing stimulus. This results provide a hint that the adaptation of transcription to the stimulus is done by adjusting the gene's times of activity. These results, however, have to be handled with care because the estimated times might be incorrect due to technical reasons.

On time over estimation

The width of the peaks is largely defined by the velocity of the RNA polymerase as it transcribes the gene. The nascent transcript stays connected to the polymerase during elongation and remains visible as explained in figure 2.2.1. A polymerase velocity of 2.5 kb/min results in an elongation time of 34 minutes of the observed transcript. This time provides a lower bound of the peak width that is detectable by this method although the actual on time might be much shorter.

Off time under estimation

To find an inactive interval it has to be flanked by two peaks. Inactive times at the beginning or the end of the observed time courses cannot be estimated. An effected that is most prominent at low stimuli because peaks there are less frequent. At high stimuli the inactive times might become shorter than the transcript elongation time thus leading to overlapping peaks that cannot be resolved.

A mathematical model resembling the stochastic mechanism of transcription would be able to overcome such short comings and model calibration provides a way to correctly estimate parameters. Nevertheless, the rough estimation of the time of gene activity already hints to important model parameters that could be dose dependent.

Summary

As summary of this exploratory data analysis it can be said that single cell transcription data show very interesting and non intuitive features. Transcriptional activity of the observed model system is dose dependent as shown by the static dose response. From the measured single cell time courses it appears obvious that transcription is a stochastic process showing a substantial amount of variability during the observation time. In addition, cells can behave very differently under the same experimental conditions. First cells are either responding to an external stimulus or not and secondly responding cells show variable behaviour reflected by a characteristic spread of the autocorrelation half lives. This cell to cell variability appeared to be dose dependent with higher variability at low stimuli. Such dose dependent features cannot intuitively be explained by simple qualitative assumptions and motivate a mathematical modelling frame work. Model fitting would allow to find parameters that are controlled by an external stimulus and thus to explain the stochastic but dose dependent single cell behaviour.

2.4 Hybrid model to understand stimulus dependent single cell transcription

A mathematical model of transcriptional (in)activation of the biological model system under study has to incorporate the features found in the experimental data. Those are the intrinsic stochasticity, the bimodal global distribution of intensity values and the characteristic ACF. At the same time a model would have to reproduce substantial cell to cell variability.

2.4.1 Model topologies of different complexity

As explained earlier in existing literature mathematical models of gene transcription were implemented with at least two promoter states, i.e. one on state where transcription can occur and one off state where no transcription takes place [Paulsson, 2005]. Switching between states usually is assumed to occur at random times and thus the process is modelled as a continuos time markov process to incorporate the necessary intrinsic stochasticity. Recent studies found that gene promoters may not only switch between two states but rather have multiple inactive states [Suter et al., 2011; Zoller et al., 2015] forming a cyclical biochemical reaction network. Fitting of a cyclical model to ChIP time courses also suggested the existence of multiple promoter states connected in a cycle.

We compare promoter cycles of different complexity in their ability to explain experimental data. Figure 2.4.1 shows model topologies corresponding to published models. In the upper part the cycle of promoter states is displayed. The smallest possible cycle of only two states is shown on the left, whereas on the right two longer cycles are shown representing models of higher complexity. The three state model has two inactive promoter states and the four state model incorporates an additional on state. By this scheme the set of model topologies can be extended by adding more states in either the active or inactive phase of the cycle. Switching between states occurs with the rates k_{on_i} and k_{off_i} respectively and is implemented as an irreversible ratchet like process as proposed in existing literature [Reid et al., 2009; Zoller et al., 2015].

Due to the structure of the artificial transcript the fluorescent signal that is visible under the microscope is delayed relative to transcriptional initiation. As described earlier RNA polymerase first has to transcribe the stem loop section of the reporter construct for the fluorescent marker to bind and visualise it (see figure 2.2.1). This generates a first

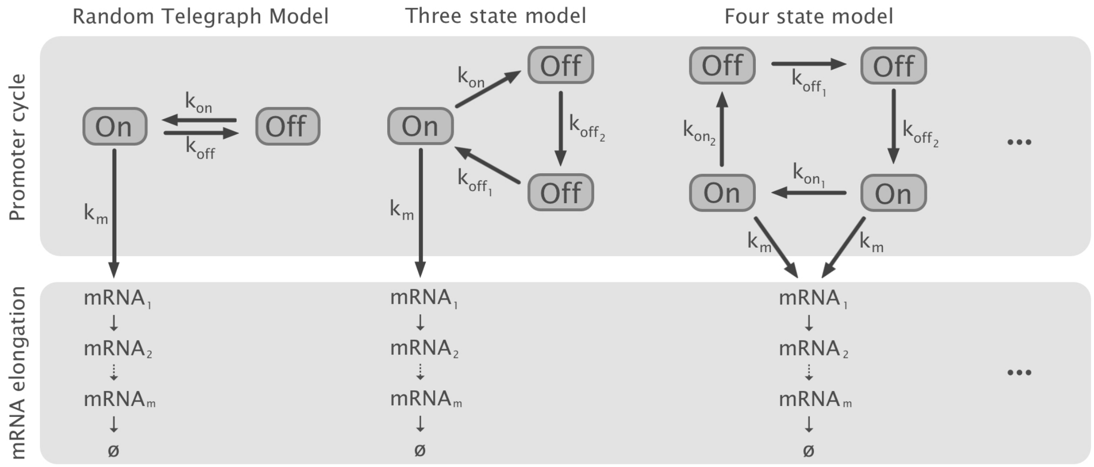

Figure 2.4.1: Model topologies of different complexities. The upper part of the figure shows three different versions of a multi state promoter cycle with increasing number of states from left to right. By adding more promoter states into the cycle more complex models can be created. Transition rates between between consecutive states are denoted by k_{on_i} and k_{off_i}, respectively. k_m denotes the transcription initiation rate by which new transcripts are created during the promoter on phase. The lower part of the figure depicts the states of RNA elongation that are necessary to incorporate the delay between transcription initiation and fluorescence signal into the model. ø denotes the termination of transcript elongation where the transcript falls off the transcription site and the fluorescent signal disappears.

delay τ_1. After a second delay τ_2 RNA polymerase has completed the transcript which then falls off the transcription site and is not contributing to the detected signal any more. A general approach to model such a delay is to introduce extra states $mRNA_i$ for RNA elongation as illustrated in the lower part of figure 2.4.1. Only a subset of the RNA states is visible under the microscope due to the position of the stem loop section within the transcript. Transition of the RNA state $mRNA_i$ to $mRNA_{i+1}$ again is stochastic and transition rates are assumed to be equal. Previously the transcriptional velocity of RNA polymerase II was estimated to ≈ 2.5 kb/min [wa Maina et al., 2014]. Assuming this velocity and the transcript length of 88 kb leads to an average RNA dwell time of 34 min at the transcription site. Thus, the life time of the individual mRNA states had to be adjusted accordingly.

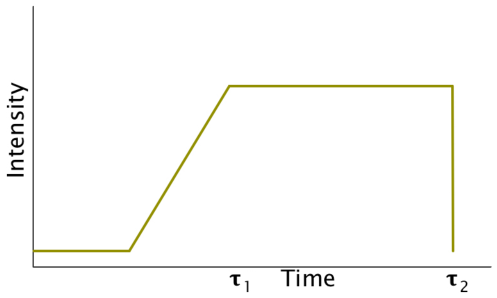

Figure 2.4.2: Deterministic temporal intensity profile of a single transcript. Scheme of the deterministic signal created by a single transcript. The two delays τ_1 and τ_2 denote the times when the transcript becomes fluorescently visible after initiation and when it is released from the transcription site. The first is defined by the position of the stem loops within the the transcript and the second by the full transcript length. The transition of the intensity from the low to the high plateau is the phase when RNA polymerase transcribes the stem loop section of the transcript. Every newly transcribed stem loop is bound by a fluorescent marker and thus increases to the visible signal. The integrated transcript carries 24 stem loops of 60 bp each. Thus the stem loop section is 1440 bp long. Assuming a elongation rate of 2500 bp/min this section is transcribed in 36 seconds which is much shorter than the imaging interval of three minutes.

2.4.2 Numerical implementation

On the biochemical level intrinsic stochasticity of transcription is caused by random interactions of reacting proteins available only in low copy numbers. Many proteins involved in gene transcription and chromatin remodelling are present in the cell nucleus only in small copy numbers, e.g up to a few thousand molecules per species [Biggin, 2011]. The binding target for these proteins is the gene promoter which is only present in the genome in two copies. This means the few molecules need to find a even fewer targets to activate them. Therefore, biochemical interactions between enzymes and template have to be modelled stochastically.

Numerically the temporal evolution of a markovian system can be simulated using the well known stochastic simulation algorithm (SSA) [Gillespie, 1976, 1977]. For details about the SSA implementation and model parameterisation please see the methods section.

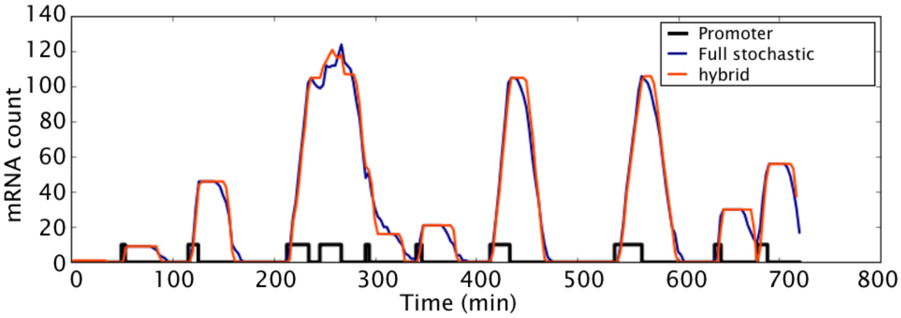

Figure 2.4.3: Hybrid stochastic-deterministic simulation provide a close approximation to full stochastic simulations. Comparison between simulated RNA counts at the transcription site for the full stochastic simulation (blue) and using the initiation events to assemble a hybrid path (orange). Promoter activity is shown in black.

Simulating the full stochastic signal led to extended computation times which significantly delays model fitting with respected simulations. To gain computational speed a hybrid stochastic-deterministic method was implemented. Instead of stochastically simulating the mRNA elongation by multiple extra states the mRNA part of the model was simulated deterministically. The deterministic signal created by a single transcript was completely defined by the structure of the transcript (length, position of the stem loops) and the velocity of the RNA polymerase as depicted in figure 2.4.2. Assuming such deterministic temporal signals was successfully used by Larson et al. to measure the elongation rate of nascent transcripts [Larson et al., 2011]. RNA polymerase velocity is assumed to be constant over the whole transcript. Thus, the position of the stem loops within the transcript defines the first delay τ_1 between transcript initiation and appearance of the fluorescence signal. The length of the transcript defines the second delay τ_2 when the transcript is finished and released from the transcription site. The stochastic part of the simulation created the time points of transcriptional initiation events and the final time course consisted of the sum over all initiated single transcript signals. Hybrid stochastic-deterministic simulations shows only minor deviations from fully stochastic simulations (see figure 2.4.3) and yielded an approximate 50 fold acceleration in computation time.

2.4.3 Intensity of single transcripts relates RNA simulations with experimental data

Stochastic simulations of the model create absolute RNA counts over time whereas the experimental read out is fluorescence light intensity. We measured the light intensity of single transcripts to allow direct comparison. The intensity of a single transcript was measured independently and used as a factor of proportionality between simulated RNA count and measured light intensity.

The measured signal is afflicted by back ground noise. The background signal was measured independently and found to exhibit a log normal distribution. Parameters of a fitted lognormal distribution to the background signal were used to add noise to the simulations.

Before the comparison with experimental data the simulated RNA time courses had to be scaled with the estimated factor and subsequently log normally distributed noise had to be added with the fitted parameters. For details please see the methods section

2.4.4 Off time modulation as a candidate mechanism to explain dose dependent modification of the global intensity histogram

Before the model was fit to data it was tested for its abilities to explain features of the data qualitatively. Extraction of active and inactive times of transcription directly from the experimental data suggested a dose dependent modulation (see figure 2.3.3). Figure 2.4.4 shows features of three simulations. A two state model was used with the parameters: on time $t_{on} = 7$ minutes and initiation rate $k_m = 5$ min^{-1} kept fixed for all three simulations. The promoter off time T was set to values of 700, 200 and 30 minutes from left to right. The upper panels show the global histograms of the simulations similar as for experimental data in figure 2.3.1. The bimodal nature exhibited by the experimental data can qualitatively be well explained. In addition, the dose dependent modulation of the peaks of the histograms as observed experimentally could be explained by modulation of the promoter off time. The displayed histograms already closely resemble those of the 5 pM, 10 pM and 20 pM data sets. This off time modulation additionally could explain the different fractions of responding and non responding cells. The variability in autocorrelation half lives of responding cells, however, could not be accounted for. The lower part of figure 2.4.4 shows the individual ACFs sorted by their

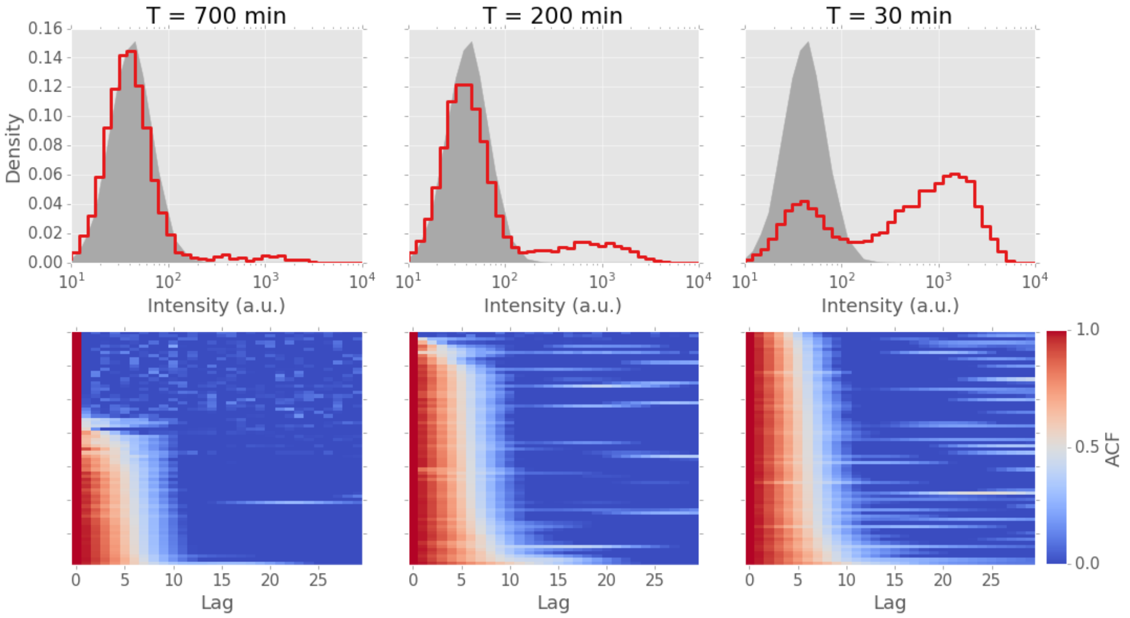

Figure 2.4.4: Stochastic simulations together with off time modulation can explain global dose dependent features The figure shows the global histogram (top) and the autocorrelations sorted by their half life as colormap for three simulations. Simulations were done using a two state model with fixed on time to $t_{on} = 7$ min and initiation rate to $k_m = 5$ min^{-1}. The off time was varied between 700, 200 and 30 minutes.

half life as colormaps similar to the middle column of figure 2.3.2 for experimental data. The characteristic triangular pattern in the lower left corner could not be resembled and the autocorrelation half lives did not show much variability.

At the current state the model can qualitatively explain several features of the data. Namely the bimodal global distribution of intensity values. In addition, modulation of the promoter off time can explain dose dependent global behaviours. Moreover, off time modulation is capable to explain one part of cell to cell variability in that it controls the fraction of non responding cells. Cell to cell variability among cells that are responsive cannot be accounted for, yet. This requires further model extensions which will be introduced in the next section.

2.4.5 Parameter perturbations can model variability among responding cells

Individual cells showed a high variability in their ACFs leading to characteristic patterns when sorted for the ACF half lives as displayed in figure 2.3.2. This cell to cell variability between responding cells can be introduced into the model by perturbations of individual model parameters. The underlying assumption is that the cellular state among different cells varies even under the same experimental conditions. Within the simulation of a data set one or more parameters of the model can be perturbed by resampling from a certain distribution, i.e. for each simulated cell an individual parameter would be sampled. Such perturbations represent extrinsic sources of variability like varying protein copy numbers among cells within one population. Perturbations are assumed to be stable over time, i.e. do not change over the course of the simulation.

Included were individual perturbations on the RNA polymerase velocity, the initiation rate, the on time and the off time. For instance, varying ATP levels are responsible for varying RNA polymerase velocity [Johnston et al., 2012], varying copy numbers of RNA polymerase can cause different initiation rates. Copy number variations in other proteins necessary to activate transcription can cause variations in on or off times. In addition to single perturbations, combinations of two perturbations were included, as well.

Perturbed RNA polymerase II velocities were sampled from a uniform distribution $v_{pol} \sim Unif(1, 5)$ kb/min. The bounds of the distribution are motivated by published values for high [Darzacq et al., 2007] and low RNA pol II velocities [Ardehali and Lis, 2009; Boireau et al., 2007] measured as population averages. All other parameters were perturbed by normal distributions. The standard deviation of that normal distribution is a measure for the strength of the perturbation. To avoid negative values the absolute value of the sampled parameter was taken.

Figure 2.4.5 displays the sorted individual autocorrelation functions for four different simulations. From left to right are shown simulations with no perturbation, perturbation of the polymerase velocity alone and two combinations of perturbation of polymerase velocity and initiation rate with small and large perturbation strength σ_{k_m}. As model again a two state model was used with initiation rate set to $\sigma_{k_m} = 5$ min^{-1}, on time to $t_{on} = 7$ minutes and off time to $T = 200$ minutes. Introducing a perturbation allows to resemble the characteristic triangular pattern of ACF decay that was observed in experimental data (see figure 2.3.2 for comparison). Thus, parameter perturbations are

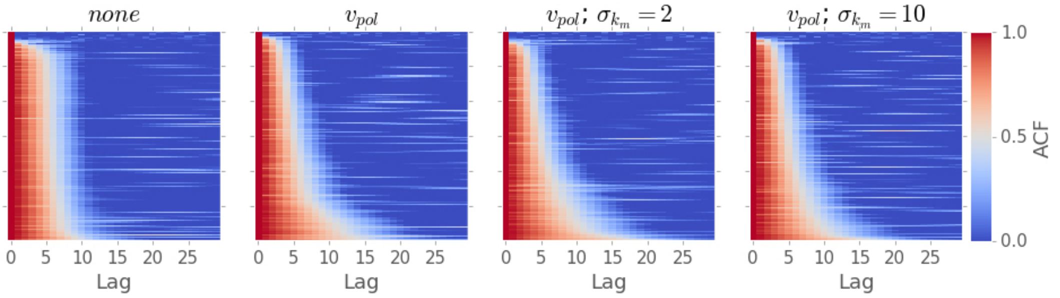

Figure 2.4.5: Parameter perturbations are necessary to explain variability in autocorrelation half life. The color maps show the autocorrelation functions sorted by their half life. Shown are four different simulations created by using a two state model with fixed parameters (t_{on} = 7 min, k_m = 5 1/min and T = 200 min) but different settings for parameter perturbations. From left to right are shown results for: no perturbation, perturbation of the polymerase velocity v_{pol} alone, two combinations of perturbation of the polymerase velocity and i) weak σ_{k_m} = 2 min^{-1} and ii) strong σ_{k_m} = 10 min^{-1} perturbation of the initiation rate.

necessary to implement into the model. Their specific influence will be highlighted in the next paragraph.

The global intensity distributions showed pronounced tails towards high intensities (see figure 2.3.1). A feature most prominent in the 5 pM and 10 pM data sets. Figure 2.4.6 investigates the influence of the type and strength of parameter perturbation on the global intensity histogram using the simulations from figure 2.4.5. The inset enlarges the right peak of the bimodal distribution. In this data representation the influence of the different types of perturbation becomes apparent. The histograms for the simulations without perturbation, perturbation of the RNA polymerase velocity alone and combined perturbation of RNA polymerase velocity together with only a weak perturbation of the initiation rate are very similar. The right peak of the histogram in case of a strong perturbation of the initiation rate together with perturbation of the polymerase velocity is flatter and broader than the other three. This appears natural as perturbation eventually leads to high initiation rates resulting in more transcripts and thus higher signals. The area of the peak remains constant because of the symmetric nature of the perturbation which is realised by a normal distribution. Hence the peak flattens. Such an effect appears desirable for example to explain the 5 pM data set where the global intensity perturbation shows a flat but rather long tail of intensities above the background.

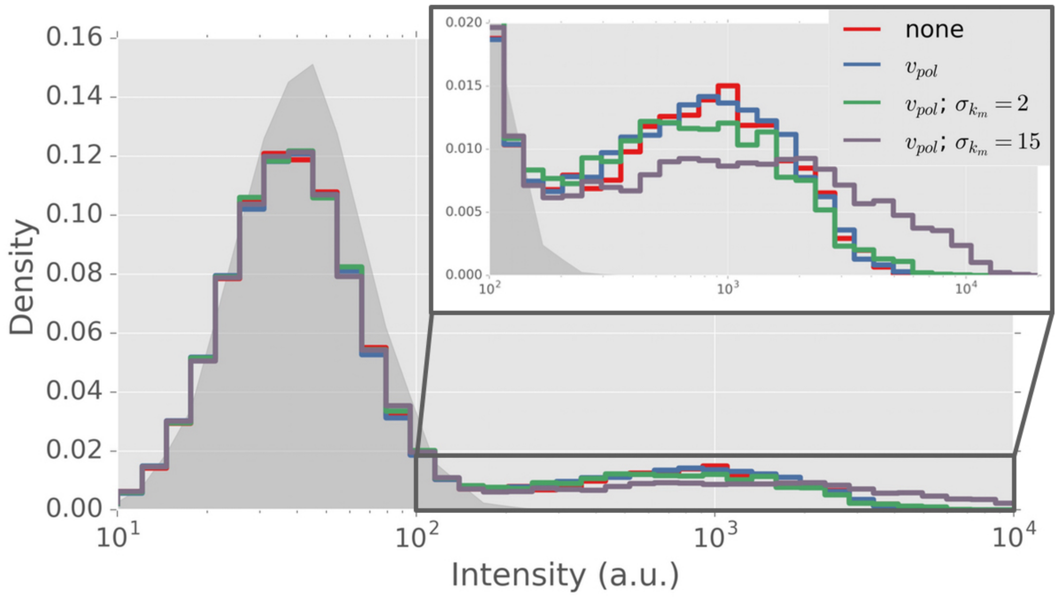

Figure 2.4.6: The type of parameter perturbation influences the global histogram. The plot shows the global histogram of the same simulations as used in figure 2.4.5. Four different perturbation conditions are shown (colored lines), no perturbation (red), perturbation of the RNA polymerase velocity alone (blue) and two combinations of perturbation of the RNA polymerase velocity and initiation rate. Perturbation strength of the initiation rate was set to a low ($\sigma_{k_m} = 2$ min^{-1}, green) and a high ($\sigma_{k_m} = 10$ min^{-1}, purple) value, respectively. The grey area denotes the distribution of the background signal. The inset enlarges the region of the histogram marked by the dark grey box.

Summary

In summary of this section it can be said that the hybrid stochastic-deterministic model is capable to qualitatively explain various features of the experimental data. The stochastic simulation alone can explain the intrinsic noise of the process of gene transcription. Varying the promoter off time allowed to explain the modulation of the global intensity histogram as observed in experimental data obtained at different experimental conditions. In addition such parameter variation seemed sufficient to explain the different fractions of cell that respond to the stimulus during the observation time. This feature appears promising in including dose dependency into a global model later. Let alone, stochastic simulation is not sufficient to fully explain cell to cell variability. Parameter perturbations as representations of sources of extrinsic noise were necessary. By this it was possible to explain variability in ACF half lives. In addition, modulating the strength of perturbations allowed for subtle but important modulation of the global intensity histogram. Calibrating the introduced model to the experimental data will yield estimations of the most likely model topology and its corresponding parameters.

2.5 Approximate Bayesian Computation for likelihood free model calibration

Calibrating the model introduced in the previous sections to experimental data consists of two tasks: model selection and parameter estimation. The former will reveal the most likely model topology and the latter the corresponding parameter values. This section will introduce a likelihood free bayesian method to calibrate stochastic models to data. In addition to parameter estimation, this method easily allows to incorporate model selection.

Incorporating the delay between gene transcription and visible fluorescence intensity under the microscope made it necessary to introduce extra states for RNA elongation into the model. This increased the systems state space substantially because of the enormous combinatorial possibilities to distribute only a small number of RNA molecules over the RNA elongation states. Hence, for a given data set D it was not possible to calculate the likelihood $L(\theta, D)$ of the parameter vector θ.

Approximate Bayesian Computation (ABC) provides a way to circumvent the calculation of a likelihood by comparing simulations with experimental data by a distance measure $\rho(D_{sim}, D_{exp})$ and thus gaining an approximate posterior distribution [Tavaré et al., 1997; Beaumont et al., 2002; Marin et al., 2012; Sunnåker et al., 2013]. In contrast to maximum likelihood model fitting a bayesian approach provides not only a point estimation of model parameters together with a confidence interval but yields a full posterior distribution of acceptable parameter estimations. In addition, ABC allows to treat the model as an extra parameter and thus include model selection into the actual fitting yielding a posterior distribution on the model as well. The simplest ABC algorithm would be a rejection algorithm where model m and a suitable parameter set θ are sampled from the prior distribution and a synthetic data set is simulated. If the distance measure ρ is below a predefined threshold ϵ the pair (m, θ) is accepted. By repeated sampling, simulating and comparing it is possible to gain a combined posterior distribution $P(m, \theta | \rho(D_{sim}, D_{exp}) \leq \epsilon)$ on the model and the corresponding parameters (see algorithm 2 in the methods sections). A computational advantage of the algorithm is that it can easily be parallelised to run on multiple processors. For a sufficiently good approximation of the posterior a small distance threshold ϵ is required making it necessary to simulate an enormous number of candidate parameter sets. A threshold set too high would result in a posterior distribution resembling the prior without revealing new insight. The actual numerical value of ϵ depends on the type of distance measure. The distance measures we used for model calibration will be explained in the next section.

2.5.1 Features of experimental data used for model calibration

A crucial step in applying ABC is the selection of a proper distance measure ρ. Experimental data sets used here consisted of about 50 time courses with 250 time points each resulting in a high dimensional distribution of measured fluorescence light intensity values. Comparing such distributions is not straightforward. To reduce the dimensionality five different metrics were combined to estimate the agreement between simulated and experimental data. Four metrics compare features of experimental data that will be introduced in the next sections and one metric compares the full multivariate distributions of experimental and simulated data. Those features were:

1. The global distribution of intensity values
2. The mean autocorrelation function
3. The distribution of ACF half lives
4. The distribution of ACF values at a lag of one
5. The maximum mean discrepancy

Global distribution of intensity values

The global intensity distribution shows a characteristic shape as shown in figure 2.3.1. This feature ignores the time course nature of the data and thus resembles a univariate distribution that easily can be compared to a simulated distribution utilising the well known Kolmogorov-Smirnov statistic. This statistic is the maximal vertical distance between the cumulative distribution functions (CDF) of the distributions to be compared, i.e. it is a number between zero and one with zero representing perfect agreement. The stability of the distribution was assessed by bootstrapping. Out of each experimental data set new data sets were created by sampling data points with replacement. For each such new data set the global CDF was calculated. The left panel in figure 2.5.1 shows the mean CDF of 1000 bootstrap data sets each for all four experimental data sets (orange lines). The ±SD error range is extremely small and thus not visible on the plots. This is due to the size of the data sets of more than 12000 data points. This allows to directly calculate the Kolmogorov-Smirnov statistic between data and simulation as the first distance metric.

