AUTHOR: **Michał Komorowski** DEGREE: **Ph.D.**

TITLE: **Statistical methods for estimation of biochemical kinetic parameters**

DATE OF DEPOSIT: ....

Statistical methods for estimation of biochemical kinetic parameters

by

Michał Komorowski

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

July 2009

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

Every YES requires an accompanying NO. Working on this thesis I said NO countless of times. First of all I would like to apologise to people who heard my NO most often. I naively hope that the microscopic contribution of this thesis to human knowledge will compensate for the distress I caused working on it.

The efforts that led to the results presented here would not have been possible without a number of people. Primarily, the gratitude towards my parents for their sacrifice and constant support throughout 29 years of my life can not be described by words.

I would like to thank my supervisors Dr Bärbel Finkenstädt and Prof. David Rand for their enormous support and an opportunity to work on exciting problems.
Without Dr Omiros Papasiliopoullos my knowledge of MCMC would have been very narrow. I thank him for introducing me to these techniques, for his time and interest in my work.
Dr Dan Woodcock was always willing to help me with my linguistic problems.

I thank the Department of Statistics for funding me with a PhD scholarship.

Finally there is a number of people who have made big efforts to keep my life in balance. I am grateful to them most of all. They have saved me and my NOs were most painful to them. I will mention only a few here. I cannot imagine last three years without two of my best friends: Piotr Janas and Piotr Zwiernik. They have found their own, unique ways to keep me going. The value of discussions with one of my closest friends Daniel Valdes Amaro has been enormous. Agnieszka and Michał Rutkowscy have never forgotten about me, they remembered during periods when I stayed silent. I will always be looking forward to our cycling trips and to the atmosphere you both miraculously create. Long jogs and discussions with Daniel Claus kept my body and mind fit. Life at the University without my Chaplaincy friends would have been hardly possible for me. I am deeply indebted to Rev. Prem Fernando.

A.M.D.G.

Declarations

The material in this thesis is original except for Chapter 1 which is background review chapter. Chapters 2 have been publish in Bioinformatics . Chapter 3 and 4 has been submitted for publication in BMC Bioinformatics and Biophysical Journal, respectively.

Chapter 2 is the result of joint efforts between Bärbel Finkenstädt, Elizabeth A. Heron, myself, Kieron Edwards, Sanyi Tang, Claire V. Harper, Julian R. E. Davis, Michael R. H. White and Andrew J. Millar and David Rand.

Experiments resulting with experimental data used in Chapters 3,5 were performed by Claire V. Harper.

Chapters 3,4,5 is the result of joint work with my supervisors Bärbel Finkenstädt and David Rand.

Chapters 2,3,4 and 5 contain section *Author contributions and chapter's structure* that specifies contribution of each author in more details.

Abstract

This thesis consists of four original pieces of work contained in chapters 2,3,4 and 5. These cover four topics within the area of statistical methods for parameter estimation of biochemical kinetic models. Emphasis is put on integrating single-cell reporter gene data with stochastic dynamic models.

Chapter 2 introduces a modelling framework based on stochastic and ordinary differential equations that addresses the problem of reconstructing transcription time course profiles and associated degradation rates from fluorescent and luminescent reporter genes. We present three case studies where the methodology is used to reconstruct unobserved transcription profiles and to estimate associated degradation rates.

In Chapter 3 we use the linear noise approximation to model biochemical reactions through a stochastic dynamic model and derive an explicit formula for the likelihood function which allows for computationally efficient parameter estimation. The major advantage of the method is that in contrast to the more established diffusion approximation based methods the computationally costly techniques of data augmentation are not necessary.

In Chapter 4 we present an inference framework for interpretation of fluorescent reporter gene data. The method takes into account stochastic variability in a fluorescent signal resulting from intrinsic noise of gene expression, extrinsic noise and kinetics of fluorescent protein maturation.

Chapter 5 presents a Bayesian hierarchical model, that allows us to infer distributions of fluorescent reporter degradation rates.

All methods are embedded in a Bayesian framework and inference is performed using Markov chain Monte Carlo.