Mean autocorrelation function

In contrast to the global histogram the autocorrelation function (ACF) incorporates the time course nature of the measured signal. It describes the similarity between measured intensities as a function of the time lag between them. The mean ACF characterises the the population average of this similarity. The ACF was calculated according to equation 2.3.2 using a sliding window approach. The observed time courses were 250 points long and a window of size 125 data points was slid along the time course with a step size of 5 data points resulting in 25 window positions. For each window position the ACF for each cell was calculated and then averaged over all cells. The second panel in figure 2.5.1 shows the ACF averaged over all window positions (orange lines) Grey areas denote the ±SD error region. The error range for the 5 pM and 10 pM data sets is larger because these cells show a substantial fraction of cells not responding to the estrogen stimulus. Not responding cells show an ACF immediately dropping to zero. Those cells cause the characteristic kink in the mean ACF at a lag of one. As the 20 pM and 1000 pM data sets hardly contain not responding cells this effect is less pronounced there causing extremely small error ranges. The distance between data and simulation was calculated by the sum of squared distance between the mean ACF of data and simulations.

$$d_{ACF} = \sum_{i=1}^{N} (acf_i^d - acf_i^s)^2 \qquad (2.5.1)$$

d and s denote data and simulation respectively. N is the maximum lag up to which the mean ACF should be considered. Here a value of $N = 51$ minute was chosen which corresponds to the 17th data point. Due to the small error ranges a weighting of the individual lag positions along the mean ACFs was neglected.

Distribution of ACF half lives

A main feature of cell to cell variability is the distribution of the half life of the ACF. Half life of the autocorrelation function here denotes the time at which the ACF drops from one to a value of one half and can be used as a measure of the variability in temporal behaviour among the cells of one data set. Non responding cells would all have a half life of 1.5 min corresponding to half the imaging interval. Responding cells display a much wider variability. Thus, this measure would mostly reflect variability among responding cells creating the triangular pattern of ACFs in figures 2.3.2 and 2.4.5.

For each window position of the sliding window calculation of the autocorrelation function the distribution of the ACF half lives was estimated. The third panel of figure 2.5.1 shows the cumulative distributions of ACF half lives. The orange lines again are the mean CDF over all window positions and the grey areas denote the ±SD error region. The steep initial increase in the CDF for the 5 pM and 10 pM is caused by the fraction of non responding cells. Shape of the CDFs for 20 pM and 1000 pM show smooth increase over the full range of values and thus indicate more subtle variability. The small error ranges here allowed to calculate the distance between simulation and data again by the Kolmogorov-Smirnov statistic as for the global histogram.

Distribution of ACF values at a lag of one

A second ACF based measure of cell to cell variability is the distribution of the ACF values at a lag of one. Non responding cells would show a value of zero, strongly responding cells a value close to one. Experimental conditions leading to half of the cells responding and half not would create a bimodal distribution with equally weighted peaks. Thus, this measure reflects the partitioning of the observed cells into responding and non responding ones.

The right column of figure 2.5.1 shows the mean CDF of the ACF values at a lag of one over all window positions of the ACF estimation (orange) ±SD error range (grey area). CDFs for the 20 pM and 1000 pM data set show a steep increase at high values indicating that only a small fraction of cells is not responding to the stimulus. CDFs for the 5 pM and 10 pM indicate more variability and show wider error ranges. This reflects the substantial fraction of cells in both data sets that is not responding. The estimated error ranges here a small, as well. Thus, the distances between simulations and data for this feature was calculated via the Kolmogorov-Smirnov statistic, too.

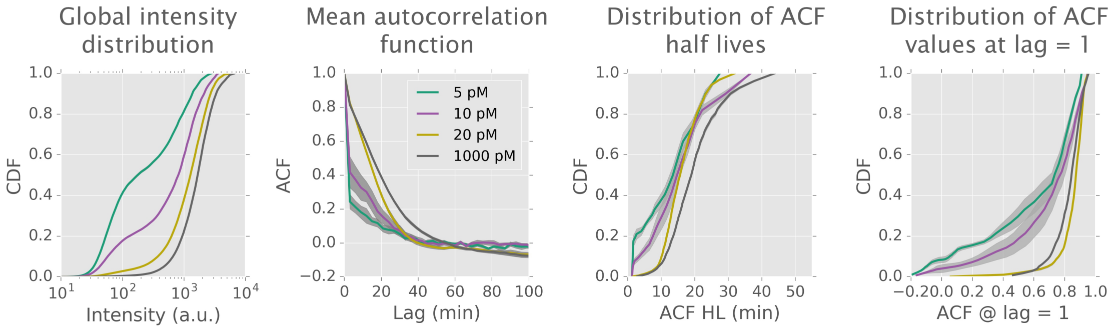

Figure 2.5.1: Data features used for model calibration. Overview over the four main features used for model calibration for all four data sets. The left panel shows the cumulative distribution functions (CDF) of the measured light intensity values. The second panel displays the mean ACF function averaged over all cells and over 25 positions of a sliding window of size 125 data points. The third panel displays the CDF of ACF half lives averaged over all obtained window positions. The last panel depict the CDF of the ACF value at a lag of one. Again this distribution was averaged over all window positions obtained by the sliding approach. See main text for explanations. Colored lines display the mean values of the different data sets as indicated by the legend and grey areas denote the ±SD region representing statistical uncertainty.

Maximum mean discrepancy

The last metric we included is the maximum mean discrepancy (MMD) which is a statistic to compare multivariate distributions [Gretton et al., 2012]. This last metric was successfully applied in an Approximate Baysian Compuation setting to fit time course measurements by Loos et al [Loos et al., 2015]. More details on the calculation of the MMD are given in the methods section.

As total distance between simulation and experimental data the sum of all five metrics was utilised. Smaller distance values represent better agreement between simulation and data. A distance value of zero, however, is highly unlikely due to the stochastic nature of the process, i.e. two simulations with the same parameters will not yield an identical outcome. Estimation of the distance of an optimal fit will be explained prior to algorithm benchmarking with synthetic data sets.

Distance measure	**Statistic**	**Interval**
Mean ACF	Sum of squared differences	$\in [0, \infty]$
Global distribution	Kolmogorov-Smirnov	$\in [0, 1]$
ACF HL	Kolmogorov-Smirnov	$\in [0, 1]$
ACF at lag = 1	Kolmogorov-Smirnov	$\in [0, 1]$
Maximum mean discrepancy	multivariate	$\in [0, 1]$

Table 2.5.1: Distance measures to compare simulations with experimental data. Distance measures to compare features of the data with simulated synthetic data sets. ACF stands for autocorrelation functions, HL for half life. For all metrics a value of zero corresponds to perfect agreement between simulation and data. As total distance the sum of all five metrics was calculated.

2.5.2 Sequential Monte Carlo Approximate Bayesian Computation

Sequential Monte Carlo Approximate Bayesian Computation (SMC ABC) provides a more efficient way to perform ABC than the simple rejection algorithm introduced above by approaching the true posterior distribution sequentially via a series of intermediate distributions [Del Moral et al., 2006; Sisson et al., 2007]. The approach implemented here follows Toni et al [Toni and Stumpf, 2009; Toni et al., 2009].

Creation of an initial population from the prior

In an initial iteration a start population of particles is sampled from the prior distributions of the model and the parameters. A particle consists of a weight, an unique model index and the corresponding parameters. For each particle a data set is simulated and compared with the experimental data. This initial round essentially is the simple rejection algorithm introduced before.

From each iteration the best 20% of particles are taken to the next round and the new population is filled up with particles that are created out of the accepted particles using proposal distributions. The particle with the largest distance within the best 20% of particles defines the threshold distance ϵ_t of the current iteration.

Creation of new particles with proposal distributions

In each iteration after the initial round the particle population has to be filled up to its initial size by creating new particles. Particle creation is done by proposing new particles in the proximity of accepted particles using proposal distributions. First a particle from the current population is selected at random based on its weight and subsequently the particle parameters including the model are changed according to the proposal distributions. For each particle a data set is simulated and compared with the experimental data. If its distance is smaller than the current maximal distance ϵ_t the particle is added to the updated population. The weight of each new particle is calculated based on the particles from the previous population and the prior probability of the particle. It is a measure for the probability of the particle to reach its position based on the positions of its predecessors combined with the prior belief of possible particle positions within parameter space.

The algorithm terminates when no improvement is gained

By subsequently iterating this algorithm approaches the true posterior by only accepting particles that are improving the distance measure with respect to the previous iteration. The algorithm terminates either when all particles are below a final stopping distance reached or when the improvement of distance between subsequent iterations is less then 5%. Details of algorithm implementation, chosen prior and proposal distributions are given in the methods section. Figure 2.5.2 depicts a graphical representation of the algorithm.

To ensure a dense initial sampling of model and parameter space a population of 50.000 candidate particles was sampled from the model and parameter prior distributions. Simulations of RNA counts were saved to file so that this population could be reused for fitting real data and algorithm benchmarking by simply adding noise according to the noise model and measure the distance to the given data. Before each SMC ABC run the 2000 best candidate particles were selected from this initial population to gain a satisfactory start population.

2.5.3 Number of free parameters and model selection

Model selection

Investigations of the number of promoter states for single cell gene transcription revealed either small (2-3) or slightly larger numbers (5-10) for different mammalian genes [Suter et al., 2011; Zoller et al., 2015]. To cover this range of states the set of models available to model selection was restricted five different topologies of the promoter cycle. Three small models consisting of 2,3 or 4 states and two large model of 10 states. The smallest

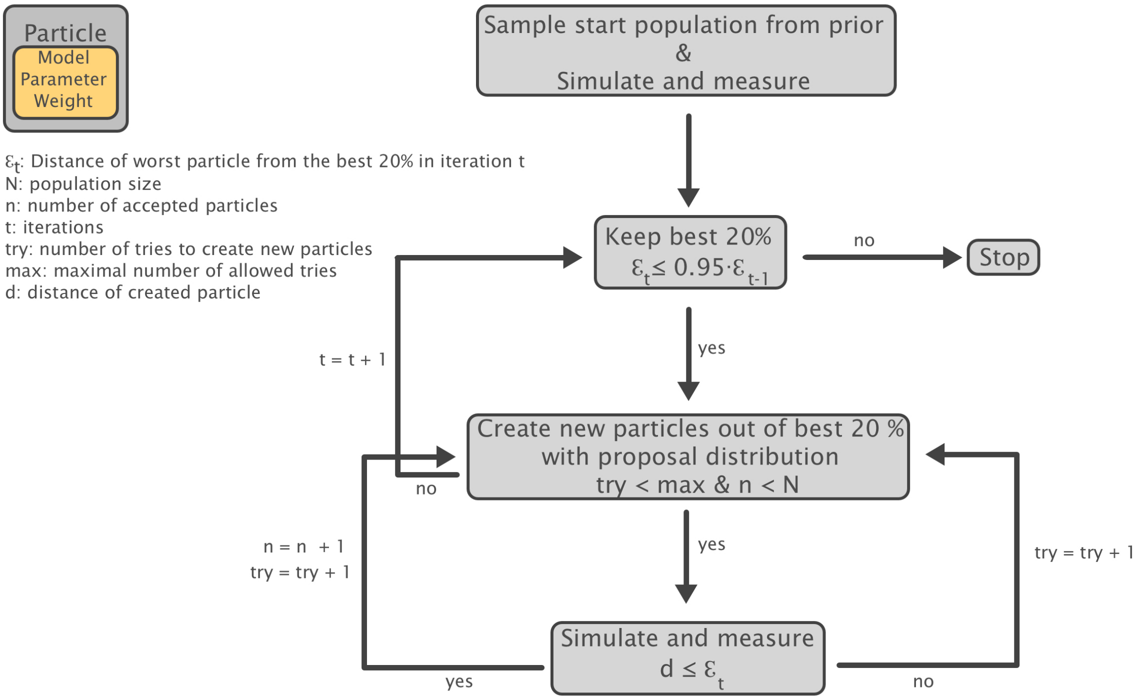
Particle
Model
Parameter
Weight
ε_t: Distance of worst particle from the best 20% in iteration t
N: population size
n: number of accepted particles
t: iterations
try: number of tries to create new particles
max: maximal number of allowed tries
d: distance of created particle
Sample start population from prior
&
Simulate and measure
Keep best 20%
ε_t ≤ 0.95·ε_t-1
no
Stop
yes
t = t + 1
Create new particles out of best 20 %
with proposal distribution
try < max & n < N
no
yes
n = n + 1
try = try + 1
try = try + 1
Simulate and measure
d ≤ ε_t
yes
no

Figure 2.5.2: Flow chart of the SMC ABC algorithm. The flow chart graphically depicts the SMC ABC algorithm described in the main text. A start population of particles is sampled from the prior distribution, simulated and measured and subsequently fed into the sequential Monte Carlo Algorithm. A particle consists of a model with corresponding parameters and its weight (see inset in the upper left). In each iteration the best 20 % of particles are kept. The worst particle of the best 20% defines the threshold distance ϵ_t of the tth iteration. Out of the best 20% of particles of the current iteration new particles are created using proposal distributions to fill up the population again. The maximal number of tries to create is limited to max. If a new particle has a distance d below ϵ_t it is added to the population until the population size reaches N again. If the maximal number of tries to create new particles is reached the population is filled up with the best particles above the threshold. The algorithm terminates when no further improvement is gained.

included model is the well known random telegraph model with only one on and one off state. A three state model has one refractory state in the off phase. In addition, a four state model has a refractory state in the on phase. As contrast to the small model two large models consisting of ten states were included, one with one and one with two on states. The set of allowed model topologies together with the corresponding number of parameters is listed in table 2.5.2.

Name	**On states**	**Off states**	**# Parameters**
Random telegraph	1	1	3
Three states	1	2	4
Four states	2	2	5
Ten states, one	1	9	11
Ten states, two	1	8	11

Table 2.5.2: Model topologies considered in model selection and their number of free parameters. The five different model topologies listed here together with their number of parameters were used during model selection. Each state is associated with a transition rate, thus the number of parameters is defined by the total number of promoter states plus the initiation rate as additional parameter.

Free parameters

Parameters describing RNA elongation were known from literature as explained earlier. Parameters to estimate by model fitting are the parameters of the promoter cycle. The waiting time the promoter spends in a single state follows an exponential distribution with the inverse of the transition rate denoting the mean life time of the corresponding state. The average total time spend in either the active (t_{on}) or inactive (T) phase is the sum of the individual life times of all states in the respective phase. While in an active state the promoter can recruit RNA polymerase which in turn creates new transcripts with the rate k_m. Total on time t_{on}, total off time T and initiation rate k_m are main parameters of the model shared by all model topologies and that were estimated by model fitting. Models with more than two states have extra parameters for the life times of the individual states within either on or off phase.

If a model topology has more than one state in either the on or off phase it will not be possible to find the correct order of these states due to the symmetry of the cycle. If for example the states within the off phase will be permuted the total off time will remain the same. The distribution of this total off time is given by the convolution of

index	**Perturbation**	
0	none	
1	RNA pol II speed	v_{pol}
2	Initiation rate	σ_{k_m}
3	On time	$\sigma_{t_{on}}$
4	Off time	σ_T
5	RNA pol II speed and initiation rate	v_{pol} and σ_{k_m}
6	RNA pol II speed and on time	v_{pol} and $\sigma_{t_{on}}$
7	RNA pol II speed and off time	v_{pol} and σ_T

Table 2.5.3: Parameter perturbations to model cell to cell variability. Listed are all eight different parameter perturbations that were included into model selection. The left column denotes the perturbation index that later will be referred to to identify specific models. v_{pol} denotes the velocity of the RNA polymerase as it transcribes the gene, σ_{k_m}, $\sigma_{t_{on}}$ and σ_T denote the standard deviation of the normal distributions where the perturbed initiation rate, on time and off time were sampled from. The latter three are extra parameters that were included into model fitting.

the exponential waiting time distributions of the individual states. Due to the symmetry of the convolution operation a permutation of the states would result in the same final distribution. This fact was reflected in model parameterisation by assuming an ordering of the waiting times within one cycle phase from long to short. For details of model parameterisation please see the methods section.

Cell to cell variability is included into the model via eight different possible perturbations on the model parameters (see table 2.5.3). Perturbations were implemented by resampling the parameter from a normal distribution for each simulated cell. The means of those normal distributions were the parameter values set as input for the simulation of the data set. The Standard deviation, i.e. the strength, of the perturbation distribution was included as a free parameter into model fitting. Thus, all perturbations, except of the RNA polymerase velocity alone increased the number of model parameters by one. Those extra parameters were named σ_{k_m}, $\sigma_{t_{on}}$ and σ_T for perturbation of the initiation rate, the promoter on and off time, respectively.

All eight different parameter perturbations could be combined with all five considered model topologies listed in table 2.5.2 resulting in a set of 40 different models available to model selection. As explained earlier model selection was included into the SMC ABC algorithm which provided a posterior distribution of the frequencies of all accepted models after the algorithm came to halt. The algorithm selects the model and corresponding parameters on their ability to explain the data, i.e. generate a low distance value. Models of higher complexity are not penalised explicitly.

2.5.4 Algorithm benchmarking

Models and parameter sets for benchmarling

To assess the ability of the SMC ABC approach to recover the correct model and parameter values different synthetic data sets where created. The full list of models used for creation of synthetic data sets is listed in table 2.5.4. Various different model topologies were selected to test the model selection capabilities of the method. Model parameters as promoter on and off times and the initiation rate were chosen according to published values [Suter et al., 2011; Zoller et al., 2015].

As discussed earlier, modulation of promoter off time provides a promising way to explain the observed dose dependency of the experimental data. To test the sensitivity of parameter estimation off time values of 30,150,300 and 700 minutes were used. 700 minutes off time are close to the experimental observation time of 750 minutes and thus observed off times of individual cells might be larger than this simply by intrinsic variability. This corresponds to cells not responding to the estrogen stimulus during the observation time. An off time of 30 minutes is in the range of transcript dwell time of approximately 35 minutes making it difficult to distinguish consecutive transcriptional bursts. In addition to promoter off time the on time was varied for benchmarking as well. An on time of two minutes as used in the last model of table 2.5.4 is below the experimental imaging interval of three minutes and will test the algorithm in that direction.

Simulations created by the models listed in table 2.5.4 showed high similarity with experimental data. Figure 2.5.3 shows the global histogram and the mean ACF for five of the ten models from table 2.5.4. All data sets show a bimodal distribution of their intensity values except model 115 700. Similar to the real data set obtained at 5 pM stimulus only little intensity lies above the background (grey curve). For the 115 30 model the right peak in the histogram is more pronounced than the left one similar to the data set acquired at 1000 pM stimulus (pink curve). In agreement with the 5 pM data set synthetic data for the 115 700 exhibits a mean autocorrelation function with a characteristic kink indicating a substantial number of cells that are not responding. The autocorrelation of the 195 model is qualitatively different in that it drops considerably

Model	#On	#Off	Perturbation	Initiation	On time	Off time
111	1	1	1	7 min^{-1}	10 min	70 min
121	1	2	1	7 min^{-1}	10 min	70 min
125	1	2	5	7 min^{-1}	10 min	70 min
225	2	2	5	7 min^{-1}	10 min	70 min
195	1	9	5	7 min^{-1}	10 min	70 min
115 30	1	1	5	7 min^{-1}	20 min	30 min
115 150	1	1	5	7 min^{-1}	10 min	150 min
115 300	1	1	5	7 min^{-1}	10 min	300 min
115 700	1	1	5	7 min^{-1}	5 min	700 min
115 2 30	1	1	5	7 min^{-1}	2 min	30 min

Table 2.5.4: Models and parameter sets used to create synthetic data sets for algorithm benchmarking. The left column denotes the name of the model that will be referred to later. Perturbations are defined according to table 2.5.3. An index of 1 indicates a perturbation of the RNA polymerase velocity and an index of 5 denotes a combined perturbation of initiation rate and RNA polymerase velocity. In the first five models the topology is varied in the last five models the on and off times.

below zero and increases back to zero afterwards. This indicates a more pronounced regularity in promoter switching between on and off states than for the smaller models. All displayed data sets, however, show some agreement to the actual experimental data justifying their use as benchmarking data sets.

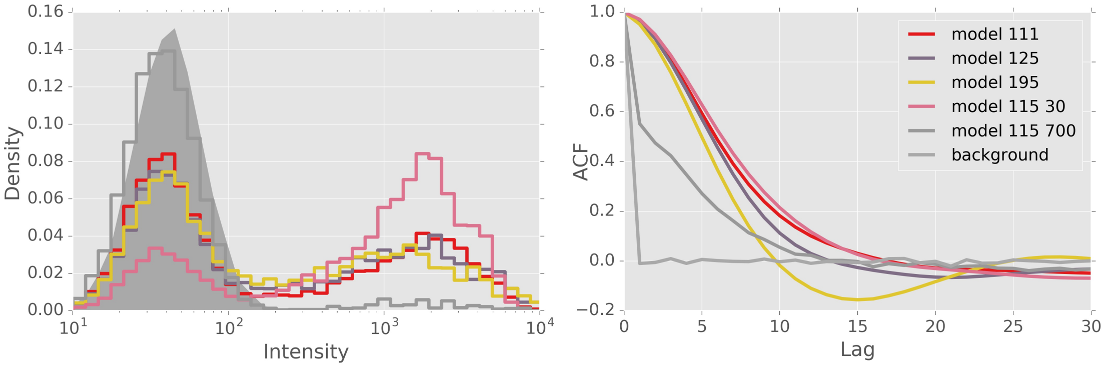

Figure 2.5.3: Benchmark data sets are similar to experimental data. The left panel shows the total distributions of five simulated data sets created with models from table 2.5.4. Each data set consists of 50 cells simulated each for 750 minutes resulting in data sets of size similar to experimental data sets. In dark grey the simulated background signal is shown. The right panel shows the corresponding mean auto correlation functions.

Distance of an optimal fit

Before running actual data fits it was necessary to define what value of the distance measure characterises a good fit. Due to the intrinsic variability of the stochastic simulations not any two data sets created with the same model and parameters would match exactly. To assess the characteristic self distance 200 data sets were simulated for three different models from table 2.5.4. From that populations random pairs were selected and their mutual distance was calculated. Box plots of the distributions of the so gained distance measures are plotted in figure 2.5.4. From these distributions a stop criterion of the SMC ABC algorithm of 0.5 was defined, i.e. if all particles of the population are below that distance the algorithm terminates.

Model fits can explain the benchmarking data sets

As explained in the previous section for each data set the 2000 best matching particles were selected from the large start population of 50.000 particles and fed into the SMC ABC algorithm. Figure 2.5.5 shows the distributions of the final distance after multiple iterations when the algorithm came to halt. The dashed grey line indicates the distance stop criterion estimated from figure 2.5.4. The majority of particles for most data sets lies below that threshold but not the full population. To asses the minimal distance that could be reached for different data sets the algorithm was run until it terminated

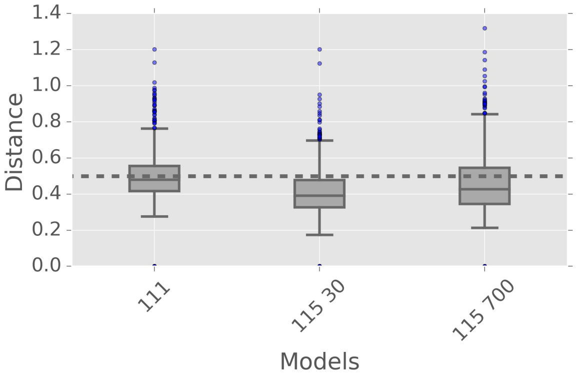

Figure 2.5.4: Distributions of mutual distances of data sets simulated with the same model. Distributions of distance measures gained by repeated simulation of data sets with the model listed in the x-axis label. Each of the three models was simulated 200 times. The distance between 1000 random pairs of data sets was calculated yielding the plotted distributions. The dashed grey line marks a distance value of 0.5 which was chosen as final stopping distance of the SMC ABC algorithm.

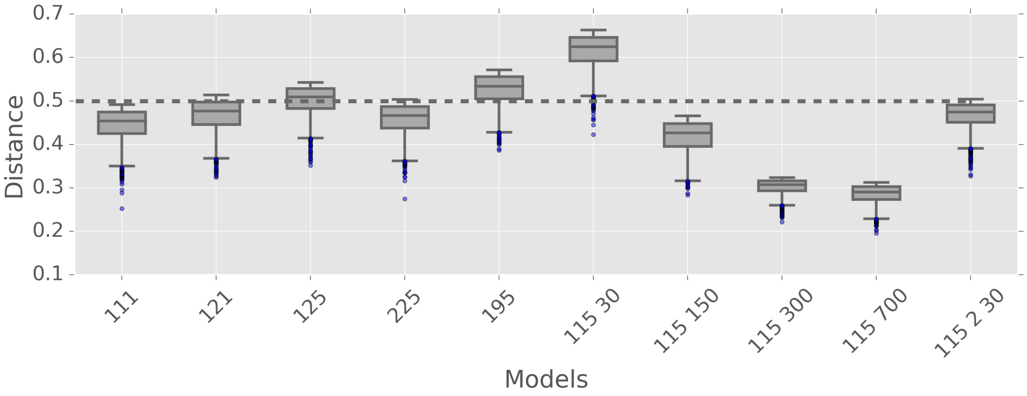

Figure 2.5.5: Model fitting yields good distance values. Box plots of the distance distributions of the final particle populations after termination of SMC ABC algorithm for all benchmark data sets listed in table 2.5.4. The dashed grey line denotes the stopping criterion set based upon figure 2.5.4.

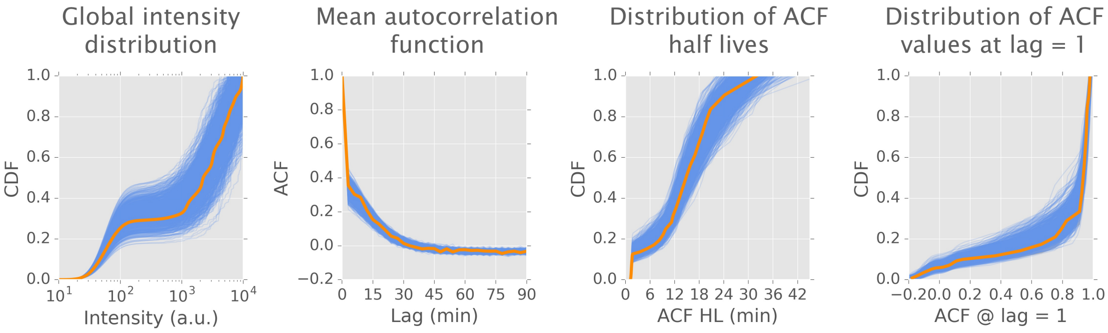

Figure 2.5.6: SMC ABC allows to fit main features of benchmarking data sets. Best fitting particles to the synthetic 115 700 data set. Each panel shows the comparison between one feature of the synthetic data set (thick orange lines) and the 1000 best particles (thin blue lines) after SMC ABC model fitting. The first panel shows the cumulative distribution of simulated intensity values. The second panel shows the mean autocorrelation, the third panel the cumulative distribution of the autocorrelation half life and the fourth panel the cumulative distribution of autocorrelation values at lag = 1.

because of lack of improvement in distance and not when it reached the stopping distance. Fitting the data set created with model 115 30 led to the largest final distances. This indicates that data sets with cells exhibiting more frequent transcription and high light intensity values seem to be more difficult to fit. In fitting real data this applies to the 1000 pM data set as will be shown later. Iterating further might eventually lead to a full population below the stop criterion but would require extremely long computation times. Fitting data sets stemming from models with longer off times in general led to smaller distance measures. This might be mostly due to the fact that the resulting distributions are less complex with most values lying in the range of the background intensity. After inspecting the final distances and the fitted lines one can conclude that the SMC ABC algorithm is able to find sufficiently good fits to data showing a close match in the main features of the data.