Abbreviations

ACF	autocorrelation function
CFP	cyan fluorescent protein
CHX	cycloheximide (inhibitor of protein biosynthesis)
DA	diffusion approximation
GFP	green fluorescent protein
LNA	linear noise approximation
MH	Metropolis-Hastings (algorithm)
MCMC	Markov chain Monte Carlo
MRE	macroscopic rate equation
ODE	ordinary differential equation
OU	Ornstein-Uhlenbeck (process)
SDE	stochastic differential equation
YFP	yellow fluorescent protein

Chapter 1

Introduction

1.1 Motivation

Systems biology is an inter-disciplinary research area that focuses on the systematic study of complex interactions in biological systems [Ehrenberg et al., 2003, Gunawardena, 2009]. It uses mathematical models to investigate how physiology and phenotype arises from molecular interactions [Gunawardena, 2009]. Two main research branches can be distinguished. The first, "omics"‘ is stimulated by high-throughput technologies such as the microarray and focuses on inferring networks from large data sets [Schena et al., 1995]. The second branch models interactions between molecules, cells and tissues and is often called "mechanistic" systems biology [Guyton et al., 1972, Hodgkin and Huxley, 1990, Kacser et al., 1995, Rapoport et al., 1974, Savageau, 1976].

Within both "omics" and "mechanistic" systems biology molecular interactions, often represented as molecular networks, play a central role [Huang, 2004]. A static, genome-wide picture of networks is of interest in "omics", whereas "mechanistic" systems biology focuses on dynamical properties of smaller networks. Among different types of molecular networks, genetic regulatory networks occupy a special place as they act as information processing machines in cells, transforming the cellular and extracellular inputs into appropriate outputs in the form of gene expression [Tkačik et al., 2008]. Rules by which these computations are performed are crucial for understanding the functioning of living cells, in particular their response to the environment and decisions about their fate [Ziv et al., 2007].

The expression of a single gene, the basic unit of a gene regulatory network, involves discrete and inherently random biochemical reactions [Guptasarma, 1995]. As a consequence regulatory networks comprise spontaneously fluctuating biochemical species. It is surprising that cells function remarkably well in the presence of noise, often performing close to the physical limits imposed by the discrete nature of the signal processing machinery [Bialek and Setayeshgar, 2005, Tkačik et al., 2008].

Dynamical properties and functionality of a network are constituted by its architecture and kinetic parameters describing velocities at which biochemical reactions occur. Therefore, there is an increased interest in determining these using a variety of available experimental techniques. Often networks are studied experimentally applying time-lapse reporters that allow for real time, in vivo measurements of concentrations of interacting biochemical species. Data obtained in this type of experiments are being used to determine the strength of interactions within networks.

Research presented in this thesis concerns statistical inference of biochemical kinetic parameters and belongs to the realm of "mechanistic" systems biology. The methodology developed here allows for estimation of kinetic parameters of gene expression from reporter assays data such as fluorescent reporter genes and PCR based assays. Methods are oriented towards single-cell data and stochasticity in gene expression so that they may contribute to the broad range of attempts to understand the ability of a living cells to grow, divide, sense and respond to its environment in the presence of random effects.

In the first part of the introductory chapter we focus on providing motivation for the methods presented in the thesis. A stimulus for this type of research comes from the area of stochastic gene expression and gene regulation that are particularly interesting due to consequences of randomness involved in cellular processes. Therefore we briefly review research in this fields.

In the second part of the introduction we shortly describe mathematical methods for modelling biochemical reactions and statistical tools for inference of biochemical kinetics parameters.

1.2 Stochasticity in biological systems

Gene expression is an inherently stochastic process [Raser and O'Shea, 2005] that leads to cell-to-cell variation in mRNA and protein levels. This randomness can be beneficial in some cases and harmful in others [Raj and van Oudenaarden, 2008]. In this section we describe the main studies that lead to a characterisation of stochasticity and explored its implications.

The experimental proof that levels of gene expression vary from cell to cell was provided by [Novick and Weiner, 1957]. The authors revealed the unpredictability of a cell's response by demonstrating that induction by lactose increased the proportion of cells expressing the beta-galactosidase enzyme rather than the expression level in each cell equally.

Another pioneering study [Ko et al., 1990] used advances in fluorescent reporter technology to examine the foundation of the variability in expression. They studied the effect of different doses of glucocorticoid on the expression of a glucocorticoid-responsive gene and observed surprisingly large cell-to-cell variability.

Theoretical investigations of the stochastic nature of gene expression was initiated by [Arkin et al., 1998, McAdams and Arkin, 1997]. Researchers modelled gene expression using a stochastic formulation of chemical kinetics proposed by [Gillespie, 1977] and predicted that in some biologically realistic parameter ranges protein numbers may fluctuate noticeably.