As an example figure 2.5.6 shows main features of simulations from the 1000 best particles (thin blue lines) of the fit to the 115 700 data set (thick orange lines). Each panel highlights one feature that was used as a distance measure. All four features could be fitted closely allowing to investigate the results further for posterior distributions of the model topologies and the kinetic parameters.

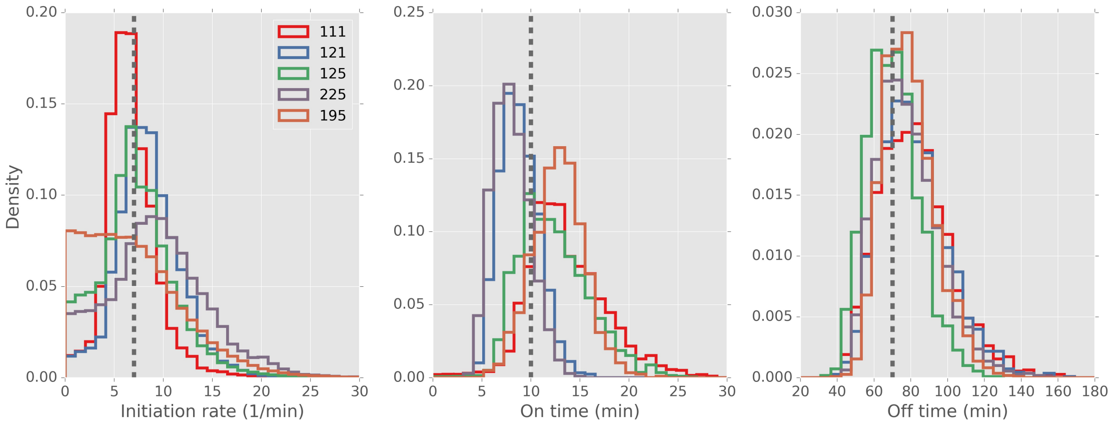

Figure 2.5.7: SMC ABC can recover true parameter values Posterior distributions of the main parameters initiation rate (left), on time (center) and off time (right) for the first five models of table 2.5.4. The colored lines denote the distributions of the recovered posterior values after SMC ABC model fitting. Dashed lines denote true parameter values.

In the next three sections the posterior distributions are going to be investigated in more detail. First the recovery of true parameter values and second the model selection will be evaluated.

Model fitting can recover the true parameter values

Figure 2.5.7 shows posterior distributions of the three main parameters initiation rate k_m (left), on time t_{on} (center) and off time T (right) for the data sets stemming from the models 111, 121, 125, 225 and 195 (colored lines) from table 2.5.4. Dashed lines indicate true parameter values. It can be clearly seen that recovery of the main parameter values is possible with the SMC ABC algorithm. Promoter on time posteriors show light under (121, 225) or over (125, 195) estimations of the true value. The mode of the posterior, however, still is located closely to the true value.

Promoter on and off times estimated directly from the experimental data indicated dose dependency (see figure 2.3.3). This suggested to test the SMC ABC algorithm towards its ability to resolve varying parameter values. Figure 2.5.8 shows posterior distributions of the varied parameters of promoter on (top) and off (bottom) time. Bins containing the true value are marked by grey rectangles. Estimating an on time of two minutes which is below the experimental time resolution of three minutes is possible. Differences for the promoter off time could be recovered accurately over a wide range

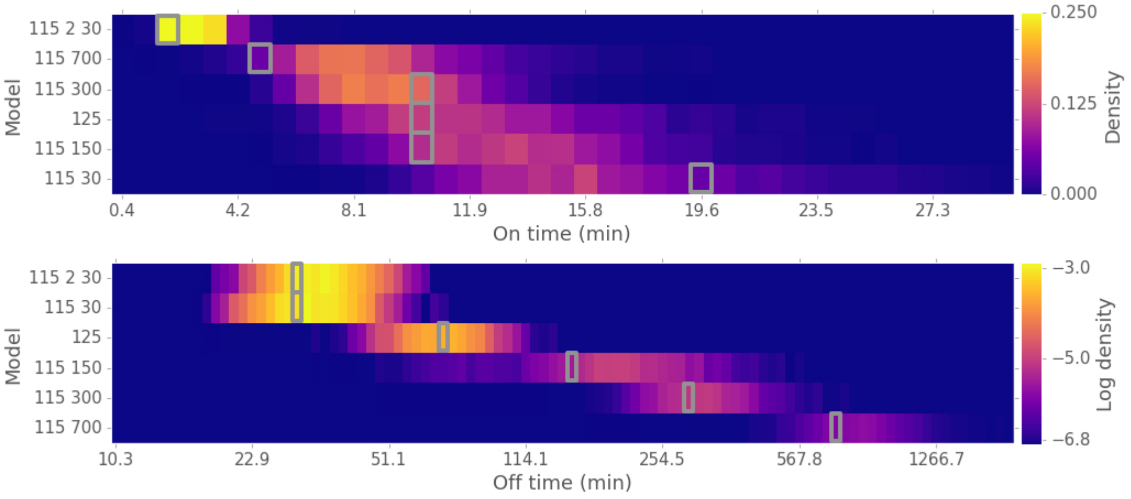

Figure 2.5.8: SMC ABC can resolve varying on and off times. Color maps show posterior distributions of promoter on time t_{on} (top) and off time T (bottom). Colors display the log density of the found posteriors with purple indicating low and yellow high density. Bins containing the true value are marked by grey rectangles.

of values up to the extreme of the experimental observation time. Both results render model fitting possible for a wide range of experimental conditions and allow to detect parameters that are modulated by the experimental stimulus.

Model selection can recover the correct model topology

In total 40 different models were available to the fitting algorithm, five topologies of the promoter cycle and eight different parameter perturbations (see tables 2.5.2 and 2.5.3). Figure 2.5.9 shows color maps of the model posterior distributions. The left panel shows a colormap of the full model posteriors for each synthetic data set (columns). On the y-axis all models are labelled using the same name convention as before by indicating by three integers the number of on states, off states and the perturbation index, respectively. True models are marked by grey rectangles. In addition to figure 2.5.9 table 2.5.5 displays the three most frequent models together with their frequency for each model fit. In general it can be said that the algorithm tends to find the correct size of the model. It is, however, difficult to distinguish models lying close together in model space, e.g distinguishing the 111 model from the 121 model. When fitting data from the 111 model the correct model was found most frequently in the model posterior distribution with 35%. Model 121, however, was ranked second with a frequency of 18% yielding a Bayes factor of approximately two.

The Bayes factor B_{12} provides a measure of how much more likely one model m_1 is over a second model m_2 by comparing their ratio of posterior frequencies, i.e. how often the respective models appear in the final SMC ABC particle population. In general the Bayes factor calculates as:

$$B_{12} = \frac{P(m_1|D)}{P(m_2|D)} \frac{\pi(m_1}{\pi(m_2} \tag{2.5.2}$$

The prior on the model $\pi(m)$ is uniform and hence the second ratio cancels leaving only the ratio of the posteriors. A Bayes Factor larger than three indicates a positive selection towards m_1 [Kass and Raftery, 1995]. The Bayes Factors here are mostly smaller than three indicating no clear favourite model as explained above for the models 111 versus 121 on the 111 data set. For the data stemming from the 195 model a large model was found but the true model was ranked at position two with only 17.4 % occurrence while 285 being the most frequent model with 55.3 %. This results in a Bayes Factor of ≈ 3 which is considered as model selection indicator favouring 285 over 195. This again indicates the difficulty to distinguish models which are closely related. The overall tendency, however, is that the algorithm is able to estimate the model size correctly.

The panels on the right hand side of figure 2.5.9 display the posterior distributions of different features of the model, i.e. the number of on (top) and off (center) states and the perturbation index (bottom). This allows to investigate which of the model structure features like number of on or off states or the parameter perturbation were not found correctly. The number of off states was correctly found most often. For larger models it was difficult to find the correct number of on states. An interesting observation was that for data sets created with a short off time of 30 minutes the model topology was found correctly but the perturbation was not. Fitting data sets created with the models 115 30 and 115 2 30 resulted in a found perturbation index of 1 instead of 5. This might indicate that for short off times perturbation of the initiation rate might not have an significant influence any more and perturbation of the RNA velocity is dominant. This effect was observed for model fits to data sets obtained at high estrogen concentrations.

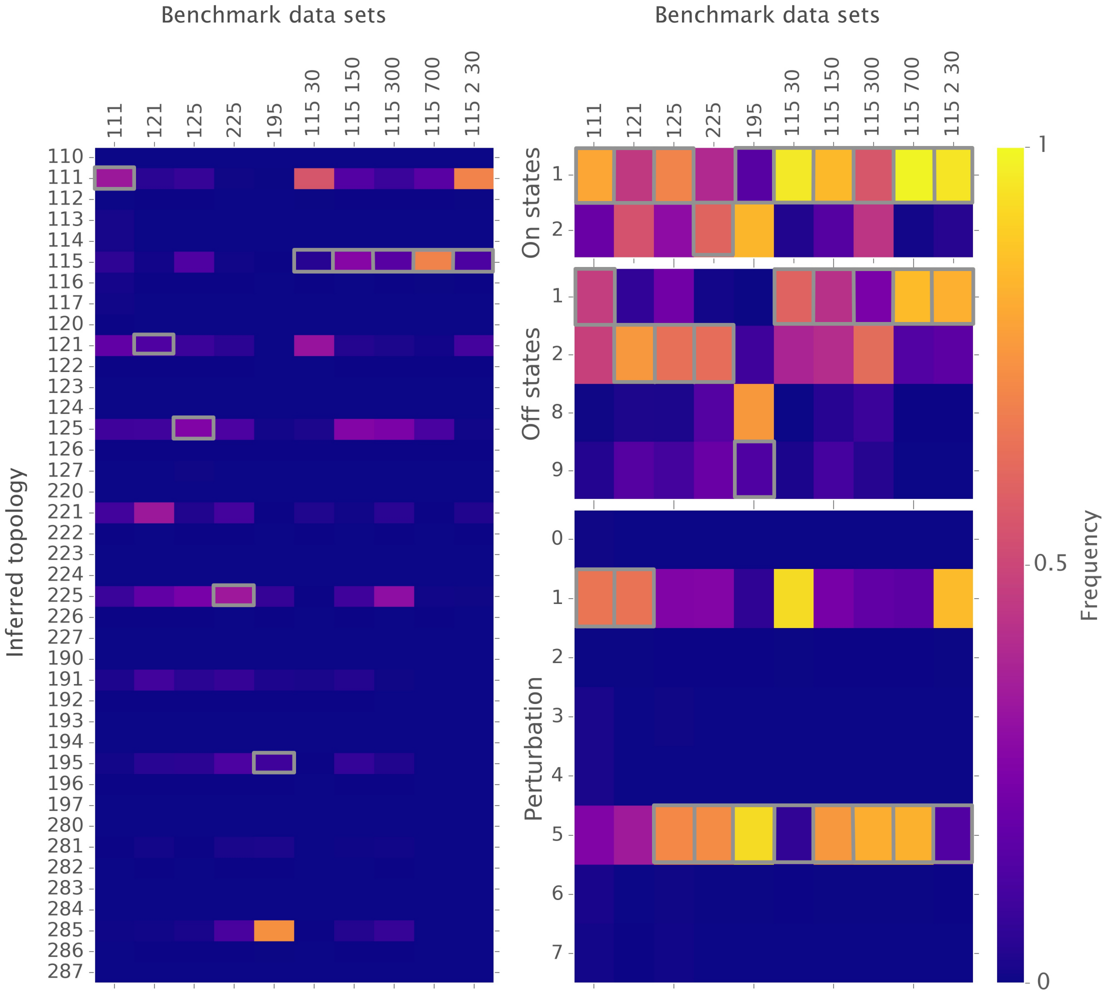

Figure 2.5.9: SMC ABC is able to estimate the correct model topology. Colormaps visualising model selection of the SMC ABC algorithm applied to synthetic test data sets. The left panel shows the frequency with which of 40 models was chosen in the posterior distribution. Each row represents one model denoted by the three digit number on the left where the first number represents the number of on states, the second the number of off states and the third the perturbation index according to table 2.5.3. The panels on the right display the frequency of how often each feature of the model was selected, i.e. how often one on state was selected for the 111 data set. The upper right panel displays the frequency of occurrence of either one or two on states. The middle right panel shows the frequencies of the allowed numbers of of states and the bottom panel the perturbation index. Brighter colors denote higher frequencies. Grey rectangles mark true models.

Model	Most frequent	2nd most frequent	3rd most frequent
111	[111], 34.3%	[121], 18.1%	[221], 10.5%
121	[221], 34.4%	[225], 17.7%	[121], 14.0%
125	[125], 26.3%	[225], 24.1%	[115], 13.8%
225	[225], 34.3%	[125], 13.8%	[195], 12.6%
195	[285], 74.5%	[195], 10.2%	[225], 7.5%
115 30	[111], 55.5%	[121], 32.5%	[115], 4.5%
115 150	[115], 27.6%	[125], 27.3%	[111], 14.6%
115 300	[225], 30.3%	[125], 25.0%	[115], 15.4%
115 700	[115], 70.1%	[111], 15.7%	[125], 12.2%
113 2 30	[111], 70.1%	[115], 12.7%	[121], 11.2

Table 2.5.5: Model selection for fitting benchmark data sets recovers the correct model topology. Most frequent models selected during benchmark SMC ABC model fitting. In the left column the names of the models that were used to generate synthetic data sets are given (see table 2.5.4). The following columns show the three most frequent models in square brackets followed by their frequency in percent.

Summary

In summary of the benchmarking of the SMC ABC algorithm it can be said that it is able to reveal both the model topology and model parameters for the given problem of fitting a stochastic model to experimental data. Synthetic data sets that were used for benchmarking showed various features of experimental data. Model selection within the algorithm was able to coarsely recover the model topology and parameter perturbation. Parameter recovery for all given data sets was precise and able to resolve parameter modulations. Thus, the SMC ABC algorithm appears as a useful method to fit the introduced stochastic model to experimental data.

2.6 Model fitting to individual data sets

Benchmarking of the model fitting SMC ABC algorithm proved its ability to both recover the correct parameter values and model topology. The algorithm was applied to experimental data the same way as to the synthetic benchmark data sets. The best 2000 particles were selected from the large start population for each data set and subsequently fed into the SMC ABC algorithm. Analysis of fitting results follows the same layout as for the benchmarking fits. First, the goodness of fit second, parameter posterior distributions and third, model selection will be assessed.

2.6.1 Model fitting can explain experimental data

Figure 2.6.8 shows the distributions of the distance measures after SMC ABC algorithm convergence for the four investigated data sets. The first five box plots in each panel show the values for the five individual metrics used to estimate the agreement between data and fit. The right most box plots show the sum of the individual metrics. The horizontal dashed grey lines denote the criterion for a good fit as estimated by the mutual distances of simulations created with the same model (see figure 2.5.4). The 5 pM data set could be fit best with the full population of particles having distances lower than the stop criterion. For the other three data sets the algorithm terminated due to a lack of improvement in distance with the highest distances between 0.9 and 1.2 for the 1000 pM data set. Fits for both 10 pM and 20 pM data sets reached distances below 0.6 being close to an optimal fit.

The pattern of the individual distance metrics is similar between the 5 pM, 10 pM and 20 pM data sets with the distance in the mean ACF being the lowest individual metric and the other four slightly larger. For the 1000 pM data set the pattern changes with the distance in the global histogram in a comparable range as for the other three data sets. The remaining four metrics are all higher than for the other data sets. Especially the difference in the distance of the mean ACF indicates that the applied model appears to show systematic deviations from the data and thus revealing the 1000 pM data set as special among the four investigated data sets.

In Figure 2.6.2 the matching of the best 1000 particles of each individual fit with the experimental data is shown for the four main features used for model fitting. Features of experimental data are represented as in figure 2.5.1 with thick orange lines representing the mean and grey regions the ±SD error region. Each of the 1000 fits is plotted as as thin blue line. All main features of the 5 pM data set can be explained well by the model (top row) with small deviations only for the global intensity distribution. Compared with the 5 pM data set the fits to the 10 pM data set show minor deviations in the global intensity distribution and the distribution of the ACF values at a lag of one. For the 20 pM data set the global intensity distribution can be explained well and deviations are visible in the distribution of the ACF half lives and the ACF values at a lag of one. The autocorrelation function of the 1000pM data set shows a slower decay than for the other data sets making it difficult for the models to explain it correctly. This leads to more pronounced deviations for the distributions of the ACF half lives and the ACF values at

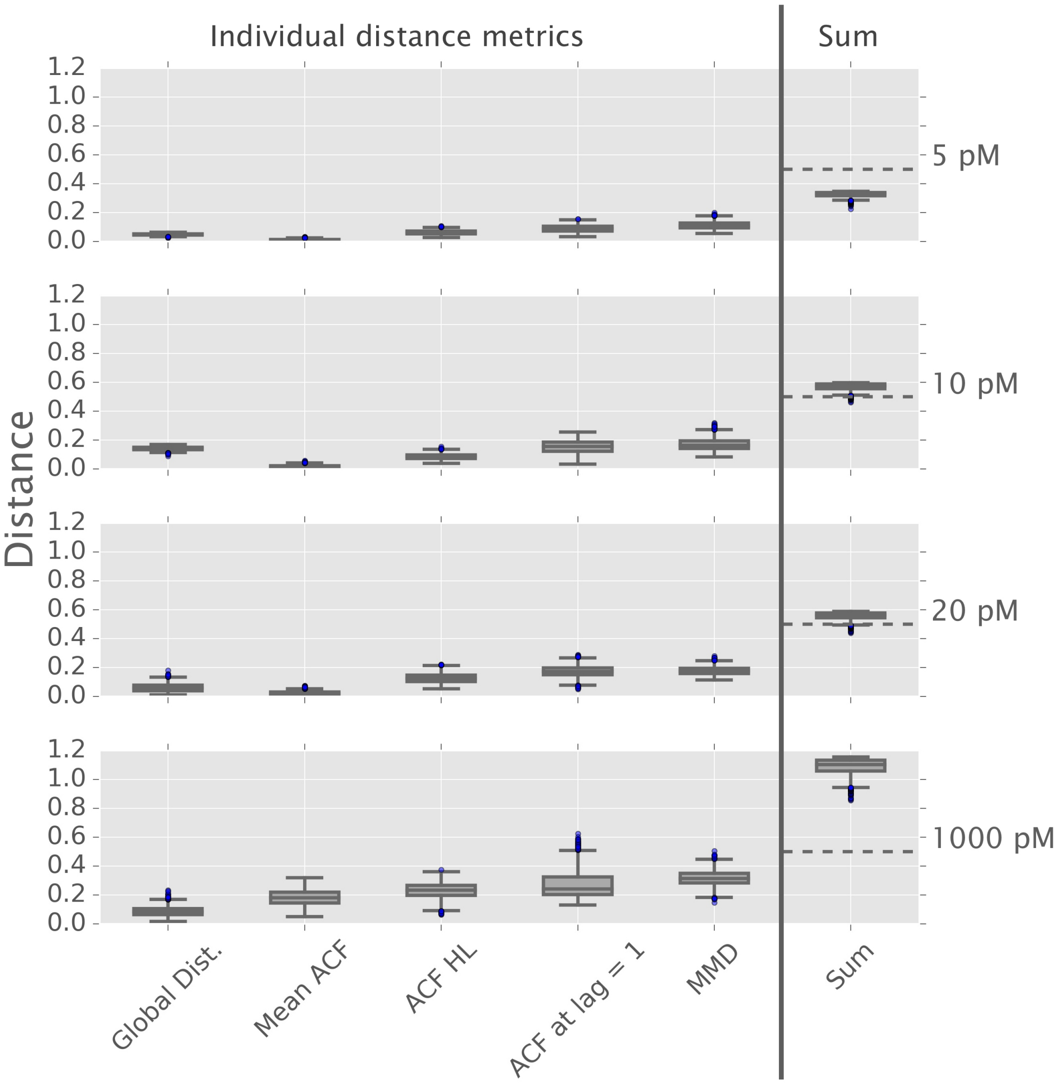

Figure 2.6.1: Final distances of fits to experimental data. Box plots show the distribution of distances of particles within the final population of the SMC ABC algorithm to different data sets. The first five box plots in each panel show the distance values for the five different metrics that where used. The last box plot shows the sum of the five metrics. The dashed grey lines indicate the distance value of 0.5 for an optimal fit as estimated in figure 2.5.4.

lag one. It can, however, be said that the overall quality of the model fits is good and thus the observed deviations appear as acceptable.

Figure 2.6.3 highlights the similarity of the experimental data and the simulations of the best particle obtained by model fitting for each data set. The representation of the data is the same as in figure 2.3.2 for the experimental data. Fitting the 5 pM and 10 pM yields correct amounts of cells that are not responding to the estrogen signal. The variability among the responding cells is similar to experimental data as well. Autocorrelation functions of simulated cells show the characteristic triangular pattern in the lower left corner.

2.6.2 Promoter off time varies with estrogen stimulus

After assessing goodness of fit in the previous section now the posterior distributions of the main model parameters will be analysed. Figure 2.6.4 shows posterior distributions of the initiation rate (top) on time (center) and off time (bottom) as color maps. The first apparent result is the clear dose dependency of the promoter off time. It varies between 800 minutes at low stimuli and less than 20 minutes at the highest stimulus. The other two displayed parameters do not show such an obvious trend and posterior distributions show significant overlap for different experimental conditions.

The burst size as the product of promoter on time and initiation rate describes the number of RNAs that is produced during during one active phase of the promoter. For other genes the burst size was found to be regulated by external stimuli [Molina et al., 2013]. Here we found no such regulation as indicated in figure 2.6.5. All four distributions show a wide overlap. The higher variability in burst size at 5 pM and 10 pM is due to the wider distributions of the initiation rate found for these data sets. This observation suggests that the transcription initiation rate and promoter on time do not vary significantly for different experimental conditions. Overall transcriptional activity appears to be controlled via the modulation of the promoter off time that shows a strong dose dependency.

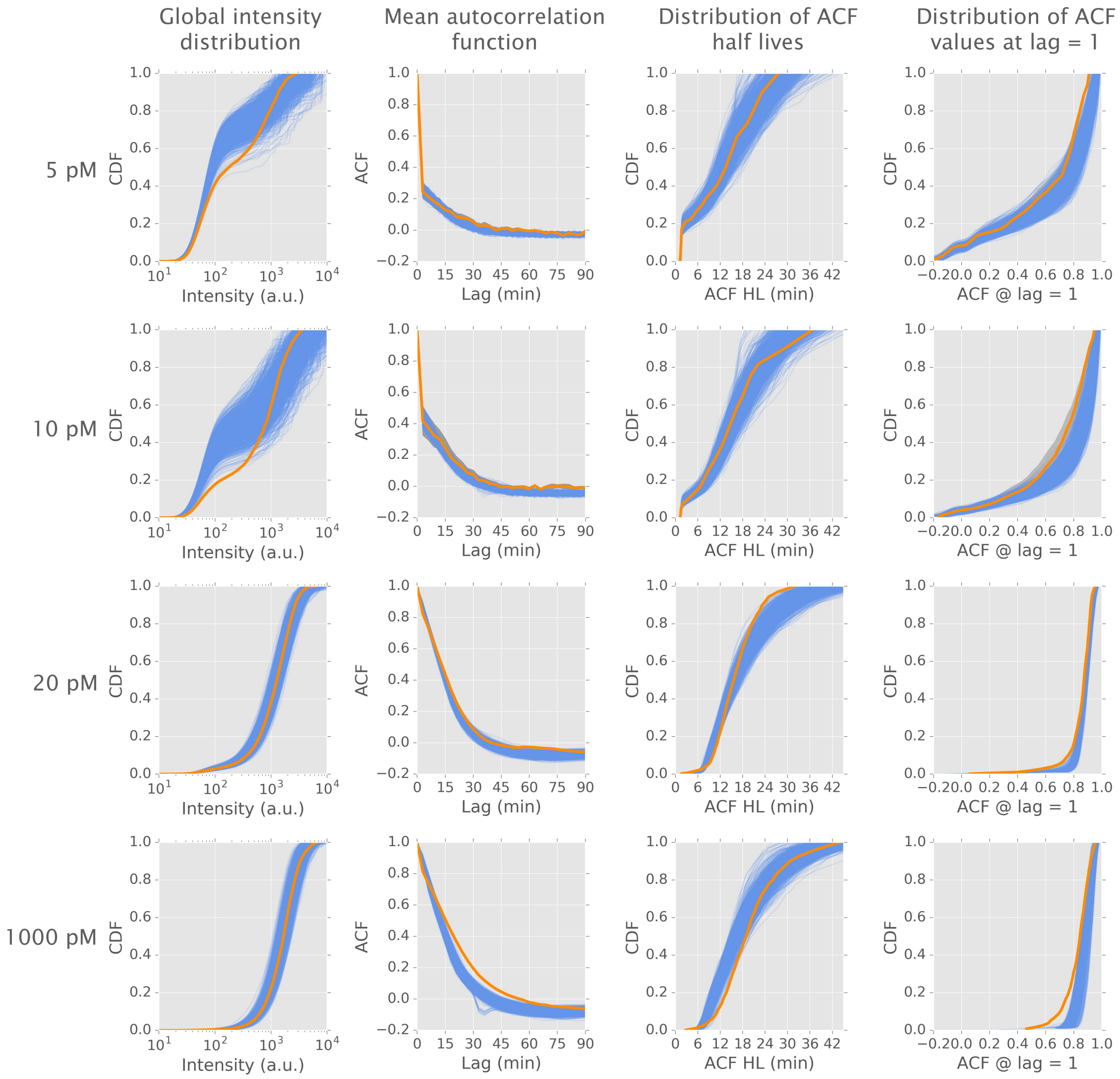

Figure 2.6.2: Model fitting allows to explain features of experimental data. The figure shows agreement of the 1000 best particles with the main features of the data. Each row of panels corresponds to the fit of one data set, from top to bottom: 5 pM, 10 pM, 20 pM and 1000 pM. The columns represent main features of the data, from left to right: the global intensity distribution, the mean autocorrelation function, the distribution of ACF half lives and the distribution of the ACF values at lag = 1. Distributions are represented by their cumulative distribution function (CDF). Features of the data are calculated according to the explanation given in the section about the SMC ABC algorithm and are plotted as in figure 2.5.1. Thick orange lines represent the mean values and grey areas denote the ±SD error region. Thin blue lines represent fits of individual particles from the final population of the SMC ABC algorithm.