1.2.1 Sources of variability in gene expression

The need for precise characterisation of stochasticity in gene expression was inspired by experiments in synthetic biology. Researchers constructed a synthetic network called "repressilator", composed of three repressors that was capable of producing oscillations in gene expression [Elowitz and Leibler, 2000]. The oscillations, however, were subject to fluctuations in their period and magnitude. The study led to the hypothesis that randomness in gene expression perturbed the functioning of an engineered genetic circuit.

Further experiments explored the origins of stochastic gene expression. The study of [Elowitz et al., 2002b] followed by the mathematical analysis by [Swain et al., 2002b] introduced the concept of extrinsic and intrinsic noise. In their experiment, they transfected two copies of the same promoter into the E. coli genome. One gene coded the cyan fluorescent protein (CFP) and the other coded the yellow fluorescent protein (YFP). Extrinsic fluctuations were defined as sources of variability that affect the expression of both copies of the gene equally in a given cell. Intrinsic fluctuations were those resulting from the randomness inherent in transcription and translation and influenced each copy of the gene independently leading to uncorrelated variations in levels of CFP and YFP.

The study by [Ozbudak et al., 2002] demonstrated that the variability in expression is depended on the underlying biochemical rates of transcription and translation. The study verified a prediction by [Thattai and van Oudenaarden, 2001], about how intrinsic noise in gene expression would change as these parameters were varied.

1.2.2 Noise and network architecture

Studies of noise in expression of a single gene were followed by investigations of how stochasticity transmits within gene regulatory networks. First, the common structure of linear transcriptional cascade [Pedraza and van Oudenaarden, 2005, Rosenfeld et al., 2005] was examined. Researchers discovered that noise can be transmitted from the upstream gene to the downstream gene substantially increasing the variability of the expression of the downstream gene. Another study showed that longer genetic cascades can filter out rapid fluctuations [Hooshangi et al., 2005].

Beside linear cascades, negative and positive feedbacks are other common motives in genetic regulatory networks. Feedback arises as the protein encoded by a gene negatively or positively influences its own transcription. The theoretical model by [Thattai and van Oudenaarden, 2001] predicted that the presence of these motives changes the magnitude of fluctuations. This effect was confirmed experimentally by [Austin et al., 2006, Becskei et al., 2001, Dublanche et al., 2006].

1.2.3 Beneficial and harmful effects of noise

Investigations of noise in genetic networks revealed its two roles in cellular processes. First, noise as a nuisance that disturbs a reliable functioning of a cellular machinery. Second, noise as a source of variability that is exploited by cells. Nonessential genes are predicted to be noisy, indicating the potential benefits of noise in this class of genes [Blake et al., 2006]. On the other hand, genes controlling the protein synthesis and degradation are predicted to be much less variable. This suggests that genes essential for cellular function require more precise control [Blake et al., 2006].

1.2.4 Stochasticity in cell fate and decision making

The other interesting aspect of noise, where its advantageous and disadvantageous consequences are revealed, appears in the context of mechanisms by which cells respond to the environment and decide about their fates. Fate decisions are of special interest to developmental biology, which studies decisions about differentiation into subtypes with specialised attributes. How cells adopt a particular fate is usually seen as a deterministic process, that results either from virtue of the cell lineage or from proximity to an inductive signal from another cell. Certain developmental decisions, however, are random, sometimes out of necessity or sometimes to explore benefits of randomness [Losick and Desplan, 2008]. For instance, Drosophila melanogaster generates neurons and glial cells by asymmetric processes of cell division. On the other hand, differentiation into alternative colour vision photoreceptors in Drosophila is generated by intrinsically stochastic biochemical reaction [Losick and Desplan, 2008].

The classical example of stochastic decision making is the lysis-lysogeny decision of bacteriophage lambda [Ptashne, 2007]. After infection of Escherichia coli, phage follows one of the two possible developmental pathways, lytic growth or lysogeny. Isogenic cells grown in the same environment, each infected with a phage particle, can produce both possible outcomes. Some cells follow lyse pathway while others become lysogens. A theoretical model of lambda infection supports the hypothesis that biochemical kinetic noise during infection is responsible for observed heterogeneity in cell fates [Arkin et al., 1998].

1.3 Systems level approach

Studies of stochasticity in living organisms stimulated two initiatives in systems biology research [Wilkinson, 2009]. First, stochastic models [Ashall et al., 2009, Chabot et al., 2007, Lipniacki et al., 2006] are being used in preference to deterministic models [Hoffmann et al., 2002, Lipniacki et al., 2004, Nelson et al., 2004] to describe biochemical network dynamics at the single-cell level. Second, advanced statistical methodology is being used to estimate parameters of both deterministic and stochastic models from time-lapsed experimental data [Chabot et al., 2007, Finkenstadt et al., 2008, Henderson et al., 2009, Heron et al., 2007]. Even a simple biological system can exhibit a range of complex dynamical behaviour [Elowitz and Leibler, 2000], therefore quantitative mathematical and statistical modelling is necessary to provide its accurate description. Traditionally, biochemical systems dynamics have been described using continuous deterministic mathematical models [Van Kampen, 2006]. As the intrinsic stochasticity of biochemical kinetics has been acknowledged it is now commonly accepted that stochastic models are necessary to properly capture heterogeneous behaviour of intracellular systems in a realistic way [Wilkinson, 2009]. Stochastic models, however, are computationally more costly and significantly more challenging to fit to experimental data.

Deterministic models

The classical approach to modelling chemical kinetics is to assume that molecular species are present in large numbers and their concentrations are measured as continuous variables [Van Kampen, 2006]. The changes in concentrations are governed by reactions that are assumed to be continuous and deterministic processes. The velocity of each reaction is given by a rate constant or a rate equation (e.g. Michaelis-Menten or Hill kinetics) [Cornish-Bowden, 1995]. Ordinary differential equations (ODEs) can be used to describe evolution of such a system. ODEs in this context are usually called macroscopic rate equation (MRE). The basic limitation of the deterministic models is that they fail to explain the behaviour of single-cell data, that are essential for understanding gene regulatory networks. The small numbers of molecules present in single cells results in random effects that contain additional information about the systems dynamics. In order to extract this information from data requires a stochastic model. The strength of stochastic effects is mainly dependent on the number of molecules in a reacting system, therefore for systems that involve small molecular numbers, stochastic models, would provide a more insightful description. In Chapter 2 we demonstrate that by using a stochastic model, one can estimate parameters that are unidentifiable in a deterministic model.

Stochastic models

A useful model of a single-cell biochemical system must therefore account for an inherent randomness of a modelled phenomena. Methods of statistical mechanics allowed for derivation of the probabilistic framework to model the dynamics of biochemical reactions. This was achieved using Poisson birth and death processes [Gardiner, 1985, Gillespie, 1992b, Van Kampen, 2006, Wilkinson, 2006] and the, so called, Chemical Master Equation that describes the probabilistic evolution of the system. The state of the system is assumed to be a vector of counts of molecules that changes due to reaction occurrence appear at discrete time points.

This model has been widely studied in the theory of stochastic processes and a number of algorithms can be used to simulate data from a stochastic chemical kinetics model [Cao et al., 2005, Gillespie, 1977, 2001, Gillespie and Petzold, 2003, Kiehl et al., 2004, Puchałka and Kierzek, 2004].

Although stochastic models based on Poisson birth and death processes provide insight that is hidden when using deterministic models, they create certain difficulties. The computational cost of a simulation is high and analytical solutions exist only for a very limited number of systems [McQuarrie, 1967]. Therefore, several types of approximations to birth and death processes have been proposed [Gardiner, 1985, Van Kampen, 2006]. Such approximations use processes that have probability distributions similar, to that of birth and death stochastic kinetic models. There are two main types of approximations, the diffusion approximation (DA) and the linear noise approximation (LNA).

The diffusion approximation provides stochastic differential equation (SDE) models where the stochastic perturbation is introduced by a state dependant Gaussian noise. The linear noise approximation can be seen as a combination of deterministic and stochastic approach because it incorporates the deterministic MRE as a model of the deterministic system and the SDEs to approximatively describe the fluctuations around the deterministic state.

Using SDE models is more convenient than Poisson birth and death models because it is usually straightforward to simulate trajectories from the SDEs using well established numerical procedures [Kloeden and E., 1999].

1.4 Statistical methods for stochastic biochemical kinetics

At the opposite end of the "mechanistic" systems biology spectrum, there is a considerable interest in using statistical methods to estimate parameters of the dynamical models of the biological systems [Gunawardena, 2009]. In general the aim is to calibrate the model so as to reproduce experimental results in the best possible way.