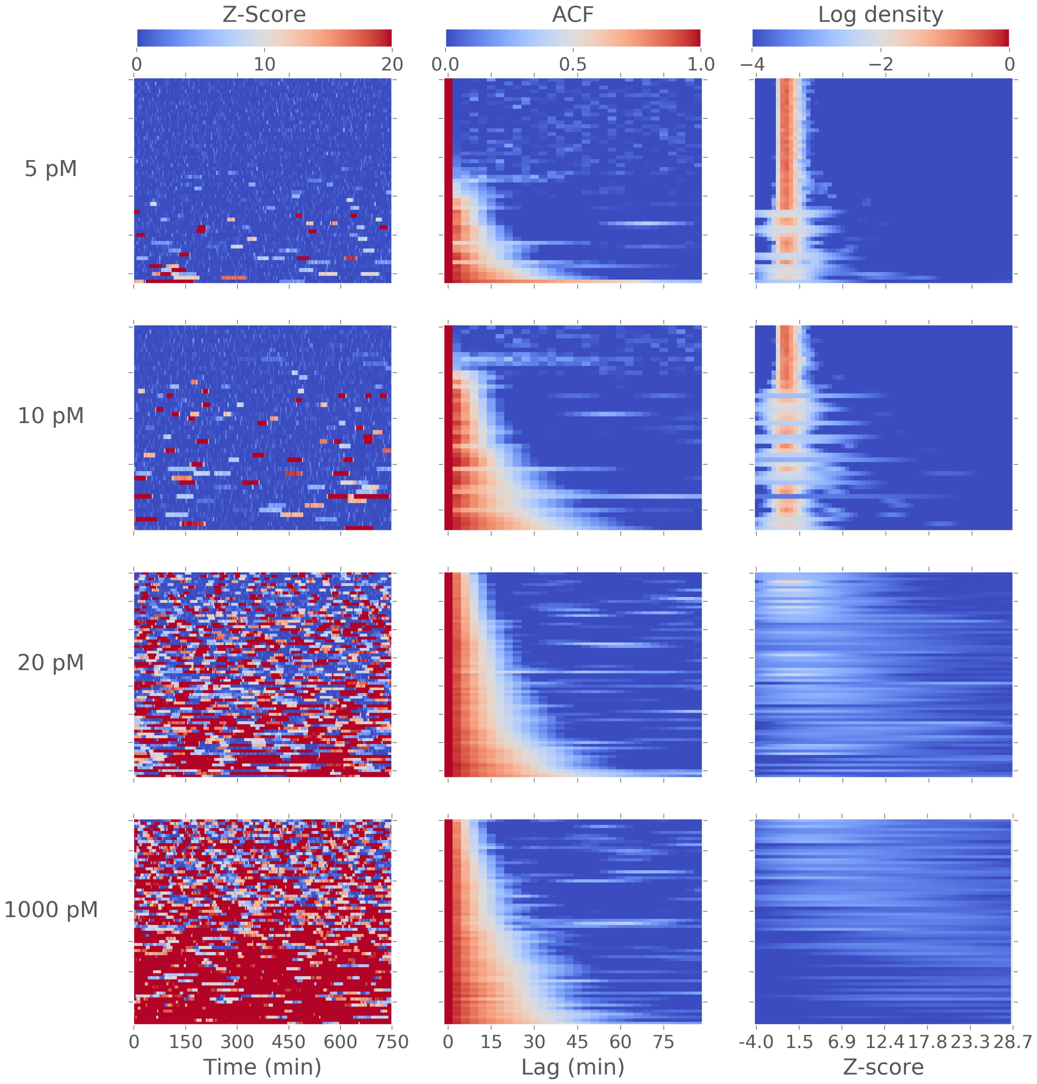

Figure 2.6.3: Model fits show close similarity to dynamic features of experimental data. Color maps to visualise dynamic features of the best fitting particle for each data set. Data representation is the same as in figure 2.3.2. The left column shows the simulations rescaled to z-scores. The middle column shows ACF of individual simulated cells and the right column shows the distribution of the z-score values of single cells obtained by a kernel density estimation. In each color map a pixel row represents a simulated single cell and cells within each data set are sorted by their ACF half lives from short (top) to long (bottom).

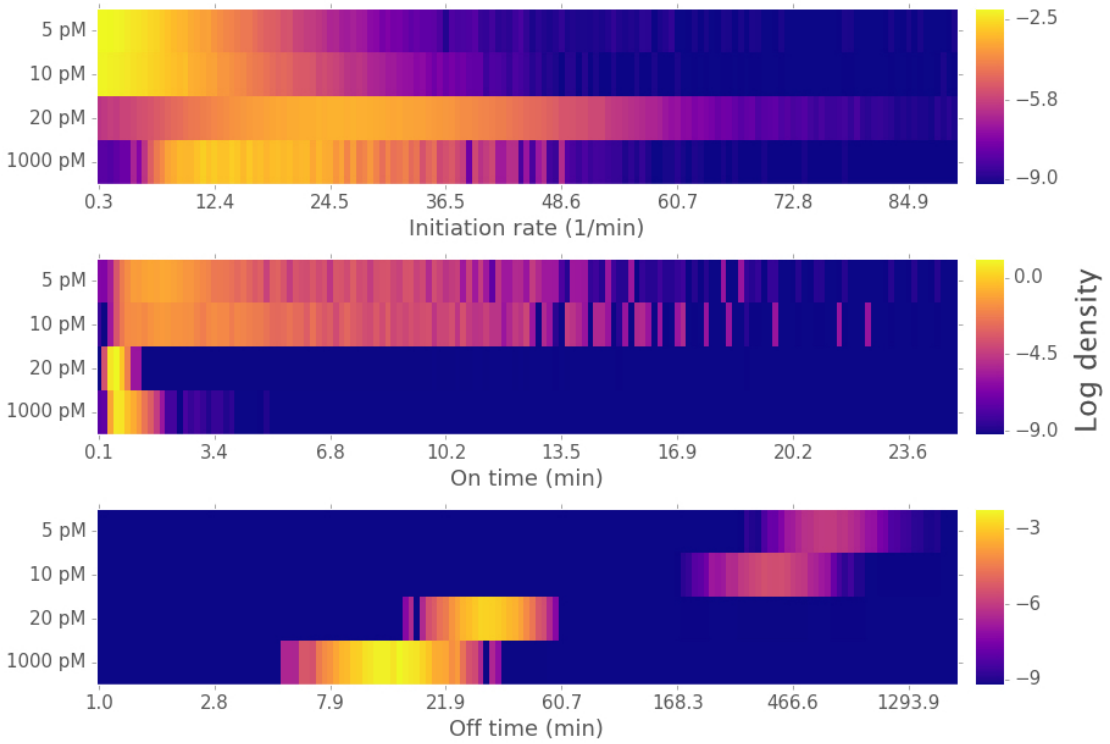

Figure 2.6.4: Promoter off time shows a clear dose dependency. Colormaps displaying the posterior distributions of the main parameters initiation rate (top), on time (center) and off time (bottom). Rows in each panel represent the results for an experimental condition denoted on the left. Brighter colors represent higher posterior densities, displayed here as the logarithm of the estimated density. Color bars indicate numerical values. Binning of the promoter off time is logarithmically and linearly for the other two parameters.

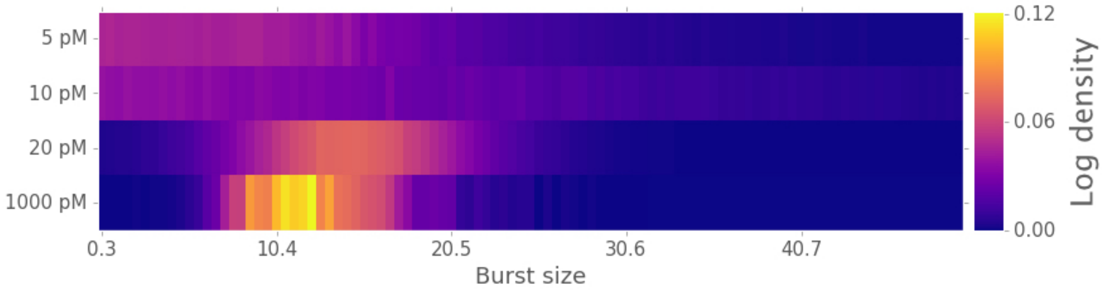

Figure 2.6.5: Burst size is not dose dependent. Colormap displaying histograms of the burst size $b = k_m t_{on}$ for the observed experimental conditions denoted on the left. Displayed is the log density of the estimated posterior distributions. Numerical values are indicated by the colorbar on the right.

2.6.3 Model selection favours small models

In general model selection favoured small models for different data sets. Figure 2.6.6 shows the posterior distributions of the model index (left panel) and the model features on states, off states and perturbation index as color maps. The fits to the 5 pM and 10 pM data sets show the same clear favourite model: a random telegraph model with a perturbation of the RNA velocity and the initiation rate (model 115). Fitting the 20 pM data set found with model 125 a three state model with the same parameter perturbation as for the other two data sets. The second and third ranked models are small models with either two or three states for the three data sets (see table 2.6.1. Bayes factors of more than six indicate a strong selection towards the most highly ranked model. Models 115 and 125 appear among the top three models for fit to the 5 pM, 10 pM and 20 pM data sets. In conclusion it can be said that small models with perturbation of initiation rate and RNA polymerase velocity can explain these three data sets.

The 1000 pM data set appeared to be different from the other three as already indicated by the higher final distance obtained by model fitting. Model selection in case of the fits to the 1000 pM was not unique in that it showed big and small models and a different type of perturbation among the top three ranked models (see table 2.6.1). In a first attempt to fit the 1000 pM data set the most favourable model was a large model with ten states, namely 191 (1000 pM, 1st; in figure 2.6.6 and table 2.6.1). The second and third ranked models are a three state model (121) and a two state model (111) respectively. The Bayes factor between the first and the second model is 3.6 indicating a less strong selection as for the other three data sets. Moreover, the large difference in topology between the first and second model make a deeper analysis of model selection necessary for the 1000 pM data set.

To assess the uniqueness of the model selection the fit to the 1000 pM data set was repeated but with a different start population of particles. For the first fit the best 2000 particles out of the 50000 candidates that were sampled from the prior were selected. For a second attempt the the second best 2000 particles were used. These fits further will be referred to as 1000 pM, 1st and 1000 pM, 2nd as indicated in figure 2.6.6 and table 2.6.1. This led to a similar picture in terms of model selection in that not one clear favourite model could be found and that among the top three models are large and small models. Thus model fitting in case of the 1000 pM data sets needed to be improved. Model selection suggests that the data set could be explained by a small model as the

Data set	Most frequent	2nd most frequent	3rd most frequent
5 pM	[115], 80.5 %	[125], 10.9 %	[111], 3.8 %
10 pM	[115], 81.0 %	[125], 12.7 %	[112], 2.3 %
20 pM	[125], 86.7 %	[225], 6.7 %	[115], 5.9 %
1000 pM, 1st	[191], 67.4 %	[121], 18.5 %	[111], 5.5 %
1000 pM, 2nd	[281], 37.1%	[221], 18.8%	[191], 14.3%

Table 2.6.1: Model selection or fitting experimental data favours small models. Most frequent models in model selection after applying SMC ABC to experimental data. Shown are in square brackets the model by its number of on and off states and its perturbation index followed by the percentage of appearance among the final SMC ABC population of 2000 particles. The last two rows show the results for the two fits to the 1000 pM data set with the best 2000 initial particles (1000 pM, 1st) and with the second best 2000 particles (1000 pM, 2nd).

other data sets. Model fitting and selection to the other three data sets with different start populations led to similar results as the findings displayed in figure 2.6.6 and table 2.6.1 indicating numerical stability.

In addition to the model structure it stands out that a combination of two parameter perturbations are necessary to explain cell to cell variability. Namely for all but the 1000 pM data set a favored perturbation index of 5 indicates a combined perturbation of the polymerase velocity and the initiation rate. The change of perturbation in case of the 1000 pM data set might be an artifact. Model fitting of benchmark data sets with short off times and a combined perturbation did not yield the correct perturbation as explained earlier. Models without any parameter perturbation did not play a significant role which indicates that cell to cell variability of the kinetic parameters plays an important role.

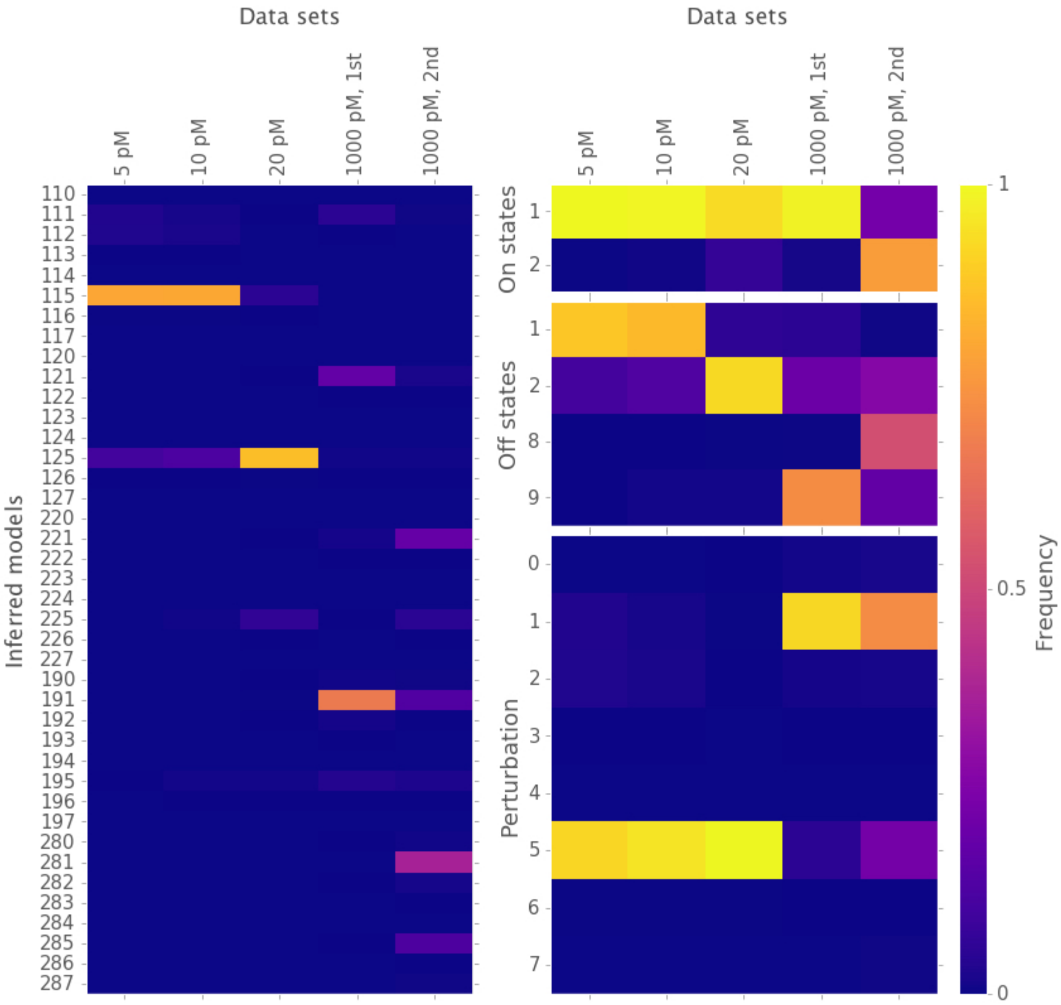

Figure 2.6.6: Model selection favours small models. Color maps visualizing model selection of the fitting algorithm to experimental data. The left panel shows the frequency with which each of the 40 models was chosen in the posterior distribution. Three digit numbers on the left denote the model with the first digit being the number of on states, the second the number of off states and the third the perturbation index according to table 2.5.3. Panels on the right display the frequency of how often each feature of the model was selected, e.g how often one on state was chosen for the 5 pM data set. The upper right panel displays the frequency of occurrence of either one or two on states. The middle panel shows the frequencies of the allowed numbers of off states and the bottom panel the perturbation index. Brighter colors denote higher frequencies as indicated by the color bar on the right.

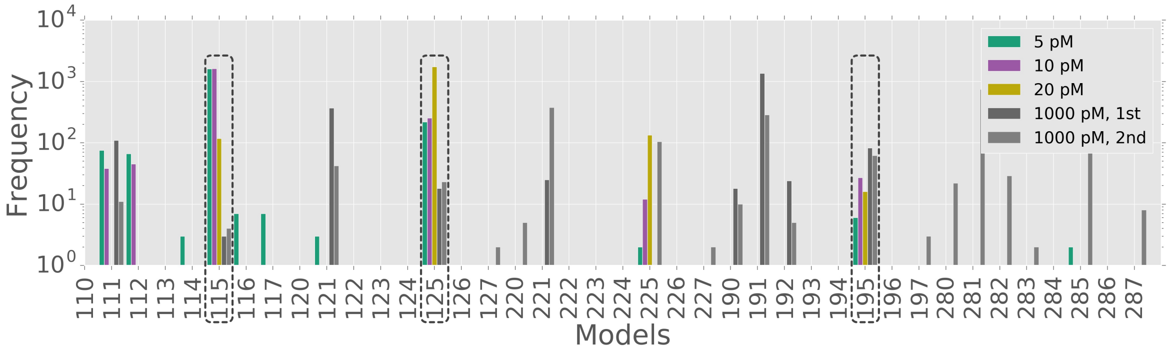

Figure 2.6.7: Model posteriors between individual fits show the biggest overlap at small models. Barplots of the frequency of appearance of all 40 models in the final particle population after the SMC ABC algorithm came to halt. Frequencies are reported in log-scale. The 1000 pM data set was fit with the best 2000 start particles (dark grey) and the second best 2000 start particles (light grey). Dashed grey boxes mark the models that appeared in all posterior populations.

2.6.4 Systematic analysis of model posteriors defines a common model topology for all doses

We wanted to reduce the number of free parameters and thus find a common model topology with only one or few parameters that are dose dependent. To find a global model topology that is able to explain all four data sets requires further analysis of the model posterior distributions for models that appear in all posteriors. Figure 2.6.7 shows the frequencies of all 40 different models in the posterior distributions of all fits. Dashed grey boxes frame the models that appear in all posteriors, namely the models 115, 125 and 195. The ten state model 195 appears less than 100 times in each posterior so that it was neglected for further analysis leaving the two small models 115 and 125 as candidates for a global model. Both models appear among the top three models for the 5 pM, 10 pM and 20 pM data sets but only infrequently for the 1000 pM data set. The inability to yield good fits to the 1000 pM data set as indicated by the higher final distance might be the reason for the less robust model selection.

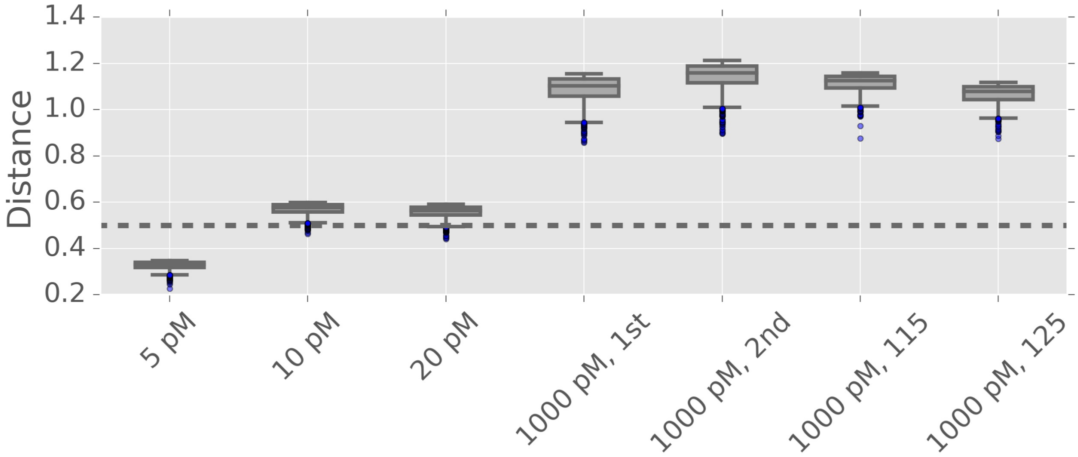

Figure 2.6.8: Repeated fitting of the 1000 pM data set yields similar distance values. Box plots show the distribution of the final distances of all particles populations after the SMC ABC algorithm came to halt. The dashed grey line marks the distance of 0.5 which is an indicator of an optimal fit as defined in figure 2.5.4. The 1000 pM data set was fit four times. The first two box plots show the distances for the fits obtained with the best 2000 start particles (1000 pM, 1st) and the second best 2000 start particles (1000 pM, 2nd). The last two box plots display distances of fits to the 1000 pM data set with fixed model topologies of 115 and 125.

To assess the abilities of the 115 and 125 models to explain the 1000 pM data set each was fit individually, i.e. the model topology was kept fix and only the model parameters were estimated. Figure 2.6.8 displays the distances of the final population of particles to all data sets. The first three box plots show again the distances of the fits to the 5 pM, 10 pM and 20 pM data sets for comparison. The last four box plots show the distances of the different fits to the 1000 pM data set. The first two represent the distances for the two fits with model selection and the last two the distances for the fits with a fixed model as indicated in the labels. Distances for all fits to the 1000 pM data set are in the same range indicating that the 115 and 125 models are able to explain the observed data to the same degree as the large models favoured by model selection. The fitted main model parameters for all four fits of the 1000 pM data set were in similar ranges. Figure 2.6.9 shows posterior distributions of the main parameters initiation rate (top) on time (middle) and off time (bottom) for the different fits to the 1000 pM data set. This shows that finding a unique model for the data set is difficult and that different model topologies result in fits with comparable quality and posterior parameter ranges.

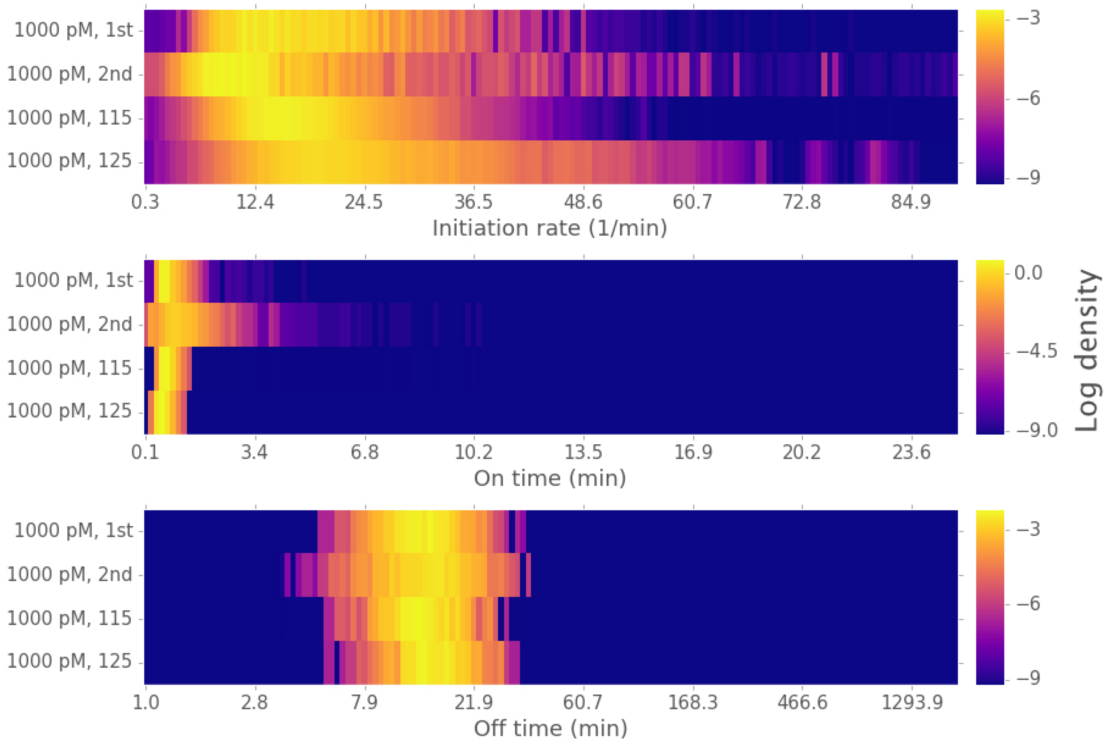

Figure 2.6.9: Repeated fitting of the 1000 pM data set yielded similar main parameters. The 1000 pM data sets was fit four times in total. Color maps show the log density of the posterior distributions of the main parameters initiation rate (top), on time (middle) and off time (bottom). The different fits were achieved with the best 2000 candidate particles as start population (1000 pM, 1st), the second best 2000 particles (1000 pM, 2nd) and with the model to be restricted to either 115 (1000 pM, 115) or 125 (1000 pM, 125).

This assessment of the model selection results suggest a common topology of a two or three state model which is capable to explain data obtained under different experimental conditions. A common model would reduce the number of free parameters and allows to explain results from different experimental conditions. For all but the 1000 pM data set a combined perturbation of polymerase velocity and initiation rate was found among the top three ranked models under all experimental conditions. Models 115 and 125 were found in the model posterior distributions of all fits and represent the most prominent candidates for a global model.

2.7 Fitting all data sets globally

The results of fitting experimental data sets individually led to the conclusion that it is possible to fit one global model to all data sets at once. Advantage of such an approach is its generality: it provides a minimal model with less parameters and avoids overfitting. Different experimental conditions can be explained by the same model which in addition will provide a mechanistic understanding of transcriptional regulation in response to alterations in estrogen concentration.

As common model topologies the 115 and 125 models will be used as discussed in the previous section. They represent a two and a three state model incorporating perturbations on the RNA polymerase velocity and on the transcription initiation rate. These models will not be subject to model selection during model fitting only model parameters will be estimated. Assessing the goodness of fit allows to discriminate between both model topologies. Transcription initiation rate k_m, promoter on time t_{on} and perturbation strength σ_{k_m} are assumed to be stimulus independent and thus represent global parameters. The 125 model has one extra parameter describing the ratio of the life times of the two off states which is assumed to be a global parameter, too. Only the promoter off time T is allowed to vary in a stimulus dependent manner as suggested by the results of the individual fits and therefore will be referred to as a local parameter. This approach reduces the number of free model parameters to three (four, for model 125) global and four local parameters. The four local parameters correspond to the four different estrogen concentrations of 5 pM, 10 pM, 20 pM and 1000 pM. To fit all four data sets by the 115 model individually requires 16 free parameters and in case of the 125 model 20. This reduced number of parameters facilitates model fitting in that the dimension of the parameter space that has to be searched is reduced. In an SMC ABC setting this means that a single particle consists of these seven (eight) parameters and the particles weight. The next sections will explain the generation of an adequate particle start population out of the results of the individual fits and the extension of the SMC ABC algorithm to fit all data sets simultaneously.

Common models	[115] & [125]
Regulated (local) parameter	Off time T
Range of initiation rate k_m	[5,35] min^{-1}
Range of on time t_{on}	[0.5,1.5] Minutes
Range of ratio of off state life times for 125 model	[0.8,1]

Table 2.7.1: Settings to filter individual fits for candidate global particles. Definition of common model and global and local parameters. For the global parameters the overlapping posterior regions were defined based on figure 2.6.4.

2.7.1 Start population for global fits from individual fitting results

The first step in fitting a global model applying the SMC ABC algorithm is to generate a sufficiently good start population of particles. This was done by filtering the results of the individual fits. First, for each condition all posterior particles favouring either the 115 or 125 model were selected. Second, the resulting particles were filtered further for overlap in the posterior distributions for the transcription initiation rate and the promoter on time. The overlap between the initiation rate posterior distributions was wide and thus the filtered range was defined to rates ranging from 5 to 35 min^{-1}. The overlapping region of the promoter on time was narrow due to the sharp on time posterior distribution found for the 20 pM data set (see figure 2.6.4 top and middle panels). The 125 model has an extra parameter u describing the ratio of the life times of the two off states. The filtered region for u was chosen to be between 0.8 and 1 based on the overlap of the posterior distributions. Table 2.7.1 summarises the filter settings to find candidate particles having common parameter values.