Methods for deterministic models

The fitting of deterministic models to time course data has a long tradition and comprises a variety of different approaches [Esposito and Floudas, 2000, Mendes and Kell, 1998, Moles et al., 2003, Ramsay et al., 2007]. The simplest method involves "distance" between the model predictions and the experimental data, and then finding parameters that minimise the distance measure. The least squares fitting approach is an example of this type [Jaqaman and Danuser, 2006]. The theoretical way to measure the discrepancy between the model and the experimental data is given by the statistical concept of the likelihood function [Silvey, 1975]. If the likelihood function is constructed an optimisation procedure can be used to obtain maximum likelihood estimates. Maximum likelihood estimates have a good property of consistency, nevertheless there are many complications related to optimisation procedure. Likelihood function may be approximately or completely flat in the neighbourhood of the optimum, demonstrating a lack of identifiability of the model parameters. Furthermore, the likelihood can be multimodal. These issues can be approached using Bayesian statistics often combined with Markov chain Monte Carlo (MCMC) techniques [Barenco et al., 2006, Brown and Sethna, 2003, Vyshemirsky and Girolami, 2008]. Bayesian inference [Gamerman and Lopes, 2006] combines prior information about the model parameters with the likelihood function that contains information present in the data. Information about model parameters contained in both the experimental data and the prior distribution is called posterior distribution. Usually a posterior distribution does not have an analytical form, nevertheless it is straightforward to construct MCMC algorithms [Gamerman and Lopes, 2006] that allow to simulate samples from it.

Methods for stochastic models

Statistical methodology for stochastic biochemical kinetic models have only recently appeared in the literature. It is probably due to the fairly sophisticated algorithms required to fit stochastic models, as the analytical form of their likelihood function is usually not available. Pursuits to tackle this problem have turned out to be successful. Here, we briefly discuss a variety of solutions that has been developed.

Derivation of a likelihood function for a stochastic model requires a probabilistic description of the biochemical system. The most exact framework to describe a biochemical kinetics system is that of Poisson birth and death processes and the corresponding CME. Few inference methods have exploited the CME to propose inference algorithms. One method, proposed by [Reinker et al., 2006], approximated the likelihood function, the other, suggested by [Tian et al., 2007] estimated it using Monte Carlo methods. Recently, also a method based on the exact likelihood [Boys et al., 2008] has been developed.

Although, methods based on the CME have the desired property of using the most accurate model they are computationally intensive and difficult to apply to problems of realistic size and complexity. In general, they are likely to be problematic also because of the requirement of the single molecule precision data, that are not available in the majority of gene expression experiments.

As stated above, the strategy based on replacing a Poison birth and death process with its approximation was a successful way to reduce computational cost of simulating trajectories of a biochemical system. The same technique has also been used to construct inference algorithms. It reduces the problem of estimating the parameters of a Poisson process to the one of estimating the parameters of a diffusion process. As the inference methods for diffusion equations are subject to intensive development [Beskos et al., 2006, Durham G. B, 2002, Elerian et al., 2001] they can be adapted to infer biochemical kinetic rates [Finkenstadt et al., 2008, Golightly and Wilkinson, 2005, Heron et al., 2007]. Although these methods are simpler in comparison to approaches based on the CME they also require sophisticated and computationally intensive Bayesian inference techniques. Simplification, however, allowed to successfully apply these methods to real biological data [Finkenstadt et al., 2008, Henderson et al., 2009, Heron et al., 2007]. Applicability of the DA based methods is unfortunately limited to simple systems, in which most of the model variables are observed. In an usual experimental setting, however, only few components of a biochemical systems can be measured in vivo. It may make MCMC simulation problematic, because unobserved variables need to be integrated out within a simulation.

The main result of this thesis is an alternative inference method that approximates Poisson birth and death process using the LNA instead of the DA. The derived framework allows for unobserved variables and measurement error without the price of an additional computational cost. It works surprisingly well even in situations in which one would not expect the LNA to be a good approximation to a birth and death process.

Chapter 2

Reconstruction of transcriptional dynamics from gene reporter data using differential equations

2.1 Author contributions and chapter's structure

This chapter is a paper by Bärbel Finkenstädt, Elizabeth A. Heron, Michal Komorowski, Kieron Edwards , Sanyi Tang, Claire V. Harper, Julian R. E. Davis, Michael R. H. White, Andrew J. Millar and David A. Rand published in Bioinformatics 2008 24(24):2901-2907. Authors contribution are as follows. BF conducted the numerical estimations for case studies 1 and 2 (ODE part). EAH and MK did numerical estimations for case study 2 (SDE) and 3, respectively, under the supervision and guidance of BF and DAR. ST contributed to the algorithm development at the early stages of the paper. KE performed experiments for case study 1 and 2, under guidance of AJM. CVH performed the experiment for case study 3 under guidance of JRED and MRHW. BF wrote the paper with assistance from MK, EAH and DAR. DAR provided help on the mathematical modeling and initiated the collaboration between the theoretical and experimental groups.