Each of the found particles already had promising values of the global parameters k_m, σ_{k_m} and t_{on}. To create a global start population good combinations with the local parameter promoter off T time were be found by a parameter scan. For each of the filtered particles multiple simulations with different candidate off times were generated and compared with the four data sets by calculating the distance. Candidate values for the off time parameter scan were selected based on the posterior distributions of the individual fits leading to a series of 19 values from short (5 minutes) to long (2000 minutes). Off times yielding the smallest distances were added to the final particles each corresponding to one estrogen concentration. In this way the most promising combinations of global and local parameters could be found.

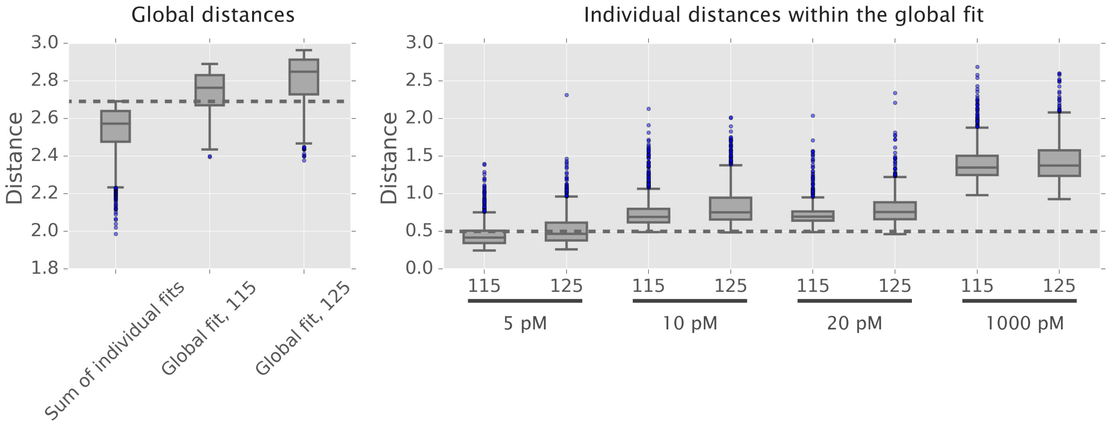

Figure 2.7.1: Global model fitting yields similar distances as individual fits Box plots showing the distribution of the final distances yielded by global fitting. The left panel shows global distances The left box plot is the sum of distances of all four individual fits displayed for comparison. The dashed line indicates the maximal distance of that distribution. The second box plot shows the distances of the global fit with model 115 and the right box plot with model 125. The right panel shows the distances to the different data sets within the global fit. The dashed line indicates a distance of one half that was estimated to be value for an optimal fit earlier.

2.7.2 Global SMC ABC fitting favours a two state over a three state model

SMC ABC model fitting could be extend to include multiple data sets at once without difficulty. As explained earlier a single particle for global fitting consisted of three or in case of the 125 model of four global (k_m, t_{on}, σ_{k_m}, u) and four local parameters (one off time T for each estrogen concentration). For each combination of the global parameters together with one local parameter a simulation was generated and compared to the appropriate experimental data set. The global distance of a candidate particle was the sum of all the four individual distances corresponding to four estrogen concentrations. Proposing new particles was done as before by individually changing parameters of accepted particles by the same proposal distributions as for the individual fits. The particle population size was reduced to 1000 particles due to long computation times. 1000 particles appeared sufficient to estimate seven (eight) parameters without model selection.

Figure 2.7.1 summarises the distribution of particle distances after the SMC ABC algorithm terminated. The left panel shows the global distances. For comparison the left box plot displays the sum of the distances of all particles of the individual fits. The dashed line marks the maximal distance of the sum of the individual fits. The second and third box plots show the final distances of the global fits of models 115 and 125, respectively. Fitting the 125 model yielded larger distances as the 115 model indicating that the two state model is sufficient to describe the data especially given that it has one parameter less and thus less freedom. The median of the global distance of the 115 model almost reached the marked distance indicating that the the individual fits are only marginally better than the global fit. Moreover, the distances to the 5 pM, 10 pM and 20 pM for the 115 model are slightly smaller than for the 125 model. This further indicates that the 115 model might be more appropriate to explain the experimental data.

Comparing the particles of the global fit with experimental data reveals that the fits are only marginally worse than fitting the data sets individually. Figure 2.7.2 shows the four main features of the experimental data compared with the 300 best particles from the global fit. The global model in general can explain the data almost equally well as the individual fits with similar but more pronounced deviations (see figure 2.6.2 for comparison). The amount of variability between the different particles (thin blue lines) is larger than before which is understandable by the reduced number of free model parameters and reflected in the slightly larger distances displayed in figure 2.7.1. Table 2.7.2 gives the 5%, 50% and 95% percentiles of the global parameter posterior distributions.

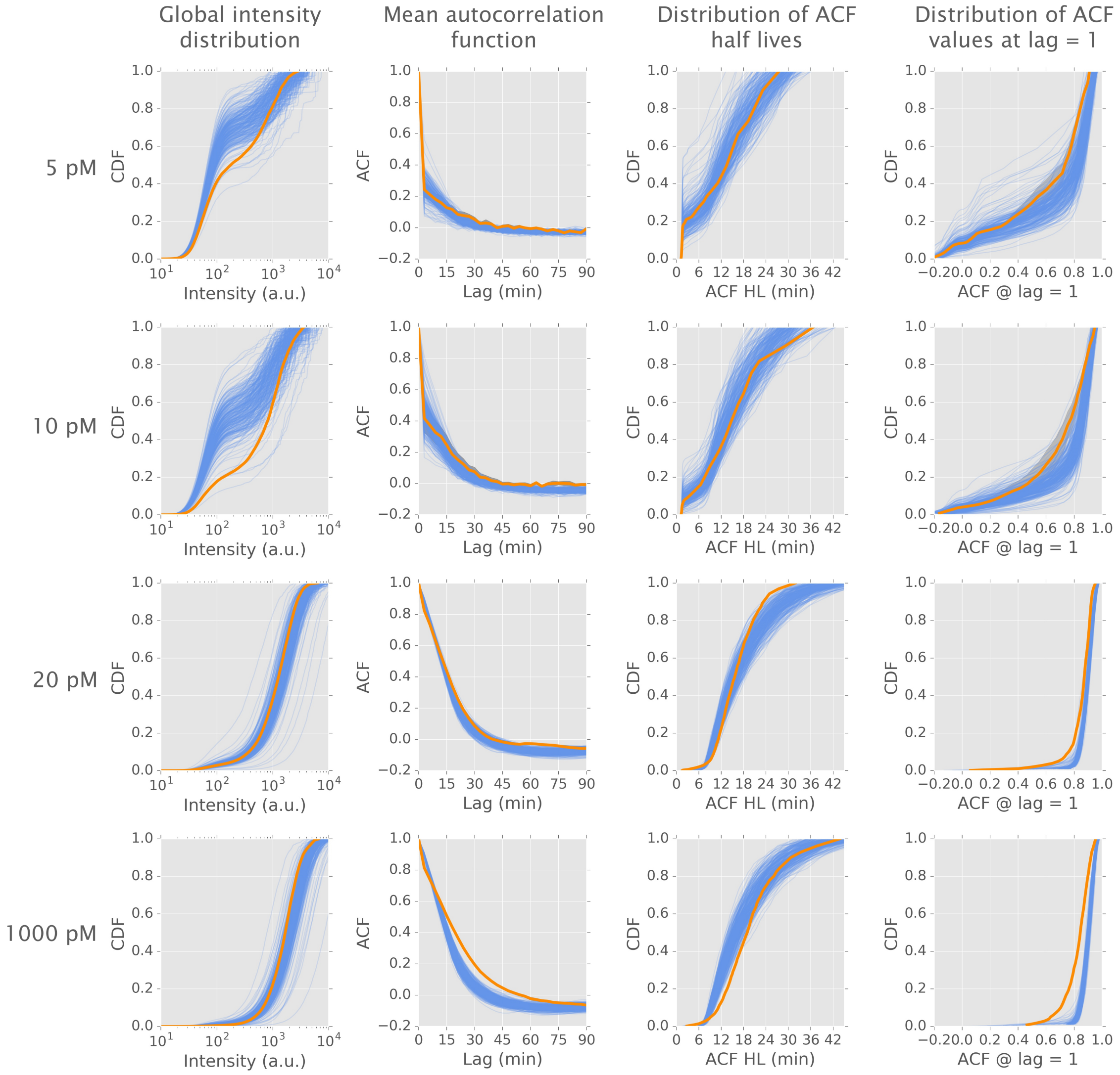

Figure 2.7.2: Global model fitting can explain features of the data in a similar quality as individual fits. Agreement of the 300 best particles from global fitting (thin blue lines) with experimental data (orange lines). Arrangement of the figure is the same as in figure 2.6.2. Columns represent features of the data, from left to right: mean ACF, global histogram, ACF half life and ACF value at lag = 1. Grey shaded areas in the ACF column denotes the ±STD region and in the global intensity histogram the background intensity.

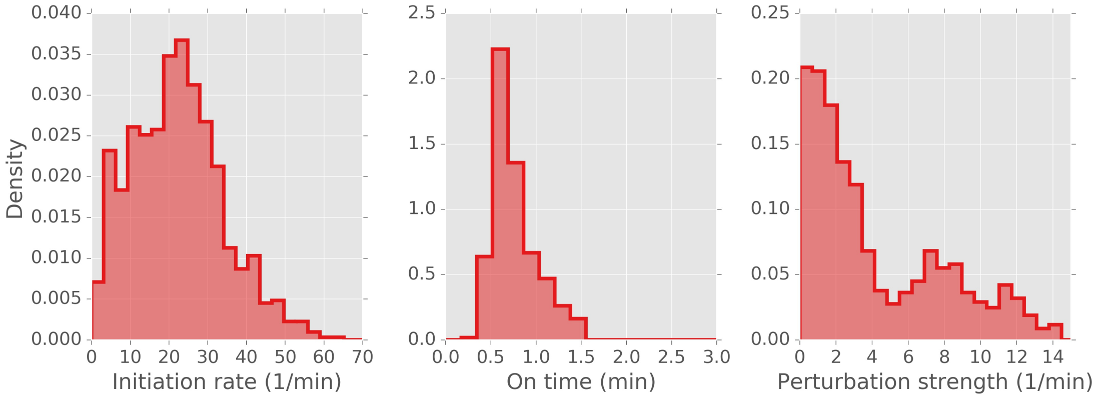

Figure 2.7.3: Promoter exhibits short on times. Posterior distributions of the global parameters initiation rate k_mm promoter on time t_{on} (middle) and strength of the perturbation on the initiation rate σ_{k_m}.

Initiation rate (1/min)			**On time (min)**			**Perturbation strength (1/min)**		
5%	50%	95%	5%	50%	95%	5%	50%	95%
4.3	21.8	43.3	0.5	0.7	1.3	0.2	2.7	11.7

Table 2.7.2: Percentiles of global posterior distributions. Table summarising the 5%, 50% and 95% percentiles of the posterior distributions of the global parameters initiation rate, promoter on time and perturbation strength.

Figure 2.7.3 displays the posterior distributions of the three global parameters of transcription initiation rate (left), promoter on time (middle) and strength of the perturbation of the initiation rate (right). All three distributions are well defined over a constrained interval. Especially the promoter on time shows a narrow distribution centered around one minute indicating a comparably short active phase of the promoter. Transcription initiation rate is centered around a value of 20 min^{-1}. This together with an on time of one minute leads on average to twenty RNA molecules per burst.

Posterior distributions of the local values of the promoter off time are shown in figure 2.7.4. As before for the individual fits a clear stimulus dependency is visible with short off times at high estrogen concentrations and very long off times for low concentrations (see figure 2.6.4 for comparison). A summary of the percentiles of the posterior distributions is given in table 2.7.3. The distributions do not vary significantly from fitting data sets individually to fitting all globally which indicates stability of the results.

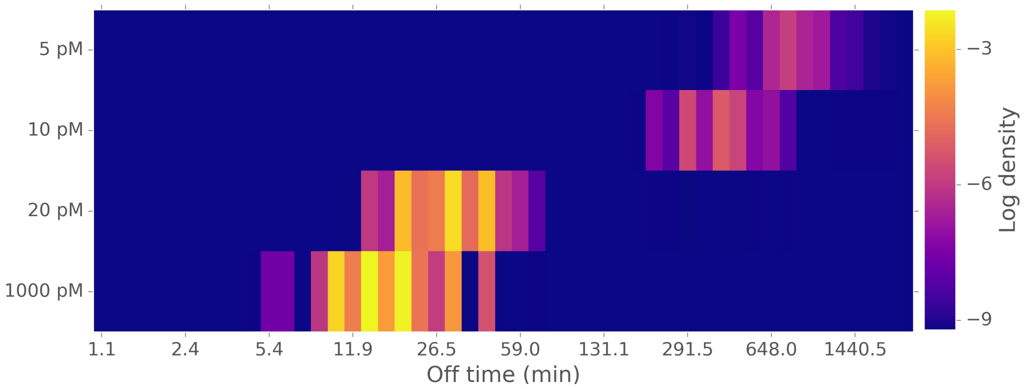

Figure 2.7.4: Switch-like modulation of promoter off times. Color map displaying the posterior off time distributions found by global model fitting. Rows represent results for individual data sets. Brighter colors indicate higher posterior densities displayed here as the logarithm of the estimated densities. Numerical values are indicated by the color bar on the right. Binning of the histograms is logarithmically. See bottom panel of figure 2.6.4 for comparison.

Summary

In conclusion can be said that the 115 model globally can explain the data well and reached a smaller distance value than the 125 model. This indicates that a two state model is the most likely to explain transcriptional stochasticity of the GREB1 gene. All main features of the experimental data are matched almost to the same degree as for the individual fits. Moreover, defining three parameters (k_m, t_{on}, σ) as globally stimulus independent and keep only the promoter off time as stimulus dependent led to similar parameter posterior distributions as fitting the data sets individually. A main point that remained open is how the promoter off time is regulated by the estrogen concentration. This will be addressed in the next section where a model extension incorporating estrogen signalling to the gene promoter is proposed.

Data set	Individual fits			Global fit		
	5%	50%	95%	5%	50%	95%
5 pM	433.7	637.9	971.4	500.0	800.0	1095.7
10 pM	243.5	382.7	583.2	270.4	400.0	300.0
20 pM	22.2	31.6	44.1	20.0	30.0	40.0
1000 pM	8.2	13.4	21.8	10.0	15.0	30.0

Table 2.7.3: Global fitting and fitting of individual data sets yield similar dose dependent off times. Table showing the 5%, 50% and 95% percentiles of the posterior off times in minutes yielded by individual fits and global fitting, respectively.

2.8 Modelling parameter dose dependency

Model fitting in the previous sections revealed that the time course data of emitted fluorescent light from active transcription sites can be described by a random telegraph model with one transcriptionally active and one inactive state. The observed dose dependency could be explained via modulation of the promoter off time, i.e. the time the gene spends in its inactive state. What remains unclear from the stochastic model is how the promoter off time is regulated by the estrogen concentration and what are the kinetics of this process.

In this section the stochastic model will be extended by a signalling pathway to explain the observed dose response behaviour and additionally to understand the response time of the cells. Response time refers to the time a system needs to react to changed experimental conditions. This extended model will be compared to data measuring the accumulation of RNA transcripts in populations of synchronised cells allowing to verify the model with data obtained in experiments independent from the previous ones.

The strategy will be as follows. First a possible estrogen signalling pathway to the GREB1 promoter will be proposed. Then the dose response relation of this pathway is fitted to the dose dependent rates $k_{off} = 1/T_{off}$ to leave the promoter off state yielded by global model fitting. The off rate dose dependency is assumed to follow the dose response relation of the signalling pathway. Parameters of the calibrated dose response relation allows to simulate the temporal dynamics of the proposed signalling pathway. Based on these simulations together with simulations of RNA accumulation of the fitted stochastic model it is possible to predict the systems response time which then will be compared to experimental data. Response time here means the time the system needs to reach half of the steady state level.

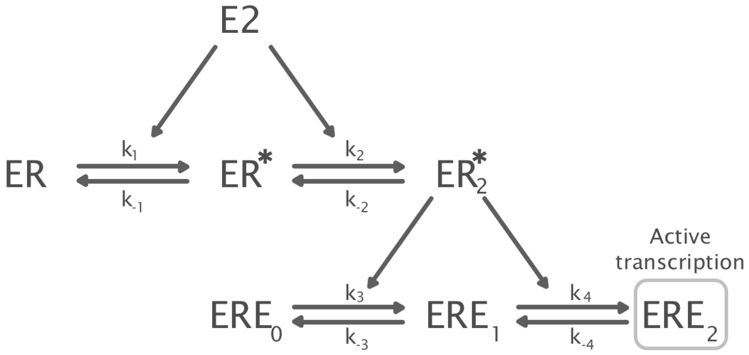

Figure 2.8.1: Two level estrogen signalling pathway. Assumed signalling path way that transmits an input estrogen signal $E2$ to the estrogen responsive elements ERE. In a first reaction level the estrogen receptor ER binds its ligand forming ER^* followed by homodimerisation to ER_2^*. Both forward reactions are assumed to be estrogen sensitive. The estrogen receptor dimers can bind to estrogen responsive elements in gene promoter regions. The GREB1 promoter carries multiple such binding sites with two very prominent EREs as revealed by a time course ChIP-seq experiment (see figure 2.2.2). Binding to both EREs is assumed to be necessary to activate gene transcription as indicated with the grey box.

2.8.1 Two step model of estrogen signalling to the GREB1 promoter

Estrogen signalling to the GREB1 promoter was assumed to occur in a two step process as depicted in figure 2.8.1. The first step is the binding of estrogen to its receptor followed by receptor dimerisation [Dahlman-Wright et al., 2006]. The second step is binding of the ligand bound receptor homodimer to estrogen responsive elements (EREs) in the promoter region of the GREB1 gene [Kumar and Chambon, 1988]. Time course chromatin immunoprecipitation experiments against the estrogen receptor alpha revealed that two EREs in close proximity to the transcription start site of the GREB1 gene show the highest signal (see the second line in figure 2.2.2). For the signalling pathway it was assumed that binding to those two EREs is sufficient to activate GREB1 transcription. Figure 2.8.1 graphically depicts the signalling path way. The model describes the deterministic behaviour of the population average. Later it will be tested against population average data which rectifies this approach.

For each level of the signalling pathway a conservation law holds as:

$$\begin{aligned} ER_{tot} &= ER + ER^* + 2ER_2^* \\ ERE_{tot} &= ERE_0 + ERE_1 + ERE_2 \end{aligned} \tag{2.8.1}$$

Where ER_{tot} is the total amount of the estrogen receptor, ER is the unbound and ER^* the ligand bound receptor. ER_2^* represents the amount of receptors bound in homodimers that can bind to EREs. ERE_{tot} is the total amount of GREB1 promoters in a population of cells. ERE_i denotes the concentration of promoters where either none ($i = 0$), one ($i = 1$) or two ($i = 2$) EREs are bound by receptor dimers.

The dynamics of each of the six species can be described by an ordinary differential equations. With the conservation laws the effective number of ODEs can be reduced to four. Applying mass-action kinetics, i.e. assuming estrogen and ER_α to be in excess, yields the following equations:

$$\begin{aligned} \frac{dER^*}{dt} &= k_1 \cdot E \cdot (ER_{tot} - ER^* - 2ER_2^*) - k_{-1} \cdot ER^* - k_2 \cdot ER^* \cdot E + k_{-2} \cdot ER_2^* \\ \frac{dER_2^*}{dt} &= k_2 \cdot ER^* \cdot E - k_{-2} \cdot ER_2^* \\ \frac{dERE_1}{dt} &= k_3 \cdot ER_2^* \cdot (ERE_{tot} - ERE_1 - ERE_2) - k_{-3} \cdot ERE_1 - k_4 \cdot ER_2^* \cdot ERE_1 + k_{-4} \cdot ERE_2 \\ \frac{dERE_2}{dt} &= k_4 \cdot ER_2^* \cdot ERE_1 - k_{-4} \cdot ERE_2 \end{aligned} \tag{2.8.2}$$

The dose response of the signalling pathway as an input-output relation is yielded by assuming steady state, i.e. the left hand side in the above equations equals zero and solving the resulting algebraic equations. For each of the two steps an individual dose-response relation can be found. The input for the first step is the estrogen concentration. The second step takes the output $ER_{2_{ss}}^*$ of the first step as input.

$$\begin{aligned} \frac{ER_{2_{ss}}^*}{ER_{tot}} &= \frac{E^2}{K_1K_1 + K_2E + E^2} \\ \frac{ERE_{2_{ss}}}{ERE_{tot}} &= \frac{{ER_{2_{ss}}^*}^2}{K_3K_4 + K_4ER_{2_{ss}}^* + {ER_{2_{ss}}^*}^2} \\ K_i &= \frac{k_{-i}}{k_i} \end{aligned} \tag{2.8.3}$$

The reaction rates $k_{\pm i}$ are not appearing in the dose-response relations but rather their ratios K_i. This indicates that while fitting an experimental dose repsonse relation the individual reaction rates cannot be identified to a common factor. Setting the total amount of receptor to $ER_{tot} = 1$ and combining both equations yields the full dose response relation of the two step system:

$$\frac{ERE_{2_{ss}}}{ERE_{tot}} = \frac{E^2}{(K_1K_2 + K_2E + E^2)\left(\frac{E^4}{(K_1K_2+K_2E+E^2)^2} + \frac{K_4E^2}{K_1K_2+K_2E+E^2} + K_3K_4\right)} \tag{2.8.4}$$

The rate to leave the promoter off state $k_{off} = 1/T_{off}$ is assumed to be proportional to ERE_2 and thus to follow the same dose response behaviour. Fitting the above equation to the mean off rates will yield the ratios of the individual back- and forward reactions K_i.

2.8.2 Fitting dose response and simulating response times

The dose response equation 2.8.4 was fitted to the mean inverse promoter off times obtained by the global model fit. The fact that only four data points are provided over the range from 5 pM to 1000 pM stimulus made it difficult to obtain only a single parameter set of the ratios K_i. Therefore, 15.000 starting points for the optimisation were created by latin hypercube sampling and subject to local optimisation. The cost function to minimise was defined as the sum of squared distances between data and fit. For further analysis the best 15% of all fits were considered. The left panel of figure 2.8.2 shows the best fit together with the data indicating that the data can be fitted well by the proposed relation although the data exhibits a very steep dose response.

With the parameters from the best model fits the equation system 2.8.2 can be simulated forward. The response time of the signalling path way is the time the ERE_2 level needs to reach half steady state level. Because fitting the dose response yielded only the ratios of backward to forward reaction rates the individual rates are only determined up to a common factor. For simulations the backward reactions were assumed to equal one and the forward reactions were scaled accordingly. By this it was possible to obtain values for the signalling response times for all fitted parameter values. As initial condition was assumed that the receptor has not yet bound any ligand ($ER^* = ER_2^* = 0$) and that both EREs are empty ($ERE_1 = ERE_2 = 0$). The middle panel of figure 2.8.2 shows

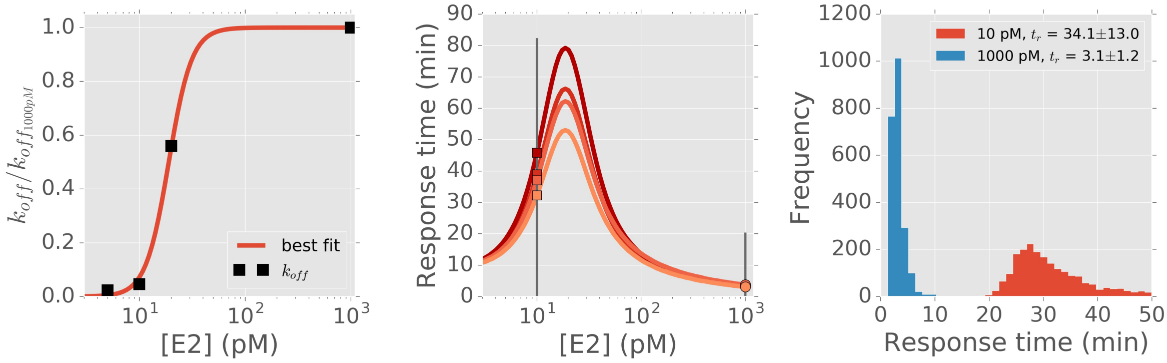

Figure 2.8.2: Fitting the off rate dose response allows to simulate the response time of the signalling pathway. The left panel shows the best fit (red line) of equation 2.8.4 to the normalised mean off rate k_{off} from the global model fit (black squares). The middle panel shows three examples from the 15% of all fits of the dose dependent signalling response time. Vertical lines mark estrogen concentrations of 10 pM and 1000 pM that were investigated experimentally. Colored squares mark the simulated values of the response times at those concentrations. The right panel displays the full distributions of simulated response times of the parameter sets from best fits. In the legend the corresponding mean values are indicated.

four example dose dependent curves of response times simulated with the fitted reaction rates. The response time of the signalling pathway is dose dependent and the peak in the curves marks the EC_{50} value of the value of the k_{off} dose response of approximately 20 pM. Experimentally two estrogen concentrations were investigated, namely 10 pM and 1000 pM as marked by vertical grey lines. Intersection points with the response time curves are marked by squares at 10 pM and by circles at 1000 pM. The right panel shows histograms of the distributions of all obtained simulated response times. The mean response time values ±SD for each distribution is shown in the figure legend. These values indicate the expected delay that the proposed signalling path way needs to bind half of the ERE pairs in a population of cells at 10 pM and 1000 pM respectively.

2.8.3 Comparing simulated response times with accumulation of RNA in cell populations

In the previous two sections an estrogen signalling pathway to the GREB1 promoter was proposed and by fitting the dose response relation of that pathway to the resulting mean k_{off} values allowed to simulate response times. In this section these simulations will be compared to two experimental data sets. The first data set measures the accumulation of RNA transcripts of the GREB1 gene in a population of initially synchronised cells. The second data set consists of time lapse imaging of transcription sites in synchronised single cells. Synchronisation was done by growing the cells in estrogen free medium for three days. This prevents the cells from division and poises the gene promoters in the off state. Thus, no new transcripts are produced and the total amount of transcripts decreases as existing transcripts decay. At time point zero the cells received a defined estrogen signal which released the blockage and transcription starts again. Accumulation of RNA allows to estimate the response time of the cells which is generated by the signalling pathway and transcriptional activation.

The resulting particles from the global SMC ABC model fit allow to simulate how fast transcripts would accumulate in a synchronised population of cells after the release of the blockage. Initially all promoters are in the off state and an estrogen signal allows the promoters to switch into the on state with a dose dependent rate k_{off}. But before this can happen the estrogen signal has to reach the gene promoters via the proposed signalling pathway. Thus, we propose that the cell's response times consist of two parts. A dose dependent first part created by the signalling path way and a dose independent second part due to the stochastic switching from the off into the on state.

Estrogen signalling can explain the dose dependent response times in cell populations

In synchronised populations of cells RNA molecules were collected at ten minute intervals. Subsequently RNA was reversely transcribed into DNA which then was detected by quantitative PCR (RT-qPCR) [Nolan et al., 2006]. As targets for PCR primers two regions within the GREB1 transcript were selected. One early region within intron two and one late region within intron 32. This allows to investigate the accumulation of both targets. The early region is transcribed soon after the initiation of transcription. The increase in detected RNA molecules measures the combined delay of signalling

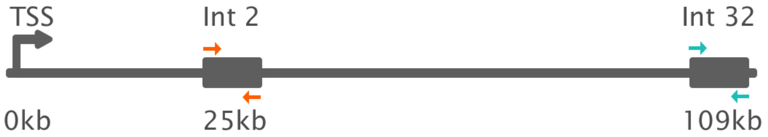

Figure 2.8.3: Primer positions within the GREB1 transcript for RT-qPCR experiments. Primer positions within the GREB1 transcript are indicated by colored arrows. Colors correspond to the plots in figure 2.8.4. Rectangles represent the position of transcript introns 2 and 32 as indicated. Below the position in kilo base pairs is given.

and transcription. The time delay between the accumulation of both targets gives an estimation of the velocity of the RNA polymerase as it elongates the transcript. The structure of the transcript of the GREB1 gene that was targeted by PCR primers is displayed in figure 2.8.3. The transcript is slightly longer than the reporter construct used before. Therefore, the simulation of single cells had to be adjusted to correctly simulate RNA elongation. RT-qPCR time course experiments were performed at 10 pM and 1000 pM. All experiments were performed by Christoph Fritzsch and Monika Kuban.

In RT-qPCR experiments the expression of the GREB1 gene was quantified relative to the house keeping gene GAPDH, a gene that is transcribed constantly and not influenced by the estrogen concentration [Barber et al., 2005]. This methods ensures robust quantification in different cell populations. For better comparison of the resulting time courses they were normalised by the mean value of the last three points of the time course when the accumulation is assumed to have reached steady state. The response time of RNA accumulation was estimated via linear interpolation between the last value below and the first value above one half in the normalised time course. Experimental data was available in technical triplicates. The ±SD region from the triplicates allowed to estimate an error range for the response time.

Estimated response times for both primer targets and both experimental conditions are displayed in table 2.8.1. The third column (RTM simulation) displays the response time of simulations applying the two state promoter model were at time zero all promoters are in the off state. This time was found to be stimulus independent. The measured response times show a dose dependency with longer response times at 10 pM estrogen. This dose dependency must be due to the signalling which was not considered in the stochastic simulations.

The three right most columns (RT-qPCR data) of table 2.8.1 show the estimated response times from experimental time courses with lower and upper bounds resulting

Stimulus	**Intron**	**RTM simulation**	**Signalling**	**RT-qPCR data**		
				low	center	high
10 pM	Int2	27.8	34.1±13.0	52.0	61.0	66.0
	Int32	47.3		74.6	78.1	93.8
1000 pM	Int2	28.7	3.1 ± 1.2	37.9	41.1	52.0
	Int32	48.5		62.0	65.2	72.5

Table 2.8.1: Estimated response times from simulation of transcription alone and RT-qPCR time courses. Displayed are the response times of simulated populations of synchronised cells, the response time of the proposed signalling pathway and measured response times from RT-qPCR time course measurements. All times are given in minutes. The low, center and high values of the measured response times were estimated from the error range given by the SD of triplicate measurements.

from the experimental error. The time difference between the early and late targets is similar under both conditions at approximately 20 minutes. Together with the distance of the primer targets of 84 kb this provides a measure of the mean RNA polymerase velocity of approximately 4.2 kb/min. This confirms the assumptions made for modelling transcriptional elongation. The velocity lies in the assumed range and it is not dose dependent.

The response times at 10 pM and 1000 pM are different indicating a dose dependent mechanism. The response times simulated with the stochastic transcription model alone are 29 minutes leaving differences of 32 and 12 minutes respectively to be filled by the response time of the signalling pathway. The simulated mean response times of the signalling pathway were 34 and 3 minutes under the constraint that all backwards reactions within the pathway were set to one (see figure 2.8.2, right panel). In case of the results from the experiment done at 10 pM the agreement is very close.

Figure 2.8.4 shows the experimental RT-qPCR time courses together with the simulated RNA accumulation time courses yielded by the stochastic model. The simulated time courses were shifted to the right by 34 minutes at 10 pM and by 3 minutes at 1000 pM. Those numbers resulted from the simulated response time with the kinetic parameters from fitting the k_{off} dose response. At 10 pM the simulation shows a clear overlap within the experimental error range for both the early and the late PCR target. The agreement between simulation and experiment for 1000 pM is less strong which might be due to the extremely steep dose response curve of the promoter off rates. A less steep curve leads to different values of the fitted K_i which in turn cause different response times. In addition,

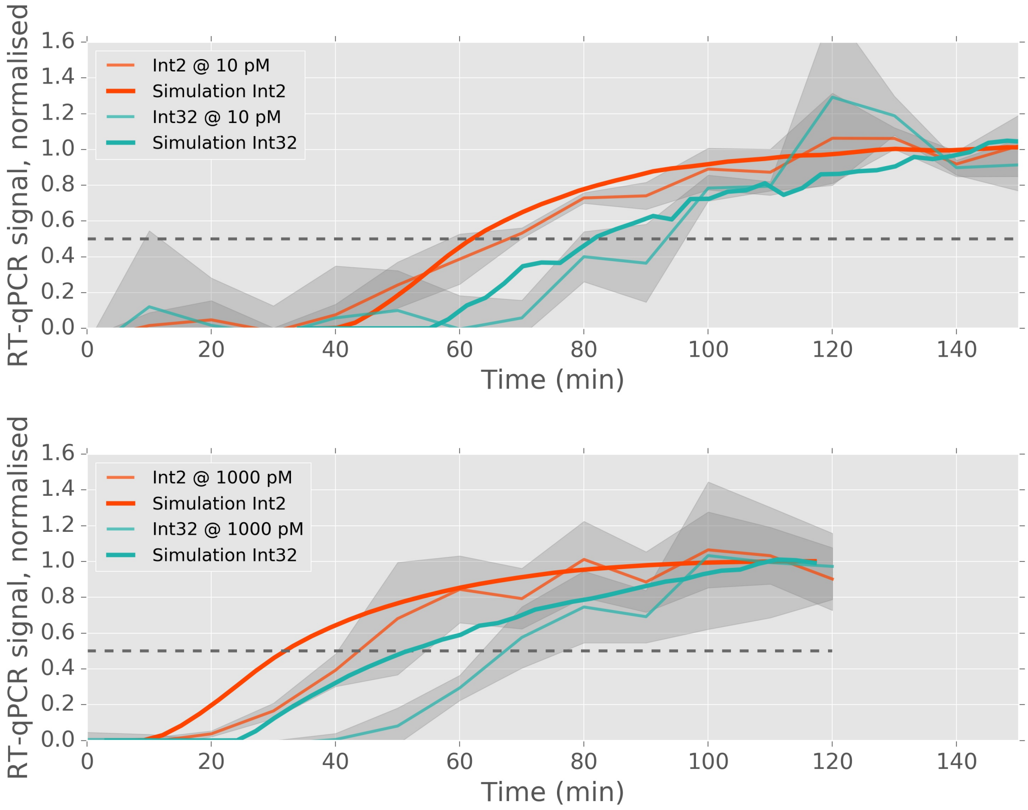

Figure 2.8.4: Comparison of simulated RNA accumulation with RT-qPCR data shows close agreement in response times. Thin colored lines display the normalised RNA accumulation over time at 10 pM (top) and 1000 pM (bottom) estrogen. Red lines represent the early PCR primer target and blue the late primer target. Grey shaded areas denote the ±SD region of technical triplicates. Thick colored lines show the simulated accumulation of RNA shifted by their dose dependent signalling response times of 34 minutes at 10 pM and 3 minutes at 1000 pM, respectively. Dashed grey lines indicates the half maximal value. Intersection points with this line mark the actual response times.

only four data points along the curve do not provide satisfactory information for the exact dose response behaviour. Despite the sparse sampling of the dose response curve the overall agreement is good and this suggests that the fitted two state model together with the proposed signalling pathway can not only explain the dose dependent promoter off time regulation but also the dynamics in terms of the response time of the full system. The response of the stochastic model is dose independent and the dose dependent delay is exclusively due to the signalling pathway.

Transcription in synchronised single cells

In a second approach to measure RNA accumulation synchronised cells were image under the microscope yielding single time courses of transcriptional activity. This allows to directly measure response in single cells whereas RT-qPCR is a population method. Moreover, PCR detects all transcripts, those still located at the transcription site and those who are already finished. In total 112 cells were imaged. The upper panel in figure 2.2.6 shows all measured time courses as a color map. Below the mean value of all cells is plotted. The dashed vertical line indicates the time point when 1000 pM estrogen were added. After that time point a clear accumulation of RNA signal is visible.

Figure 2.8.5 compares this curve (red) with the RT-qPCR curve from the early PCR target from the lower panel of figure 2.8.4 (blue) and with the simulation from the same plot (yellow). Both the RT-qPCR curve and the simulation are within the error range of the mean curve from single cells. The simulation shows an almost perfect agreement with the response time of the mean of the single cells. This is another confirmation that the stochastic model of gene transcription coupled with the deterministic estrogen signalling pathway makes reliable predictions about the systems behaviour.

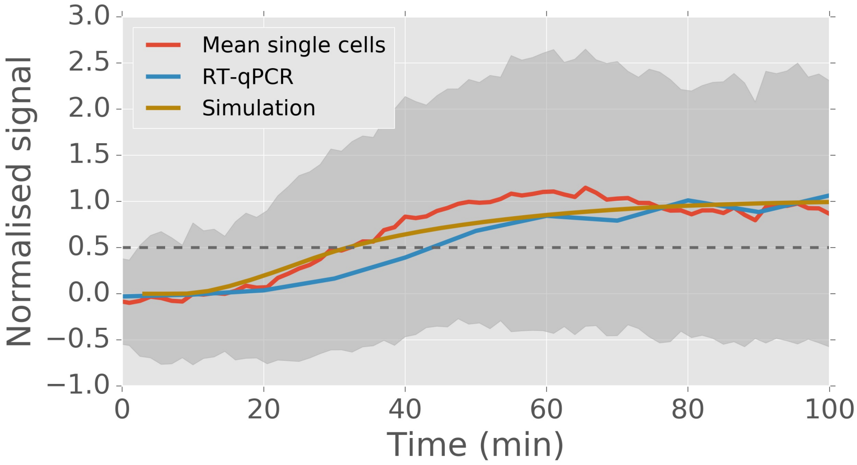

Figure 2.8.5: Accumulation of RNA in single cells and cell populations can be described by the combined model of estrogen signalling and stochastic transcription. The three different curves show accumulation of RNA measured by time lapse microscopy in single cells (red) and RT-qPCR in cell populations (blue) compared with simulations (orange). The grey region denotes the ±SD error region of the mean signal from single cells. All curves where normalised to one for better comparison.

3 Discussion

This work is concerned with the study of estrogen induced transcription of the GREB1 gene in estrogen sensitive MCF-7 breast cancer cells. Transcripts of the gene were labelled fluorescently allowing direct imaging of nascent transcripts in live cells. Time lapse microscopy enabled to study the stochastic dynamics of transcription under different experimental conditions. We calibrated stochastic models of different complexity to time course data to reveal the mechanism behind GREB1 transcription in single cells.

The observed time courses of transcriptional activity clearly indicated that transcription of the GREB1 gene occurs in bursts similar as many other eukaryotic genes [McAdams and Arkin, 1997; Raj et al., 2006; Suter et al., 2011; Zoller et al., 2015]. Moreover, the global intensity histograms at different estrogen concentrations and static dose response relations showed a clear estrogen dependency of GREB1 transcription with higher levels of transcription at higher estrogen levels consistent with previous findings [Rae et al., 2005; Laviolette et al., 2014]. A novel observation was that the shape of those histograms was bimodal with one peak representing the background signal and one peak representing active transcription. Increasing the estrogen stimulus characteristically pronounces the peak of active transcription while at the same time the background peak becomes less prominent.

We directly observed nascent transcripts which is in contrast to previous attempts to measure transcriptional activity in single cells which incorporated the use of short lived luciferase proteins as reporters rather than the transcript itself [Suter et al., 2011; Harper et al., 2011; Molina et al., 2013; Zoller et al., 2015]. This made it necessary to include translation of RNA into protein into model based interpretation of the data requiring strong assumptions of the process. RNA translation into a protein is a highly regulated and complex process and thus observation of the protein hides important details of transcriptional initiation. In addition, our approach of labelling nascent transcripts was implemented by a short DNA insertion into an endogenous gene. This conserves the chromatin environment and should not interfere with the natural behaviour of the cell. Moreover, the model system enables to study the process at different experimental conditions using the estrogen concentration as an input influencing transcription of GREB1.

In the introduction two major questions of this work were posed, namely:

1. Which mechanisms can explain the stochastic nature of gene transcription?
2. How does stochastic gene transcription change upon changing environmental conditions?

Our results show that a stochastic model of only two states can explain experimental data of transcription in single cells better than more complex models with multiple internal states. The two promoter states represent transcriptional activity and inactivity and the promoter stochastically switches between these two states. Thus, the answer to the first question is a stochastic two state or random telegraph model. In addition to the model structure, our results showed that dose dependent regulation of transcription is achieved by off time modulation. At low estrogen concentrations the gene spends more time in its inactive state than at high concentrations whereas promoter off time and initiation rate are not subject to a dose dependent regulation. This modulation of a single parameter of the model answered the second question and allowed to explain transcriptional activity over a wide range of stimuli from 5 pM to 1000 pM.

This chapter will discuss these results in the light of the current knowledge of transcriptional activation. The following sections will discuss the random telegraph model and the biological role of its two states of transcriptional activity. Subsequently, estrogen dependent transcriptional regulation, cell to cell variability and the applied likelihood free bayesian model calibration algorithm will be discussed.

3.1 The random telegraph model

A class of models of different complexity were calibrated to experimental data and subject to model selection. Model selection favoured the smallest possible model topology that allows transcriptional regulation namely a two state or random telegraph model. Such a model stochastically switches between the states of transcriptional activity and inactivity. It facilitates a high variability in the numbers of transcripts per cell by fostering transcriptional bursts [Paulsson, 2005]. Modulation of the switching rates between model states allows to adopt the transcriptional level to external stimuli [Munsky et al., 2012].

Several previous studies found promoters exhibiting stochastic behaviour compatible with a three or more state model where the promoter off phase consists of two or more distinct states leading to a peaked off time distribution [Suter et al., 2011; Harper et al., 2011; Zoller et al., 2015]. The topology of the switching process between active and inactive transcription, however, appears to be highly gene specific [Zoller et al., 2015]. So far transcriptional dynamics were measured only for specific examples of genes. Single cell RNA sequencing would allow to investigate the distributions of transcript copy numbers in single cells genome wide [Tang et al., 2010]. Such distributions could be compared to simulated distributions and thus calibrated to models with different numbers of internal states. This would allow to investigate topologies of promoter states for all genes.

The two state or random telegraph model not only is the simplest model that enables transcriptional regulation by a dose dependent modulation of model parameters but it is at the same time the model exhibiting the widest variability or noise. Mathematically the stochastic models we calibrated to data are described by the theory of markov chains. In a markovian system like the random telegraph model the waiting times to remain in one state are exponentially distributed [Van Kampen, 1992]. Thus, the off time of the two state or random telegraph model is exponentially distributed and the mean off time equals the standard deviation of the off time which is an inherent characteristic of the exponential distribution. This feature leads to a coefficient of variation of the promoter off time of one. In case of multiple internal states during the off phase the total off time distribution is defined by the convolution of the waiting time distributions of the individual states. The resulting off time distribution in contrast to an exponential distribution is peaked and, more importantly, the variability is reduced with a coefficient of variation smaller than one. This more accurately defined off time should be reflected on the RNA level by a reduced variability, i.e. a smaller coefficient of variation of the RNA copy number per cell. Thus, transcription of GREB1 with its two states appears not to be tuned for noise suppression.

3.1.1 The discrepancy between deterministic transcription cycles in cell populations and stochastic transcription in single cells

Binding and release of transcription factors to estrogen regulated gene promoters was observed to occur in highly regular and deterministic fashion as so called transcription cycles (see figure 1.3.1) [Métivier et al., 2003, 2004, 2008]. This high regularity is in stark contrast to the stochastic transcription observed in single cells. Similar to previous experiments by Métivier et al. on the binding and release of transcription factors to the promoter we synchronised single cells by estrogen starvation. We found that after receiving an estrogen signal cells quickly upregulate transcription but we did not see cycles in the production of nascent transcripts (see figure 2.2.6) which revealed a major discrepancy between these two types of data.

Calibration of cyclic ratchet like ODE models revealed that several hundred states are necessary to explain the observed deterministic transcription cycles on the epigenetic level. To the contrary, in single cells only a small number of states was found to be sufficient to explain experimental data. On the epigenetic side the biological advantage of such a ratchet like process is that it accelerates the assembly of large protein complexes on the chromatin in contrast to equilibrium binding [Rybakova et al., 2015]. As explained earlier such a multi state process strongly can suppress noise whereas the few states of transcriptional activity give room for high variability. The latter might be favourable in changing environmental conditions. The connection between both levels, however, remains unclear.

The experimental methods to acquire the two types of data are different in that they measure different molecular processes. Chromatin immunuprecipiation measures the promoter state and we measured the production of nascent transcripts by time lapse microscopy. Chromatin immunoprecipitation as a population method detects only promoters that are bound by a specific factor and thus cannot distinguish variability among cells. A conclusion to reconcile regular population and stochastic single cell behaviour might be that between the deterministic epigenetic regulation and actual stochastic transcription are rate limiting steps that mask the regularity at the transcriptional level. For instance, transcription might be slow relative to changes in the epigenetic state. In addition, not every transcription initiation event leads to a complete transcript [Darzacq et al., 2007].

The highly regular transcription cycles were observed for the ps2 gene. Published results on GREB1 transcriptional cycling revealed a less pronounced and less regular cycling [Sun et al., 2007]. Thus, a second explanation that we did not observe cycles in transcription might be that GREB1 does show no cycles or at least less regularity on the promoter level. Our attempts to produce time course ChIP experiments on the GREB1 promoter did not show such highly regular oscillations in contrast to the ps2 gene. After an initial rise in signal with a maximum after 20 minutes the signal decreases again and shoes no further regularity.

3.1.2 The biological nature of the promoter states

The biological nature of the model states, i.e. what characterises the states of the random telegraph model on the chromatin level, is not clear. Transcriptional activity is mostly controlled via the accessibility of the DNA to transcription factors. Transcriptionally active chromatin is open and inactive closed. Opening and closing of chromatin is characterised by the nucleosome density along the chromatin fiber. Two mechanisms are responsible to modulate chromatin accessibility and thus nucleosome density: histone (de)acetylation and nucleosome repositioning. It was found that acetylation of the GREB1 promoter and gene is induced by estrogen stimulation which in turn leads to higher transcription levels [Sun et al., 2007]. Positioning of nucleosomes in the promoter region is responsible for the accessibility of DNA to the binding of transcription factors and RNA polymerase. When the nucleosomes are removed or slid out of the promoter region accessibility is highest. Consistent with the role of nucleosomes for stochastic switching it was found that the accessibility of the yeast gene PHO5 DNA is controlled by nucleosome positions and that the nucleosome configuration changes stochastically leading to stochastic changes in transcriptional activity and transcriptional bursts [Brown et al., 2013]. Nucleosome positioning is performed by so called chromatin remodelling factors for instance the SWI/SNF complex [Medina et al., 2005]. The BRG1 subunit of SWI/SNF showed cyclic engagement to the ps2 promoter [Métivier et al., 2003] (see figure 2.1.2). Assuming a similar mechanism for GREB1 indicates that upon estrogen stimulation the GREB1 gene locus undergoes substantial chromatin modifications leading to pronounced transcriptional activity. It is, however, unclear how to map the many known chromatin modifications to only two levels of transcriptional activity.

As acetylation is involved in GREB1 activation, experiments perturbing the processes of either acetylation or deacetylation by inhibiting the corresponding enzymes would provide further insight into the biological nature of the promoter states. For instance, sodium butyrate is an inhibitor of histone deacetylases, i.e. enzymes that erase acetyl groups from histones [Davie, 2003]. HDAC inhibition by sodium butyrate leads to histone hyperacetylation. It was found that HDAC inhibition in estrogen induced transcription leads to reduced transcriptional activity which was interpreted as an interuption of the proposed transcriptional cycle [Reid et al., 2005]. Fitting the random telegraph model to data from cells treated with sodium butyrate or other inhibitors of epigenetic factors would allow to investigate the influence of epigenetic perturbations on the dynamics of single cell transcription.

We made the simplifying assumption that transcriptional activity is binary, i.e. it either is on or off. Recently Corrigan et al. suggested an alternative interpretation of the modulation of transcriptional activity. Instead of having only few discrete states of activity cells can modulate the transcription over a continuum of states each having slightly different transcriptional activity [Corrigan et al., 2016]. This would leave the cells much more room to fine tune their transcriptional response to environmental changes. It can, however, not resolve the puzzle of reconciling the contrasting behaviour at the population and single cell level. In addition, we did not find evidence for a dose dependent regulation of the transcription initiation rate for GREB1. This further indicates that transcriptional regulation might be highly gene specific.

3.2 Off time regulation by signalling to estrogen responsive elements

We showed that estrogen controls the transcription of GREB1. Static dose response curves exhibited typical sigmoidal behaviour with higher transcription levels at high stimuli. In case of the 5 pM data set almost half of the cells that were imaged in a time resolved manner did not show a response during an observation time of 12 hours (see the three top panels in figure 2.3.2). This ratio changed substantially for higher estrogen concentrations towards smaller fractions of non responding cells. Different hypotheses appear plausible to explain such behaviour: with increasing estrogen stimulus the gene on time could increase, the gene off time could decrease or the initiation rate could increase. Also a combination of the three is possible. Model fitting found that this dose dependent behaviour can be explained by a dose dependent modulation of the promoter off time. The off time posterior distributions showed no overlap for different experimental conditions whereas the posteriors of the on time and initiation rate showed substantial overlap (see figure 2.6.4). Thus, a single stimulus dependent parameter of the model was able to explain a main feature of the data. Accordingly, we found that a simple global model could fit data sets obtained at different estrogen stimuli under the assumption that all parameters except the off time are stimulus independent. A similar observation was made for another steroid hormone. Ponasterone showed a similar dose dependent modulation of the transcriptional off times with long off times at low stimulus levels and vice versa [Larson et al., 2013]. Moreover, ponasterone induced transcription was observed to occur in bursts with the target genes randomly switching between active and inactive states [Larson et al., 2013]. Thus, such a dose dependent promoter off time regulation appears to be common in steroid hormone induced transcription.

3.2.1 The stochastic model combined with an estrogen dependent signalling pathway can predict dose dependent response times

The dose dependency of the gene off time raised the question how the off time is regulated by the estrogen stimulus. Cells have to respond fast to environmental changes, i.e. adapt transcription to the current needs. The random telegraph model with the fitted parameters has a response time of about 30 minutes for the GREB1 gene. The nuclear translocation of the estrogen receptor upon estrogen stimulation takes approximately the same time [Spona et al., 1980]. This timing motivates a combined model of estrogen signalling and estrogen driven transcription to understand the cells response times. To do so we formulated a signalling pathway considering the following mechanisms: Estrogen is a ligand to the estrogen receptor alpha (ER_α). Ligand bound receptors form dimers and directly bind to estrogen responsive elements (ERE) in the genome acting as transcription factors [Kumar and Chambon, 1988]. The GREB1 promoter carries multiple EREs [Sun et al., 2007]. Two EREs lying closest to the transcription start site showed the strongest signal of ER_α binding in a time course ChIP-seq experiment (see figure 2.2.2). Concurrent binding to these two EREs in close proximity to the transcriptional start site was assumed to be sufficient to activate transcription (see figure 2.8.1). This led to the proposition of a two level signalling pathway to explain the off time dose dependency of GREB1. The first level of the signalling pathway represents the ligand binding and dimerization of the receptor and the second level the binding of the receptor dimers to the two major EREs. We tested this hypothetical signalling pathway on data of the accumulation RNA in synchronised cells.

An important question with regard to the signalling path way was how fast cells can react to an estrogen signal. Assumption of the described coarse signalling pathway together with the calibrated stochastic model allowed to predict the accumulation of RNA in synchronised cells. This accumulation was measured by two time course methods i) by RT-qPCR quantifying the amount of RNA in a large population of cells and ii) by imaging single cells. Both methods allowed to estimate the response times of the system by measuring the time the cells needed to achieve the half maximal value of RNA content.

The observed response times were found to consist of two parts: The first part is created by the signalling pathway and describes the time the estrogen signal needs to reach the promoter. This signalling response time is dose dependent with short times at high stimuli. The second part of the total response time is created by the stochastic model. This part describes the time the cells need to switch from the inactive state into the active and start transcribing. The response time of the stochastic model is dose independent whereas final amount of RNA depends on the stimulus. The response time of the stochastic model is rate limiting in that is much longer than the response time of the signalling pathway at high stimuli. At low stimuli both parts of the total response time are approximately equal.

Treatment of the cells with antiestrogens like ICI or tamoxifen would allow to further test the hypothesised signalling pathway. For instance, titration of estrogen and an antiestrogen in a two dimensional dose response and subsequent measurement of the cells response times would allow to assess the signalling dynamics.

3.3 Transcriptional heterogeneity

3.3.1 Cell to cell variability

Stochastic simulations can reproduce the inherent stochasticity of transcription. But when implementing the stochastic model we found to describe the observed data it was necessary to incorporate cellular heterogeneity. Heterogeneity means that cells under the same experimental conditions can exhibit a wide variability in behaviour. This variability to a certain extent stems from the stochastic dynamics alone which we refer to as intrinsic noise. But this variability was not sufficient to explain the experimental data. We found that kinetic parameters like the initiation rate and the polymerase velocity vary from cell to cell (see figures 2.4.5 and 2.4.6). Kinetic parameters of the stochastic models are linked to the internal state of the cell via protein concentrations that can show substantial variability among cells. An observation known as extrinsic noise. Cell to cell variability was incorporated in the stochastic model by resampling or perturbing model parameters or combinations of two parameters. Model selection favoured models incorporating a combined parameter perturbation, namely of the RNA polymerase velocity and the transcription initiation rate.

RNA polymerase velocity describes how fast new nucleotides are added to the elongating nascent transcript. Transcription by RNA polymerase is dependent on the supply with energy by ATP molecules. It was found that cells with higher mitochondrial mass exhibit higher transcriptional activity than cells with lower mitochondrial mass relating the energy metabolism with transcription. Varying numbers of mitochondria are due to stochastic segregation during mitosis [das Neves et al., 2010; Johnston et al., 2012]. In general, partitioning of cell organelles and molecular components into two daughter cells during mitosis is highly stochastic. Many molecules in a cell are present only in low copy numbers causing high relative fluctuations after segregation [Huh and Paulsson, 2011]. Such fluctuations provide a strong source of extrinsic noise and are an explanation for varying initiation rates.

For future work the individual contributions of intrinsic and extrinsic variability are interesting to study. This can be done with a cell line carrying the fluorescent reporter construct in two alleles of the same gene. This would allow to study the stochastic fluctuations of two transcription sites integrated into the same cellular environment, i.e. that share the same extrinsic but not intrinsic sources of noise. Correlations between both signals would indicate that they are not independent from each other and thus driven by the same extrinsic noise sources. Moreover, such a system allows to study the effect of environmental conditions on both alleles [Bar-Even et al., 2006; Hilfinger and Paulsson, 2011; Rinott et al., 2011].