Sections 2.2 - 2.6 are followed by supplementary section 2.8 that contains details of mathematical modeling and statistical methods.

2.2 Abstract

Promoter driven reporter genes, notably luciferase (luc) and green fluorescent protein (gfp), provide a tool for the generation of a vast array of time-course data sets from living cells and organisms. The aim of this study is to introduce a modeling framework based on stochastic and ordinary differential equations that addresses the problem of reconstructing transcription time course profiles and associated degradation rates. The dynamical model is embedded into a Bayesian framework and inference is performed using Markov chain Monte Carlo algorithms. We present three case studies where the methodology is used to reconstruct unobserved transcription profiles and to estimate associated degradation rates. We discuss advantages and limits of fitting either stochastic or ordinary differential equations and address the problem of parameter identifiability when model variables are unobserved. We also suggest functional forms such as on/off switches and stimulus response functions to model transcriptional dynamics and present results of fitting these to experimental data.

2.3 Introduction

Imaging data from luciferase (LUC) and green fluorescent protein (GFP) reporters combined with fluorescent tagging of protein can provide very high quality data with good temporal resolution [Millar et al., 1995, Nelson et al., 2004]. In this case the actual imaging time series is approximately proportional to the abundance of an artificial protein. The underlying transcriptional dynamics are unobserved and are masked by two degradation processes, namely of reporter mRNA and reporter protein. In this study we address the problem of back-calculating from the observed protein activity to the hidden transcriptional dynamics where it is of interest to estimate the associated rates of degradation as part of the analysis. We formulate a probability model based on (stochastic) differential equations which provides the mechanistic rules for the back-calculation. In practise heterogeneous data sets may be available from different experiments which contain information about the transcription process and model parameters. Data sources may be of different quality and time resolution, as well as from single cells or an aggregated population of cells. Longitudinal measurements are discrete in time and can be irregularly spaced or on different time scales for different variables. Other realistic shortcomings of the data are that time course measurements may not correspond to the same biological sample, or data on different variables may not be matched in time which would be preferable for fitting a multivariate dynamical model. As the quality and quantity of such data sets supports more or less complex modeling approaches we consider both stochastic and ordinary differential equations with measurement noise. Information on rate constants may be incorporated through prior distributions in a Bayesian approach. We first describe the models and the statistical methods used for its inference. Then we present three case studies each with the aim of reconstructing transcription and inferring any identifiable degradation rates from reporter gene data using available heterogeneous sources of data. These case studies serve to demonstrate the adaption of the methodology to different experimental scenarios.

2.4 Models and Inference

It is now well understood that, because of the stochastic nature of reaction events and the presence of internal noise due to the fluctuations in the molecular environment of the cell, regulatory and signalling systems are intrinsically stochastic. To develop a stochastic model one can attempt to model the individual stochastic events involved such as binding of the transcription factors, the assembly and initiation of the polymerase and transcription. Although an exact simulation algorithm of the corresponding stochastic processes is provided by [Gillespie, 1977, 1992a] such models are too detailed for there to be any hope of fitting to current data with its limitations. Stochastic differential equations (SDEs) provide a good approximation of molecular population systems when one can assume that there is a macroscopic time scale for which (a) the event rates can be regarded as constant and (b) there are many events of each type. An example of formulating and fitting an autoregulatory feedback system with transcriptional delay as a system of SDEs can be found in [Heron et al., 2007]). However, if the data are too sparsely sampled in time to reveal information about the volatility process, or if measurements are not realizations of the same continuous stochastic process in a cell, then the assumption of SDEs can be problematic in estimation. Simpler modeling approaches based on ODEs to represent the mean process with an additional stochastic error may provide a useful vehicle for estimation purposes at least in systems that have relatively regular and stable dynamics. The formulation of ODEs to model the dynamics of molecular population processes has become a widespread tool in systems biology (see, for example, systems studied in [Goldbeter, 2002, Jensen et al., 2003, Locke et al., 2005a,b]), and early statistically less rigorous attempts in obtaining kinetic parameters from GFP reporter data can be found in [Ronen et al., 2002] and [Kalir et al., 2001].

Here we consider the following dynamic model as the mechanistic backbone for the reconstruction of transcription profiles from reporter protein data

$$dM/dt = \tau(t) - \delta_M M(t), \quad dP/dt = \alpha M(t) - \delta_P P(t), \tag{2.1}$$

where M denotes the abundance of mRNA molecules and P denotes the abundance of the corresponding protein. The first equation describes the dynamics of mRNA molecules where transcription is given by a non-negative function $\tau(t)$. The second equation states that the protein is synthesized at a rate proportional to the abundance of mRNA. The mRNA and the protein are degraded (or leave their molecular compartment otherwise) at time scales with mean $1/\delta_m$ and $1/\delta_p$, respectively. The aim is to infer the transcription function $\tau(t)$ and possibly other rate constants of the system given time series data proportional to one or both variables of the system. Suppose that we measure M, P proportionally to their population size, $s_M M(t)$ for the mRNA and $s_P P(t)$ for the reporter protein. Reparameterizing (2.1) gives a scaled model which is identical to (2.1) with scaled terms for α and τ (see supplementary section 2.8). However, degradation rates are not affected by scaling. Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^T = \{M(t_i), P(t_i)\}_{i=1}^T$ denote experimental time series data observed at discrete time points. In order to obtain a likelihood function that incorporates the mechanistic rules in (2.1) we consider two approaches. One is the SDE approach where (2.1) is formulated as an appropriate system of stochastic differential equations. This approach is rigorously modeling the volatility of the stochastic dynamics of the kinetic processes provided that the assumptions of the SDE approximation itself are valid. It is very challenging to incorporate additional measurement error unless its variance is known or assumed. The second is the mean ODE approach where we assume that a solution path to (2.1) represents the mean of a stochastic process whilst the modeler makes assumptions about the probability distribution of the residual process. This approach is less exact than the SDE approach in modeling the volatility of the underlying stochastic interaction between molecules. On the other hand it naturally deals with measurement error and might also be useful for fitting to data sets which do not comply with the SDE assumption, for example, if data points are averages over replicates, come from different samples and/or represent a population of cells. We now introduce the two approaches and their likelihood derivation in more detail.

SDE approach: Here, M and P are random variables of molecular population sizes and the rates of increase and decrease in model (2.1) are event probabilities of birth and death processes at the individual molecular level. One can derive the following Itô SDEs (see supplementary section 2.8)

$$\begin{aligned} dM &= \zeta_M(t,\theta)dt + \sigma_M(t,\theta)dW_M \\ dP &= \zeta_P(t,\theta)dt + \sigma_P(t,\theta)dW_P, \end{aligned} \tag{2.2}$$

where $\zeta_M(t,\theta) = \tau(t) - \delta_m M(t)$, $\quad \zeta_P(t,\theta) = \alpha M(t) - \delta_P P(t)$, and $\sigma_M(t) = s_M^{1/2}(\tau(t) + \delta_m M(t))^{1/2}, \sigma_P(t) = s_P^{1/2}(\alpha M(t) + \delta_P P(t))^{1/2}$ are drift and volatility functions, respectively and W_M and W_P are independent Wiener processes [1]. Here and throughout the Chapter θ is used to denote a vector of model parameters. If M

[1] The Wiener process, or Brownian motion, is a continuous-time stochastic process that has independent normally distributed increments.

and P are indirect measurements of molecular populations in the sense that they are proportional to molecular abundance with factors s_M, s_P then these factors arise as additional parameters in the volatility functions and their estimation will be extremely useful allowing us to calibrate the model to the population level. Given data $\mathbf{Y}$ the likelihood function for the diffusion process is

$$L_{\mathsf{SDE}}(\theta; \mathbf{Y}) = \prod_{i=1}^{T-1} f(\mathbf{y}_{i+1}|\mathbf{y}_i; \theta) \tag{2.3}$$

where $f(\mathbf{y}_{i+1}|\mathbf{y}_i; \theta)$ denotes the transition density of $\mathbf{y}_{i+1}$ given $\mathbf{y}_i$, that is the joint probability distribution of $M(t_{i+1})$ and $P(t_{i+1})$ given present values, under parameter vector θ. The exact transition density function for solutions of SDEs is rarely available in analytical form and usually approximations have to be considered. If the time-step $\Delta t_i = t_{i+1} - t_i$ is small then a good approximation is given by assuming that, conditional on past values,

(*) Increments $\mathbf{y}(t_{i+1}) - \mathbf{y}(t_i)$ are bivariate normal with mean vector $\zeta(t_i)\Delta t_i$ and variance matrix $\Sigma(t_i)\Delta t_i$ where $\zeta(t_i) = (\zeta_M(t_i), \zeta_P(t_i))$, $\Sigma(t_i) = \text{diag}(\sigma_M^2(t_i), \sigma_P^2(t_i))$ are the drift and volatility functions as defined above.