3.3.2 Consequences of transcriptional variability

Given that GREB1 is transcribed in stochastic bursts the number of transcripts per cell is expected to be highly variable among cells. The coefficient of variation as the ratio between the standard deviation of transcripts per cell and the average number provides a measure of transcriptional noise. The transcriptional noise in the observed populations of single cells was found to decrease with the stimulus, i.e. coefficient of variation decreased with increasing estrogen concentrations from a value of approximately two at 5 pM to a value below one at 1000 pM. These numbers indicate very high variability of the number of nascent transcripts per cell.

The strong noise at the level of nascent transcripts raises the question how pronounced is the variability at the levels processed transcripts or proteins. The half life of GREB1 transcripts in mouse embryonic stem cells was found to be 4.4 hours [Sharova et al., 2009] which is shorter than the almost 12 hours promoter off time that was found at 5 pM stimulus. Thus, assuming a similar half life of GREB1 transcripts in MCF-7 cells the number of transcripts per cell at low stimuli can be expected to be more variable. At high stimuli the off time is much shorter than the transcript half life which should dampen transcriptional noise [Friedel et al., 2009].The transmission of noise from the RNA level to the protein level mostly depends on protein degradation, i.e. the protein's half life. Short lived proteins will follow the temporal RNA variability more closely than long lived ones [Raj et al., 2006]. Thus, assuming a short GREB1 protein half life the highest variability in protein copy number would be expected at low estrogen concentrations.

GREB1 is essential for progression of estrogen dependent breast cancers [Rae et al., 2005; Hodgkinson and Vanderhyden, 2014; Laviolette et al., 2014]. Thus, variability in GREB1 protein copy number is expected to have an effect on breast cancer tumor progression. Estrogen signalling can be inhibited by antiestrogens like ICI or tamoxifen which are common treatments against breast cancer. ICI prevents the shuttling of ER_α between cytoplasma and nucleus [Dauvois et al., 1993] and thus binding of the receptor to DNA to acitvate transcription. Even at low estrogen concentrations we observed a substantial fraction of cells showing transcription. Moreover, transcriptional variability at 5 pM estrogen stimulus was highest. Thus, we suspect to suppress GREB1 transcription completely is unlikely as a significant number of outlier cells still transcribing GREB1 would be expected. Indeed it was found that individual patients exhibit highly variable responses to tamoxifen treatment. Moreover, the response to antiestrogens can be highly nonlinear and thus requires individual treatment [Lebedeva et al., 2012]. Thus, cell to cell variability is an important point in cancer treatment [Sun and Yu, 2015].

3.4 The SMC ABC algorithm

We calibrated stochastic models to single cell transcription data using an likelihood free bayesian approach. The length of the GREB1 transcript made it necessary to introduce extra states representing transcript elongation into the model to include delays between the actual transcription and the observed signal. Those extra state extended the model's state space enormously and prohibited the calculation of the likelihood of parameters θ given data D. Approximate Bayesian Computation (ABC) is a powerful method to calibrate models where a likelihood cannot be calculated or is computationally very expensive [Tavaré et al., 1997; Beaumont et al., 2002; Sunnåker et al., 2013]. Moreover, ABC allows to directly integrate model selection by treating the model itself as a variable [Toni et al., 2009].

The choice of a distance measure $\rho(D_{sim}, D_{exp})$ that describes the similarity of simulated and experimental data is crucial for the application of ABC methods. This is especially important in the case of model selection [Marin et al., 2014]. We manually selected four features of the experimental data that we compared with simulations to estimate the similarity between data and simulations. In addition to these four features we used a multivariate statistic capable to compare high dimensional distributions. In principle this statistic as distance metric should be sufficient for model fitting [Loos et al., 2015] but inclusion of the four other features increased the quality of the fits substantially in that it yielded more narrowly defined posterior distributions. The number of features could be increased but we found the quality of fit and moreover, the model selection qualities of the algorithm sufficient.

The iterative Sequential Monte Carlo Approximate Bayesian (SMC ABC) algorithm as a particle based Monte Carlo method [Sisson et al., 2007] in addition to the distance measure has two further crucial points: i) the creation of new particles in each iteration and ii) the setting of a threshold schedule ϵ_t for the subsequent iterations. For particle creation we used log normal proposal distributions in accordance with Zoller et al. as this was successful in a similar analysis [Zoller et al., 2015]. The shape parameter of those distributions were fixed and adjusted manually prior to model fitting. It is, however, possible to adopt the proposal distributions to the current generation of particles which should lead to a more effective search in the parameter space [Filippi et al., 2013] and probably more effective convergence.

The threshold schedule ϵ_t defines the level of the distance measure in the tth iteration below which particles are accepted and used for the next iteration. We used an adaptive method by accepting always the best 20% of the particles and did not predefine a fixed schedule. Such an adaptive percentile based method proved useful in different settings [Moral et al., 2011; Drovandi and Pettitt, 2011], although, algorithm convergence may depend on the choice of the percentile [Silk et al., 2013].

Altogether we showed by algorithm benchmarking that the chosen settings for the SMC ABC algorithm are sufficient to fit stochastic models to the given time course data. Model selection for three out of four data sets found reproducibly a unique model. Both facts underline the applicability and numerical stability of the implemented SMC ABC algorithm.

4 Methods

All modelling and data analysis were implemented in the python programming language [Van Rossum and Drake, 2001] in combination with the IPython package [Perez and Granger, 2007]. Furthermore the NumPy [van der Walt et al., 2011] , SciPy [Jones et al.], scikit-learn [Pedregosa and Varoquaux, 2011] and Matplotlib [Hunter et al., 2007] libraries for scientific computation and visualisation were used.

4.1 Extending the cost function to experimental errors to fit reaction blocks

For each factor the corresponding reaction block was fitted independent of the other factors. A reaction block is the subset of states within the states of the transcription cycle where the factor is present at the promoter of the gene. Such a block is described by three parameters: start, end and amplitude. Start and end denote the promoter state where the factor binds or releases, respectively. The amplitude of the block describes the fraction of promoters within the population contributing to the measured ChIP signal.

Data for the factors listed in table 2.1.2 was taken from Metivier et al [Métivier et al., 2003] where no experimental error range is provided. Thus, we included the experimental error σ as free parameter into model fitting. Assuming a gaussian distributed experimental error allows maximum likelihood estimation of the model parameters. Including the the experimental error, however, led to a change in the usual χ^2 cost function. The likelihood of a parameter vector θ given the data D is:

$$L(\theta, D) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i}} \cdot e^{-\frac{(y_i^D - y_i^M)^2}{\sigma_i^2}} \tag{4.1.1}$$

y_i^D and y_i^M are the experimental data points and the corresponding simulated points of the model, respectively. Minimising the negative log likelihood while considering σ_i to be part of the parameter vector θ led to the following cost function

$$cost = \sum_i \frac{(y_i^D - y_i^M)^2}{\sigma_i^2} + \ln \sigma_i \tag{4.1.2}$$

The second term balances the size of the experimental error against the ability of the model to explain the data. It was assumed that σ_i is a sum of a relative error a and an absolute error b.

$$\sigma_i = a y_i^D + b \tag{4.1.3}$$

Including the experimental error increased the number of free parameters per factor to five. Each factor was fitted independently. A temporal order of the binding events of the individual factors was not assumed prior to model fitting. By applying latin hypercube [McKay et al., 1979] sampling multiple starting points for optimisation were created which were then used for local optimisation. As local optimisation algorithm the SLSQP method [Kraft, 1988] as implemented in the SciPy library was used. Bounds were defined for each parameter to avoid biologically unplausible results (see table 2.1.1). A block has to start at or after the first state of a cycle. For the end, in addition, a cyclic boundary was defined, i.e. when the end of a cycle was found at a state index higher than the number of states in the cycle it was shifted to the next cycle accordingly. This avoids to define a special state that marks the beginning of a cycle.

4.2 The stochastic simulation algorithm for single cell transcription

To simulate the stochastic behaviour of the gene promoter the well know stochastic simulation algorithm (SSA) was utilised as introduced by D. Gillespie [Gillespie, 1977, 1976] (see algorithm 1). The SSA allows to numerically create exact realisations of system paths and thus to investigate system behaviour for various parameter sets. Here the direct method of the SSA was applied. Due to the binary nature of the occupancy of the individual promoter states and the cyclic topology of the reaction network it was not possible to implement a computationally faster but approximate variant of the SSA like $\tau-$leaping or alike [Gillespie, 2001, 2007].

Algorithm 1: Stochastic simulation algorithm

```
Initialise: $t = 0$, simulation time $t_f$; rate constants $k_m$, $k_i \ldots$, initial state $x$;
while $t \leq t_f$ do
    calculate hazards $h_i(x, k_i)$ and combined hazard $h_0 = \sum h_i(x, k_i)$
    sample two random numbers $r_1$, $r_2 \sim Unif(0, 1)$
    Time to next event: $t' = 1/h_0 \cdot \ln(1/r_1)$
    Next reaction: smallest index $j$: $r_2 \cdot h_0 \geq \sum_{i=0}^{j} h_i(x, k_i)$
    Update state according to reaction $j$
    set $t = t + t'$
end
```

The SSA by sampling random numbers simulates the time to the next reaction event and the actual reaction that is occurring. Both of which depend on the current state of the system which reflects the markovian property. Reaction hazards $h(x, k_i) = c_i k_i$ in the system investigated here correspond to the reaction rates k_i and not on the abundance c_i of the different species. The promoter can only be in one of its states at a time (i.e. $c_i = 0, 1$) and progresses with the rate associated to this state. RNA production is not dependent on the present number of RNA molecules and occurs with rate k_m when the promoter is in an active state.

4.3 Intensity of single transcripts and noise model

To connect experimental data with the stochastic model a noise model was calibrated. Fluorescence intensity was assumed to be proportional to the number of transcripts at the transcription site. The factor of proportionality α is the fluorescence intensity of a single transcript which was measured independently. First the linear relation between illumination intensity and detected fluorescence light intensity was calibrated. Secondly at high illumination intensities it was possible to detect and measure the fluorescence light from single transcripts diffusing in the nucleoplasm. With the linear relation estimated before it was possible to rescale the signal from single transcript to the illumination settings used for long term live microscopy. By this a factor of $\alpha = 32.8$ was estimated for the used imaging set up.

As a measure of background light intensity a site within the cell nucleus distant from the transcription site but of similar size was tracked in each image. A log normal distribution was fitted to the distribution of the gained background intensity values as

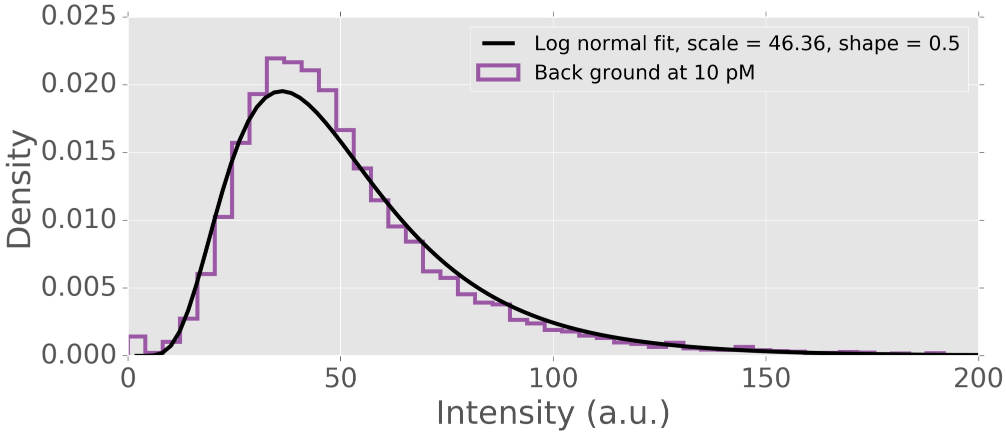

Figure 4.3.1: The background signal follows a log normal distribution. Global distribution (blue) of measured background intensities at 10 pM stimulus. A fit of a lognormal distribution is shown in black wth the corresponding parameters in the legend.

shown in figure 4.3.1. The fitted parameters of the distribution of the background signal were similar under all experimental conditions (see table 4.3.1). With the so calibrated noise model the probability to observe signal S given M transcripts at the transcription site is given by:

$$\begin{aligned} P(S|M) &= \alpha \cdot M + noise \\ noise &\sim LN(scale, shape) \end{aligned} \tag{4.3.1}$$

Where the noise is sampled from a lognormal distribution with the scale and shape parameter obtained by fitting. For sampling the random number generator implemented in the SciPy package was used.

Stimulus	Scale	Shape
5 pM	48.7	0.5
10 pM	46.4	0.5
20 pM	47.3	0.5
1000 pM	43.5	0.5

Table 4.3.1: Parameters of the fitted noise model model are similar for different experimental conditions. Parameters of the fitted lognormal distribution of the back ground signal for different experimental conditions

4.4 Model parametrisation

Model parameterisation closely follows the approach of Zoller et al [Zoller et al., 2015]. The transcriptional model as shown in figure 2.4.1 is parameterised by the number of active states N, the number of inactive states M, the perturbation index p, the initiation rate k_m, the transition rates k_{on_i} of the active states and the transition rates k_{off_i} of the inactive states. The first three parameters describe the model structure and the remaining parameters the kinetics.

A more practical approach to parametrise the kinetic part can be expressed by the parameters b, t_{on}, T, π_i, p_j. Where $b = k_m t_{on}$ is the burst size, t_{on} and T are the total on and off times respectively. $\pi_i = t_{on_i}/t_{on}$ and $p_j = t_{off_i}/T$ are the fractions of time spent in the ith active or jth inactive state. Therefore the total on and off times are given by: $t_{on} = \sum t_{on_i} = \sum 1/k_{on_i}$ and $T = \sum t_{off_j} = \sum 1/k_{off_j}$. The fractions π_i and p_j follow the constraint $\sum \pi_i = \sum p_j = 1$

Due to the symmetry of the promoter cycle the order of states within either the active or inactive phase is not identifiable. Therefore it is only possible to recover distributions of the total time spent in either phase because that distribution is the convolution of the distributions of the life times of the individual states. Since convolution is symmetric the order of states within the active or inactive promoter phase cannot be resolved.

To take the symmetry of the promoter cycle into account an ordering condition on the π_i and the p_i is assumed: $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_N$; $p_1 \geq p_2 \geq \cdots \geq p_M$. Such an ordering can be achieved by the following parametrisation:

$$\begin{aligned} p_2 &= u_1 p_1 \\ p_3 &= u_2 p_2 = u_1 u_2 p_1 \\ &\dots \\ p_M &= u_{M-1} p_{M-1} = p_1 \prod_{j=1}^{M-1} u_j \end{aligned} \tag{4.4.1}$$

where $u_j \in [0,1]$. Together with the constraint $\sum p_j = 1$ the number of free parameters u_j and p_j is the same and the latter can be expressed by the former. With the system of equations:

$$\begin{aligned} p_M &= 1 - \sum_{j=1}^{M-1} p_j = 1 - p_1\left(1 + \sum_{l=1}^{M-2} \prod_{j=1}^{l} u_j\right) \\ p_M &= p_1 \prod_{j}^{M-1} u_j \end{aligned} \tag{4.4.2}$$

p_1 can be expressed as:

$$p_1 = \frac{1}{1 + \sum_{l=1}^{M-1} \prod_{j=1}^{l} u_j} \tag{4.4.3}$$

With equation 4.4.1 all other p_j can be calculated as:

$$p_k = \frac{\prod_{j=1}^{k-1} u_j}{1 + \sum_{l=1}^{M-1} \prod_{j=1}^{l} u_j} \tag{4.4.4}$$

An equivalent parametrisation was used for the π_i.

4.5 Approximate Bayesian Computation

Bayesian model fitting in contrast to usual frequentist maximum likelihood estimations do not only provide a point estimation and a confidence interval of the most likely parameters but a full posterior distribution. Posterior distributions of model parameters given observed data $P(\theta|D)$ can be obtained by applying Bayes rule:

$$P(\theta|D) = \frac{P(D|\theta)\pi(\theta)}{P(D)} \tag{4.5.1}$$

It is, however necessary to calculate the likelihood of the data given the parameters $P(D|\theta)$. For complex models this can be computationally costly to calculate if not even mathematically impossible. For the models used here the dimensionality of the state space can be extremely large due to the combinatorial possibilities to distribute RNA molecules over the extra states that describe RNA elongation. This prohibited the calculation of a likelihood and required the application of a likelihood free model fitting approach.

Forward simulation of the system for given parameters θ yields a synthetic data D_{sim} set which can be compared to the experimental data set D_{exp} by a distance measure

$\rho(D_{sim}, D_{sim})$. If ρ is sufficiently small θ is accepted and thus an approximate posterior distribution can be yielded $P(\theta|\rho(D_{sim}, D_{exp}) \leq \epsilon)$. This obviously depends on the desired distance ϵ. If ϵ is large many proposed parameter θ will be accepted but the approximation of the posterior will be bad. A smalll ϵ in contrast leads to a low acceptance rate requiring a large number of simulations.

Algorithm 2: ABC rejection algorithm

```
Initialise: Population size $N$; counter $n = 0$;
while $n \leq N$ do
    Sample model $m \sim \pi(m)$
    Sample $\theta' \sim \pi(\theta|m)$
    Simulate a data set $D_{sim} \sim f(D|\theta')$
    If $\rho(D_{sim}, D_{exp}) \leq \epsilon$ accept $\theta'$ and set $n = n + 1$
end
```

Here simple rejection ABC as described in algorithm 2 was used to create a large start population of candidate parameter sets. From the parameter priors 50000 candidates were drawn and simulated mRNA counts were saved. For each data set the best 2000 candidates were selected by adding noise according to the noise model to the candidates and subsequent distance calculation. The population found this way was then used as start population for the Sequential Monte Carlo Approximate Bayesian Computation (SMC ABC) algorithm described in the next section.

4.5.1 Sequential Monte Carlo Approximate Bayesian Computation

Sequential Monte Carlo Approximate Bayesian Computation provides a way to sequentially approach the true posterior distribution via a sequence of intermediate distributions. In each iteration of the algorithm 3 the current allowed distance measure ϵ_t is decreased. The first iteration consists of the simple rejection algorithm creating a start population of candidate particles. A particle consists of a weight, a model index and the corresponding model parameters. The 20% best particles are taken to the next iteration and the most distant of the 20% best defined the distance measure ϵ_t of the current round. The population is filled up again by new particles created out of the existing particles by proposal distributions. Iterating in this way in each round the approximation of the posterior distribution is improved and getting closer to the true posterior.

Algorithm 3: Sequential Monte Carlo Approximate Bayesian Computation

```
Initialise:Population size $N$, data $D_{exp}$, iteration counter $t = 0$, Maximal number
  of iterations $T$, $try = 0$, maximal number of tries to create new particles $try_{max}$,
  stop criterion $\epsilon_{stop}$;
if $t = 0$ then
    sample $m$ and $\theta'$ from priors $\pi(m)$ and $\pi(\theta|m)$
    set all weights to $w_{t=0} = 1/N$
else
    sample particle $(m, \theta)$ from the best 20% particles of the previous population
      with weights $w_{t-1}$
    Perturb $(\theta, m)$ to obtain $(\theta', m') \sim K(\theta', m'|\theta, m)$
    $try = try + 1$
    Simulate a candidate data set $D'_{sim} \sim f(D'|\theta', m')$
    if $\rho(D'_{sim}, D_{exp}) \geq \epsilon_t$ & $try \leq try_{max}$ then
        create new candidate particle
    else
        add $(\theta', m')$ to the population of particles and calculate its weight as:

            $$w_t^{(i)} = \frac{\pi(\theta')}{\sum_{j=1}^{N} w_{t-1}^{(j)} K(\theta_{t-1}^{(j)}|\theta')}$$

    end
    If $i < N$ set $i = i + 1$ and sample a new particle from the previous population
end
Normalize the weights
if $max\{\rho_i(D_{sim}^i, D_{exp})\} \leq \epsilon_{stop}$ then
    stop
else
    start new iteration
end
if $t < T$ then
    set $t = t + 1$
else
    stop
end
```

The algorithm comes to halt if it either reaches a predefined stop distances, i.e. all particles are below that distance, or if improvement in distance between consecutive iterations is smaller than 5%.

4.5.2 Model and parameter prior distributions

As with all bayesian approaches a prior belief about the model topology and the corresponding parameters has to be specified. The full prior is a multivariate distribution of all parameters that is difficult to sample from. A common way to circumvent this is to factorise the full prior of the system into a product of priors of the individual parameters. In addition to the parameters a prior on the model m itself has to be specified as well.

$$\pi(m, \theta_m) = \pi(m)\pi(k_m)\pi(\tau)\pi(T)\pi(\mu_i|m)\pi(u_j|m) \tag{4.5.2}$$

For the model itself and the parameters μ and u_i that describe the fraction of time that the system spends in individual states of the on or off phase and the strength of the parameter perturbation an uniform prior was used. The prior initiation rate was sampled from a lognormal distribution. As prior distributions for the total promoter on and off times exponential distributions were used. All priors and the corresponding parameterisations are summarised in table 4.5.1.

Parameter	**Prior**	**Parametrisation**
Model	Uniform	$m \sim Unif(1, 40)$
Initiation rate	Lognormal	$k_m \sim LN(5, 0.8)$
On time	Exponential	$\tau \sim Exp(-1/10)$
Off time	Exponential	$T \sim Exp(-1/70)$
Relative duration of single on state	Uniform	$\mu \sim Unif(0.8, 1)$
Relative duration of individual off state	Uniform	$u_i \sim Unif(0.8, 1)$
Parameter perturbation strength	Uniform	$\sigma_{(k_m,\tau,T)} \sim Unif(1, 8)$

Table 4.5.1: Prior distributions of model and parameter values. The full prior was factorised according to equation 4.5.2 and the table lists the individual factors and the corresponding parameterisation.

4.5.3 Proposal distributions to create new candidate particles

To explore the model and parameter space new particles have to be created out of accepted particles. This was done by so called proposal distributions that randomly change the position of one particle in the model and parameter space. To move a particle in model space only certain jumps are allowed depending on the current position. These jumps are depicted in figure 4.5.1. The left hand side of the figure displays the jumps in model topology and the right hand side the jumps in perturbations. Each topology could be combined with each perturbation. Thus, for each position in model space a certain set of moves is allowed. In each iteration to move a particle in model space a uniform random number $r \in [0, 1]$ was drawn. If $r > 0.6$ a new model was sampled uniformly from transitions allowed for the current model.

If a new particle had less promoter states after the proposed move than before the fastest states were omitted, i.e. the corresponding μ or u_i were deleted from the parameter vector. In the opposite case, when new states had to be added new values for μ or u_i were sampled from their prior. The parameters describing the strength of perturbations for the initiation rate, the on time and the off time $\sigma_{(k_m,\tau,T)}$ where proposed with a normal distribution $N(\sigma'_{(k_m,\tau,T)}, 1)$ in case when the perturbation after the model change remained the same. In case the perturbation changed a new $\sigma_{(k_m,\tau,T)}$ was sampled from its prior.

Proposal distributions of the kinetic parameters again closely follows the work of Zoller et al [Zoller et al., 2015]. The parameters b, τ and T where changed with a lognormal distribution $LN(\theta; \theta', \sigma_\theta)$ with individual scale parameters σ_θ. As proposal distribution for the u_i a beta distribution $q_\beta(u, \alpha(u'), \beta(u'))$ was used. Where the parameters of the distribution are defined as: $\alpha(u') = 1 + \lambda u'$ and $\beta(u') = 1 + \lambda(1 - u')$ with $\lambda = 2$. A value of two proved useful in benchmark tests of the algorithm.

4.5.4 Maximum mean discrepancy to compare multivariate distributions

Maximum mean discrepancy is a multivariate statistic that allows to compare high dimensional distributions p and q [Gretton et al., 2012] and is defined as.

$$MMD[F, p, q] := \sup_{f \in F} \Big(E_p[f(p)] - E_q[f(y)] \Big) \tag{4.5.3}$$

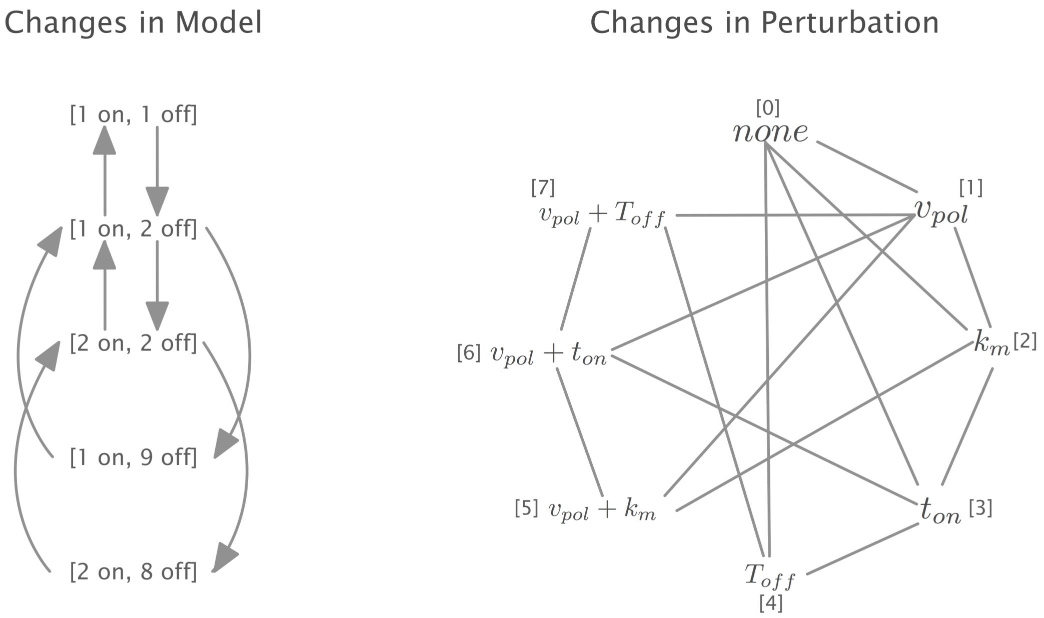

Figure 4.5.1: Network of allowed moves in model space to propose new particles. On the left the considered model topologies are shown, depicted by their number of on and off states. Arrows indicate allowed moves. The right illustration shows allowed moves in parameter perturbations. There are in total eight different perturbations indicated by a perturbation index ranging from 0 to 7 (see table 2.5.3 for details, perturbation indices are given in square brackets). Each line indicates an allowed jump from one perturbation to another. Jumps are allowed in both directions with equal probabilities. Arrow heads are omitted for better visibility. Each perturbation could be combined with each model topology resulting in 40 models in total. Each model is associated with a certain set of allowed moves that was used to explore the model space.

If p and q are equal, MMD is zero. F is a class of functions $f : \chi \longrightarrow \mathbb{R}$ as the unit ball in a universal reproducing kernel Hilbert space H. For samples $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_m)$ of p and q an empirical estimation of the MMD is given by

$$MMD[F, X, Y] := \sup_{f \in F} \Big(\frac{1}{n} \sum_{i=1}^{n} f(x_i) - \frac{1}{m} \sum_{j=1}^{m} f(y_j) \Big) \tag{4.5.4}$$

A kernel $k(x, y) = \Phi(x)^T \Phi(y)$ allows to rewrite MMD by the mean embedding $\mu_p := E_p[\Phi(x)]$ as $MMD[F, p, q] = \sup_{f \in F} \Big\langle \mu_p - \mu_q, f \Big\rangle = \|\mu_p - \mu_q\|_H$. With $\mu_X = \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i)$ and $k(x, y) = \langle \Phi(x), \Phi(y) \rangle$ the empirical estimation of the MMD can be obtained via

$$MMD[F,X,Y] = \left(\frac{1}{n^2}\sum_{i\neq j}^{n} k(x_i,x_j) + \frac{1}{m^2}\sum_{i\neq j}^{m} k(y_i,y_j) - \frac{2}{nm}\sum_{i,j=1}^{n,m} k(x_i,y_j)\right)^{\frac{1}{2}} \tag{4.5.5}$$

A working python script to calculate MMD between two data sets was published by Vincent Van Ash alongside to his PhD thesis [Van Asch, 2012]. This implementation was used here. In addition the MMD as single distance measure was used by Loos et al to estimate model parameters in an SMC ABC setting from time course measurements [Loos et al., 2015].

4.5.5 Bayes factors as measure of model selection

In contrast to a frequentist approach of model selection the Bayes Factor allows not only say if the null hypothesis has to be rejected but instead can be used to argue in favour for a model m. A Bayes Factor between two models m_1 and m_2 with given data D and posterior distributions $P(m_{1,2}|D)$ is calculated as:

$$B_{1,2} = \frac{P(m_1|D)}{P(m_2|D)} \cdot \frac{\pi(m_2)}{\pi(m_1)} \tag{4.5.6}$$

In case of uniform priors as we used here on the model the second term cancels reducing the Bayes factor to the simple ratio of the two models in the posterior. According to Kass & Raftery a Bayes factor of three and above is considered to be positive in favour to model m_1 and agains m_2 [Kass and Raftery, 1995]. Higher values indicate stronger selection.

5 Appendix

5.1 Python code for model fitting

Attached to this thesis is a CD containing python scripts and IPython notebooks that were used to analyse data, calibrate models and to generate the presented figures. In addition, this CD contains the measured imaging time course data used for model calibration and RT-qPCR time course data used to estimate response times. Alternatively, the code can be downloaded from GitHub under:

`https://github.com/baumgast/gene_transcription_SMC_ABC`

References

B. Alberts, A. Johnson, J. Lewis, D. Morgan, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, 6 edition, 2014. ISBN 9780815344322.

M. Ardehali and J. T. Lis. Tracking rates of transcription and splicing in vivo. *Nature structural & molecular biology*, 16(11):1123–1124, 2009. doi: 10.1038/nsmb1109-1123.

A. Bar-Even, J. Paulsson, N. Maheshri, M. Carmi, E. K. O'Shea, Y. Pilpel, and N. Barkai. Noise in protein expression scales with natural protein abundance. *Nature Genetics*, 38(6):636–643, 2006. doi: 10.1038/ng1807.

R. D. Barber, D. W. Harmer, R. a. Coleman, and B. J. Clark. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiological genomics*, 21(3):389–95, 2005. doi: 10.1152/physiolgenomics.00025.2005.

M. A. Beaumont, W. Zhang, and D. J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–35, dec 2002. doi: GeneticsDecember1, 2002vol.162no.42025-2035.

M. D. Biggin. Animal transcription networks as highly connected, quantitative continua. *Developmental cell*, 21(4):611–26, oct 2011. doi: 10.1016/j.devcel.2011.09.008.

S. Boireau, P. Maiuri, E. Basyuk, M. de la Mata, A. Knezevich, B. Pradet-Balade, V. Bäcker, A. Kornblihtt, A. Marcello, and E. Bertrand. The transcriptional cycle of HIV-1 in real-time and live cells. *The Journal of cell biology*, 179(2):291–304, oct 2007. doi: 10.1083/jcb.200706018.

C. R. Brown, C. Mao, E. Falkovskaia, M. S. Jurica, and H. Boeger. Linking stochastic fluctuations in chromatin structure and gene expression. *PLoS biology*, 11(8):e1001621, 2013. doi: 10.1371/journal.pbio.1001621.

C. Carlberg. The impact of transcriptional cycling on gene regulation. *Transcription*, 1 (1):13–16, jan 2010. doi: 10.4161/trns.1.1.11984.

M. Chalfie, Y. Tu, G. Euskirchen, W. W. Ward, and D. C. Prasher. Green fluorescent protein as a marker for gene expression. *Science*, 263(5148):802–805, feb 1994.

H. H. Chang, M. Hemberg, M. Barahona, D. E. Ingber, and S. Huang. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194): 544–7, may 2008. doi: 10.1038/nature06965.

L. T. Chow, R. E. Gelinas, T. R. Broker, and R. J. Roberts. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell*, 12(1):1–8, sep 1977. doi: 10.1016/0092-8674(77)90180-5.

J. R. Chubb, T. Trcek, S. M. Shenoy, and R. H. Singer. Transcriptional pulsing of a developmental gene. *Current biology : CB*, 16(10):1018–25, may 2006. doi: 10.1016/j.cub.2006.03.092.

A. A. Cohen, N. Geva-Zatorsky, E. Eden, M. Frenkel-Morgenstern, I. Issaeva, A. Sigal, R. Milo, C. Cohen-Saidon, Y. Liron, Z. Kam, L. Cohen, T. Danon, N. Perzov, and U. Alon. Dynamic proteomics of individual cancer cells in response to a drug. *Science (New York, N.Y.)*, 322(5907):1511–6, dec 2008. doi: 10.1126/science.1160165.

S. Comşa, A. M. Cîmpean, and M. Raica. The Story of MCF-7 Breast Cancer Cell Line: 40 years of Experience in Research. *Anticancer research*, 35(6):3147–54, jun 2015.

A. M. Corrigan, E. Tunnacliffe, D. Cannon, and J. R. Chubb. A continuum model of transcriptional bursting. *eLife*, 5:1–38, feb 2016. doi: 10.7554/eLife.13051.

A. Coulon, M. L. Ferguson, V. de Turris, M. Palangat, C. C. Chow, and D. R. Larson. Kinetic competition during the transcription cycle results in stochastic RNA processing. *eLife*, 3:1–22, oct 2014. doi: 10.7554/eLife.03939.

J. A. Dahl and P. Collas. A rapid micro chromatin immunoprecipitation assay (microChIP). *Nature protocols*, 3(6):1032–45, jan 2008. doi: 10.1038/nprot.2008.68.

K. Dahlman-Wright, V. Cavailles, S. a. Fuqua, V. C. Jordan, J. a. Katzenellenbogen, K. S. Korach, A. Maggi, M. Muramatsu, M. G. Parker, and J.-A. Gustafsson. International

Union of Pharmacology. LXIV. Estrogen receptors. *Pharmacological reviews*, 58(4): 773–81, dec 2006. doi: 10.1124/pr.58.4.8.

X. Darzacq, Y. Shav-Tal, V. de Turris, Y. Brody, S. M. Shenoy, R. D. Phair, and R. H. Singer. In vivo dynamics of RNA polymerase II transcription. *Nature structural & molecular biology*, 14(9):796–806, sep 2007. doi: 10.1038/nsmb1280.

R. P. das Neves, N. S. Jones, L. Andreu, R. Gupta, T. Enver, and F. J. Iborra. Connecting variability in global transcription rate to mitochondrial variability. *PLoS biology*, 8 (12):e1000560, 2010. doi: 10.1371/journal.pbio.1000560.

S. Dauvois, R. White, and M. G. Parker. The antiestrogen ICI 182780 disrupts estrogen receptor nucleocytoplasmic shuttling. *Journal of cell science*, 106 (Pt 4:1377–88, dec 1993.

J. R. Davie. Inhibition of histone deacetylase activity by butyrate. *The Journal of nutrition*, 133(7 Suppl):2485S–2493S, jul 2003.

P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 68(3):411–436, 2006. doi: 10.1111/j.1467-9868.2006.00553.x.

C. C. Drovandi and A. N. Pettitt. Estimation of parameters for macroparasite population evolution using approximate bayesian computation. *Biometrics*, 67(1):225–33, mar 2011. doi: 10.1111/j.1541-0420.2010.01410.x.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic gene expression in a single cell. *Science (New York, N.Y.)*, 297(5584):1183–6, 2002. doi: 10.1126/science.1070919.

a. M. Femino. Visualization of Single RNA Transcripts in Situ. *Science*, 280(5363): 585–590, apr 1998. doi: 10.1126/science.280.5363.585.

S. Filippi, C. P. Barnes, J. Cornebise, and M. P. H. Stumpf. On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. *Statistical Applications in Genetics and Molecular Biology*, 12(1):87–107, 2013. doi: 10.1515/sagmb-2012-0069.

C. C. Friedel, L. Dölken, Z. Ruzsics, U. H. Koszinowski, and R. Zimmer. Conserved principles of mammalian transcriptional regulation revealed by RNA half-life. *Nucleic acids research*, 37(17):e115, sep 2009. doi: 10.1093/nar/gkp542.

N. Geva-Zatorsky, N. Rosenfeld, S. Itzkovitz, R. Milo, A. Sigal, E. Dekel, T. Yarnitzky, Y. Liron, P. Polak, G. Lahav, and U. Alon. Oscillations and variability in the p53 system. *Molecular systems biology*, 2:2006.0033, jan 2006. doi: 10.1038/msb4100068.

D. T. Gillespie. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *Journal of computational physics*, 434: 403–434, 1976.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of computational physics*, 93555(1):2340–2361, 1977.

D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716, 2001.

D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual review of physical chemistry*, 58:35–55, jan 2007. doi: 10.1146/annurev.physchem.58.032806.104637.

I. Golding, J. Paulsson, S. M. Zawilski, and E. C. Cox. Real-time kinetics of gene activity in individual bacteria. *Cell*, 123(6):1025–36, dec 2005. doi: 10.1016/j.cell.2005.09.031.

E. L. Greer and Y. Shi. Histone methylation: a dynamic mark in health, disease and inheritance. *Nature reviews. Genetics*, 13(5):343–57, may 2012. doi: 10.1038/nrg3173.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schoelkopf, and A. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

C. V. Harper, B. Finkenstädt, D. J. Woodcock, S. Friedrichsen, S. Semprini, L. Ashall, D. G. Spiller, J. J. Mullins, D. A. Rand, J. R. E. Davis, and M. R. H. White. Dynamic analysis of stochastic transcription cycles. *PLoS biology*, 9(4):e1000607, apr 2011. doi: 10.1371/journal.pbio.1000607.

A. Hilfinger and J. Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108(29):12167–72, jul 2011. doi: 10.1073/pnas.1018832108.

K. M. Hodgkinson and B. C. Vanderhyden. Consideration of GREB1 as a potential therapeutic target for hormone-responsive or endocrine-resistant cancers. *Expert opinion on therapeutic targets*, 18(9):1065–76, 2014. doi: 10.1517/14728222.2014.936382.

D. Huh and J. Paulsson. Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences of the United States of America*, 108(36):15004–9, sep 2011. doi: 10.1073/pnas.1013171108.

J. Hunter, D. Dale, and M. Droettboom. Matplotlib, 2007.

F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of molecular biology*, 3:318–56, jun 1961. doi: 10.1016/S0022-2836(61)80072-7.

M. Jeschke, S. Baumgärtner, and S. Legewie. Determinants of cell-to-cell variability in protein kinase signaling. *PLoS computational biology*, 9(12):e1003357, jan 2013. doi: 10.1371/journal.pcbi.1003357.

M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, and E. Charpentier. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (New York, N.Y.)*, 337(6096):816–21, aug 2012. doi: 10.1126/science.1225829.

I. G. Johnston, B. Gaal, R. P. das Neves, T. Enver, F. J. Iborra, and N. S. Jones. Mitochondrial variability as a source of extrinsic cellular noise. *PLoS computational biology*, 8(3):e1002416, 2012. doi: 10.1371/journal.pcbi.1002416.

E. Jones, T. Oliphant, P. Peterson, and E. Al. SciPy: Open source scientific tools for Python.

R. E. Kass and A. E. Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, jun 1995. doi: 10.1080/01621459.1995.10476572.

T. B. Kepler and T. C. Elston. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical journal*, 81(6):3116–36, dec 2001. doi: 10.1016/S0006-3495(01)75949-8.

B. N. Kholodenko. Cell-signalling dynamics in time and space. *Nature reviews. Molecular cell biology*, 7(3):165–76, mar 2006. doi: 10.1038/nrm1838.

I. Kirmes, A. Szczurek, K. Prakash, I. Charapitsa, C. Heiser, M. Musheev, F. Schock, K. Fornalczyk, D. Ma, U. Birk, C. Cremer, and G. Reid. A transient ischemic environment induces reversible compaction of chromatin. *Genome Biol*, 16(1):246, 2015. doi: 10.1186/s13059-015-0802-2.

M. S. Ko. A stochastic model for gene induction. *Journal of theoretical biology*, 153(2): 181–94, nov 1991.

M. S. Ko, H. Nakauchi, and N. Takahashi. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *The EMBO journal*, 9(9):2835–42, sep 1990.

D. Kraft. A software package for sequential quadratic programming. In *Tech. Rep. DFVLR-FB 88-28*. DLR German Aerospace Center âĂŤ Institute for Flight Mechanics, Koeln, 1988.

C. Kreutz, A. Raue, D. Kaschek, and J. Timmer. Profile likelihood in systems biology. *The FEBS journal*, 280:2564–2571, 2013. doi: 10.1111/febs.12276.

V. Kumar and P. Chambon. The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. *Cell*, 55(1):145–56, oct 1988. doi: 10.1016/0092-8674(88)90017-7.

M. Lacroix and G. Leclercq. Relevance of breast cancer cell lines as models for breast tumours: An update. *Breast Cancer Research and Treatment*, 83(3):249–289, 2004. doi: 10.1023/B:BREA.0000014042.54925.cc.

D. R. Larson, D. Zenklusen, B. Wu, J. a. Chao, and R. H. Singer. Real-time observation of transcription initiation and elongation on an endogenous yeast gene. *Science (New York, N.Y.)*, 332(6028):475–8, apr 2011. doi: 10.1126/science.1202142.

D. R. Larson, C. Fritzsch, L. Sun, X. Meng, D. S. Lawrence, and R. H. Singer. Direct observation of frequency modulated transcription in single cells using light activation. *eLife*, 2013:e00750, jan 2013. doi: 10.7554/eLife.00750.

L. A. Laviolette, K. M. Hodgkinson, N. Minhas, C. Perez-Iratxeta, and B. C. Vanderhyden. 17β-estradiol upregulates GREB1 and accelerates ovarian tumor progression in vivo. *International journal of cancer*, 135(5):1072–84, sep 2014. doi: 10.1002/ijc.28741.

G. Lebedeva, A. Yamaguchi, S. P. Langdon, K. Macleod, and D. J. Harrison. A model of estrogen-related gene expression reveals non-linear effects in transcriptional response to tamoxifen. *BMC systems biology*, 6(1):138, jan 2012. doi: 10.1186/1752-0509-6-138.

L. Lemaire, F. Jay, I.-H. Lee, K. Csilléry, and M. G. B. Blum. Goodness-of-fit statistics for approximate Bayesian computation. pages 1–30, jan 2016.

V. Lemaire, C. F. Lee, J. Lei, R. Métivier, and L. Glass. Sequential recruitment and combinatorial assembling of multiprotein complexes in transcriptional activation. *Phys Rev Lett*, 96(19):198102, 2006.

B. Li, M. Carey, and J. L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–19, 2007. doi: 10.1016/j.cell.2007.01.015.

C. Loos, C. Marr, F. J. Theis, and J. Hasenauer. Approximate Bayesian Computation for Stochastic Single-Cell Time-Lapse Data Using Multivariate Test Statistics. In *Computational Methods in Systems Biology, 13th International Conference, CMSB 2015 Nantes, France, September 16âĂŞ18, 2015 Proceedings*, pages 52–64. Springer International Publishing, 2015. ISBN 978-3-319-23400-7. doi: 10.1007/978-3-319-23401-46.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012. doi: 10.1007/s11222-011-9288-2.

J.-m. Marin, N. S. Pillai, C. P. Robert, and J. Rousseau. Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859, nov 2014. doi: 10.1111/rssb.12056.

M. Markićević, R. Džodić, M. Buta, K. Kanjer, V. Mandušić, Z. Nešković-Konstantinović, and D. Nikolić-Vukosavljević. Trefoil factor 1 in early breast carcinoma: a potential indicator of clinical outcome during the first 3 years of follow-up. *International journal of medical sciences*, 11(7):663–73, 2014. doi: 10.7150/ijms.8194.

H. H. McAdams and A. Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 94(3):814–9, feb 1997.

M. D. McKay, R. J. Beckman, and W. J. Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 21(2):239, may 1979. doi: 10.2307/1268522.

P. P. Medina, J. Carretero, E. Ballestar, B. Angulo, F. Lopez-Rios, M. Esteller, and M. Sanchez-Cespedes. Transcriptional targets of the chromatin-remodelling factor SMARCA4/BRG1 in lung cancer cells. *Human molecular genetics*, 14(7):973–82, apr 2005. doi: 10.1093/hmg/ddi091.

R. Métivier, G. Penot, M. R. Hübner, G. Reid, H. Brand, M. Kos, and F. Gannon. Estrogen receptor-alpha directs ordered, cyclical, and combinatorial recruitment of cofactors on a natural target promoter. *Cell*, 115(6):751–63, dec 2003.

R. Métivier, G. Penot, R. P. Carmouche, M. R. Hübner, G. Reid, S. Denger, D. Manu, H. Brand, M. Kos, V. Benes, and F. Gannon. Transcriptional complexes engaged by apo-estrogen receptor-alpha isoforms have divergent outcomes. *The EMBO journal*, 23 (18):3653–66, sep 2004. doi: 10.1038/sj.emboj.7600377.

R. Métivier, G. Reid, and F. Gannon. Transcription in four dimensions: nuclear receptor-directed initiation of gene expression. *EMBO Reports*, 7(December 2005):1–7, feb 2006. doi: 10.1038/sj.embor.7400626.

R. Métivier, R. Gallais, C. Tiffoche, C. Le Péron, R. Z. Jurkowska, R. P. Carmouche, D. Ibberson, P. Barath, F. Demay, G. Reid, V. Benes, A. Jeltsch, F. Gannon, and G. Salbert. Cyclical DNA methylation of a transcriptionally active promoter. *Nature*, 452(7183):45–50, mar 2008. doi: 10.1038/nature06544.

N. Molina, D. M. Suter, R. Cannavo, B. Zoller, I. Gotic, and F. Naef. Stimulus-induced modulation of transcriptional bursting in a single mammalian gene. *Proceedings of the National Academy of Sciences of the United States of America*, 110(51):20563–8, dec 2013. doi: 10.1073/pnas.1312310110.

P. D. Moral, A. Doucet, and A. Jasra. An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation. *Stat. and Comput.*, (December 2008):1–12, 2011. doi: 10.1007/s11222-011-9271-y.

K. Munk. *Taschenlehrbuch Biologie Genetik*. Georg Thieme Verlag, Stuttgart, 2010. ISBN 9783131916419. doi: 10.1055/b-002-29650.

B. Munsky, G. Neuert, and A. van Oudenaarden. Using Gene Expression Noise to Understand Gene Regulation. *Science*, 336(6078):183–187, apr 2012. doi: 10.1126/science.1216379.

M. C. Neale and M. B. Miller. The use of likelihood-based confidence intervals in genetic models. *Behavior genetics*, 27(2):113–20, mar 1997. doi: 10.1023/A:1025681223921.

T. Nolan, R. E. Hands, and S. Bustin. Quantification of mRNA using real-time RT-PCR. *Nature protocols*, 1(3):1559–82, jan 2006. doi: 10.1038/nprot.2006.236.

A. L. Paek, J. C. Liu, A. Loewer, W. C. Forrester, and G. Lahav. Cell-to-Cell Variation in p53 Dynamics Leads to Fractional Killing. *Cell*, 165(3):631–42, apr 2016. doi: 10.1016/j.cell.2016.03.025.

J. Paulsson. Models of stochastic gene expression. *Physics of Life Reviews*, 2(2):157–175, jun 2005. doi: 10.1016/j.plrev.2005.03.003.

J. Peccoud and B. Ycart. Markovian modelling of gene product synthesis. *Theoretical Population Biology*, 1995.

F. Pedregosa and G. Varoquaux. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.1007/s13398-014-0173-7.2.

F. Perez and B. E. Granger. IPython : A System for Interactive Scientific Computing. *IEEE*, pages 21–29, 2007.

J. M. Rae, M. D. Johnson, J. O. Scheys, K. E. Cordero, J. M. Larios, and M. E. Lippman. GREB 1 is a critical regulator of hormone dependent breast cancer growth. *Breast cancer research and treatment*, 92(2):141–9, jul 2005. doi: 10.1007/s10549-005-1483-4.

A. Raj and A. van Oudenaarden. Single-molecule approaches to stochastic gene expression. *Annual review of biophysics*, 38:255–70, jan 2009. doi: 10.1146/annurev.biophys.37.032807.125928.

A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, and S. Tyagi. Stochastic mRNA synthesis in mammalian cells. *PLoS biology*, 4(10):e309, oct 2006. doi: 10.1371/journal.pbio.0040309.

A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics (Oxford, England)*, 25(25):2–5, aug 2009. doi: 10.1093/bioinformatics/btp358.

J. Rausenberger and M. Kollmann. Quantifying origins of cell-to-cell variations in gene expression. *Biophysical journal*, 95(10):4523–8, nov 2008. doi: 10.1529/biophysj.107.127035.

G. Reid, R. Métivier, C.-Y. Lin, S. Denger, D. Ibberson, T. Ivacevic, H. Brand, V. Benes, E. T. Liu, and F. Gannon. Multiple mechanisms induce transcriptional silencing of a subset of genes, including oestrogen receptor alpha, in response to deacetylase inhibition by valproic acid and trichostatin A. *Oncogene*, 24(31):4894–907, jul 2005. doi: 10.1038/sj.onc.1208662.

G. Reid, R. Gallais, and R. Métivier. Marking time: the dynamic role of chromatin and covalent modification in transcription. *The international journal of biochemistry & cell biology*, 41(1):155–63, jan 2009. doi: 10.1016/j.biocel.2008.08.028.

R. Rinott, A. Jaimovich, N. Friedman, R. R, J. A, and F. N. Exploring transcription regulation through cell-to-cell variability. *Proceedings of the National Academy of Sciences*, 108(15):6329–34, apr 2011. doi: 10.1073/pnas.1013148108.

K. N. Rybakova, F. J. Bruggeman, A. Tomaszewska, M. J. Moné, C. Carlberg, and H. V. Westerhoff. Multiplex Eukaryotic Transcription (In)activation: Timing, Bursting and Cycling of a Ratchet Clock Mechanism. *PLOS Computational Biology*, 11(4):e1004236, 2015. doi: 10.1371/journal.pcbi.1004236.

Y. Shang, X. Hu, J. DiRenzo, M. a. Lazar, and M. Brown. Cofactor dynamics and sufficiency in estrogen receptor-regulated transcription. *Cell*, 103(6):843–52, dec 2000.

L. V. Sharova, A. a. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S. H. Ko. Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 16(1):45–58, feb 2009. doi: 10.1093/dnares/dsn030.

D. Silk, S. Filippi, and M. P. H. Stumpf. Optimizing threshold-schedules for sequential approximate Bayesian computation: Applications to molecular systems. *Statistical Applications in Genetics and Molecular Biology*, 12(5):603–618, 2013. doi: 10.1515/sagmb-2012-0043.

S. A. Sisson, Y. Fan, and M. M. Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104 (6):1760–1765, 2007. doi: 10.1073/pnas.0607208104.

H. D. Soule, J. Vazguez, A. Long, S. Albert, and M. Brennan. A human cell line from a pleural effusion derived from a breast carcinoma. *Journal of the National Cancer Institute*, 51(5):1409–16, nov 1973. doi: 10.1093/jnci/51.5.1409.

J. Spona, H. Leibl, and C. Bieglmayer. Nuclear translocation of estrogen-receptor complex and stimulation of RNA synthesis by estrogens of different biological potencies in the female rat pituitary. *Biochimica et biophysica acta*, 607(2):189–200, apr 1980. doi: 10.1016/0005-2787(80)90071-4.

T. J. Stasevich, Y. Hayashi-takanaka, Y. Sato, K. Maehara, Y. Ohkawa, K. Sakata-sogawa, M. Tokunaga, T. Nagase, N. Nozaki, J. G. Mcnally, and H. Kimura. Regulation of RNA polymerase II activation by histone acetylation in single living cells. *Nature*, 516(7530):272–275, 2014. doi: 10.1038/nature13714.

J. Sun, Z. Nawaz, and J. M. Slingerland. Long-range activation of GREB1 by estrogen receptor via three distal consensus estrogen-responsive elements in breast cancer cells. *Molecular endocrinology (Baltimore, Md.)*, 21(11):2651–62, nov 2007. doi: 10.1210/me.2007-0082.

X.-X. Sun and Q. Yu. Intra-tumor heterogeneity of cancer cells and its implications for cancer treatment. *Acta pharmacologica Sinica*, 36(10):1219–27, oct 2015. doi: 10.1038/aps.2015.92.

M. Sunnåker, A. G. Busetto, E. Numminen, J. Corander, M. Foll, and C. Dessimoz. Approximate Bayesian Computation. *PLoS Computational Biology*, 9(1), 2013. doi: 10.1371/journal.pcbi.1002803.

D. M. Suter, N. Molina, D. Gatfield, K. Schneider, U. Schibler, and F. Naef. Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science (New York, N.Y.)*, 332(472):472–474, 2011. doi: 10.1126/science.1198817.

F. Tang, C. Barbacioru, E. Nordman, B. Li, N. Xu, V. I. Bashkirov, K. Lao, and M. A. Surani. RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature protocols*, 5(3):516–35, mar 2010. doi: 10.1038/nprot.2009.236.

S. Tavaré, D. J. Balding, R. C. Griffiths, and P. Donnelly. Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–18, feb 1997.

M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98 (15):8614–9, jul 2001. doi: 10.1073/pnas.151588598.

T. Toni and M. P. H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 2009. doi: 10.1093/bioinformatics/btp619.

T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society, Interface / the Royal Society*, 6(31):187–202, 2009. doi: 10.1098/rsif.2008.0172.

V. B. Van Asch. *Domain Similarity Measures: On the use of distance metrics in natural language processing*. PhD thesis, Universiteit Antwerpen, 2012.

S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2):22–30, mar 2011. doi: 10.1109/MCSE.2011.37.

N. G. Van Kampen. *Stochastic processes in physics and chemistry*, volume 11. Elsevier, 1992. ISBN 0444893490. doi: 10.2307/2984076.

G. Van Rossum and F. L. Drake. *The python reference manual*. PythonLabs, 2001. URL `www.python.org`.

C. wa Maina, A. Honkela, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence, and M. Rattray. Inference of RNA polymerase II transcription dynamics from chromatin immunoprecipitation time course data. *PLoS computational biology*, 10(5):e1003598, may 2014. doi: 10.1371/journal.pcbi.1003598.

J. Watson, T. Baker, S. P. Bell, A. Gann, M. Levine, and R. Losick. *Molecular Biology of the Gene*. Prentice Hall, 6 edition, 2007. ISBN 0321507819.

D. Zenklusen, D. R. Larson, and R. H. Singer. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nature structural & molecular biology*, 15(12): 1263–71, dec 2008. doi: 10.1038/nsmb.1514.

B. Zoller, D. Nicolas, N. Molina, and F. Naef. Structure of silent transcription intervals and noise characteristics of mammalian genes. *Molecular Systems Biology*, 11(7): 823–823, jul 2015. doi: 10.15252/msb.20156257.