Thus, for sufficiently small sampling intervals Δt_i the likelihood function can be approximated by a product of the form

$$L_{\mathsf{SDE}}(\theta; \mathbf{Y}) = \prod_{i=1}^{T-1} \Phi(\mathbf{y}(t_{i+1}) - \mathbf{y}(t_i); \zeta(t_i)\Delta t_i, \Sigma(t_i)\Delta t_i) \tag{2.4}$$

where $\Phi(x; \mu, \Sigma)$ denotes the bivariate normal density function with mean vector μ and variance matrix Σ. Justifications for this approximation are given in [Kloeden and E., 1999].

Mean ODE approach: Suppose there is a solution path $\mu(t; \theta) = (M(t), P(t); \theta)$ to the system in (2.1) from unknown initial conditions (M_0, P_0). Then a natural probabilistic model is to assume that $\mathbf{Y}$ has a joint distribution with mean function $\mu(t; \theta)$ and a variance function $\sigma^2(t; \theta)$. The distribution function and variance are specified according to assumptions that the modeler makes about the residual process and measurement error. If the error process is assumed independent then the likelihood in the mean ODE approach is

$$L_{\mathsf{ODE}}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{T} g(\mathbf{y}_i | \mu(t_i), \sigma^2(t_i), \theta), \tag{2.5}$$

where θ now incorporates initial conditions (M_0, P_0) and g is a suitably chosen probability distribution.

Inference: By Bayes' theorem the posterior distribution is

$$\pi(\theta | \mathbf{Y}) \propto L(\theta | \mathbf{Y}) \pi(\theta), \tag{2.6}$$

where L is the likelihood function, derived for either the ODE or SDE approach, and $\pi(\theta)$ are prior densities of model parameters. Sampling from the posterior distribution is usually achieved using Markov chain Monte Carlo (MCMC), where each element of θ is updated by using an appropriately constructed Metropolis-Hastings acceptance/rejection scheme based on either random walk or independence proposals [Gamerman and Lopes, 2006]. The reason for choosing a Bayesian approach combined with a MCMC algorithm is twofold: Firstly, the Bayesian methodology is flexible allowing for portability of inference results between different experimental studies in a well defined way and this is highly relevant to studies in systems biology. Secondly, the probabilistic imputation of missing data and/or unobserved variables can be implemented in a straightforward way as part of an MCMC sampler.

Discrete data and unobserved variables: Molecular time series data are discretely measured and it cannot be guaranteed that the sampling interval is small enough for the approximation (*) to work well. A remedy suggested in econometric applications of SDEs [Durham G. B, 2002, Elerian et al., 2001] is to augment the observed data by introducing a number of latent or unobserved data points, called a *bridge*, in-between the measurements with the aim of creating a virtual fine discrete time grid for which the assumption in (*) is valid. The bridges are treated as missing or latent data. Let Y^* denote the collection of all latent data. We wish to sample from the joint posterior $f(\theta, Y^* | Y)$ of the parameters θ and the latent variables Y^* given the data Y, using the fact that, by Bayes' theorem,

$$\pi(\theta, Y^* | Y) \propto L(Y^*, Y | \theta) \pi(\theta) \tag{2.7}$$

where $L(Y^*, Y|\theta)$ is the approximated augmented likelihood. This is achieved by sampling in turn from the full conditional densities of $\theta|Y^*, Y$ and $Y^*|\theta, Y$ [Tanner and Wong, 1987]. Thus, in the framework of an MCMC, one can generate proposal bridge processes and accept these with an appropriately constructed acceptance probability. In practice we have used (see [Heron et al., 2007]) a bridging method based on an independence sampler suggested by [Elerian et al., 2001](see supplementary section 2.8). The treatment of other forms of missing data such as unobserved variables as part of the inference algorithm is theoretically the same. In practise, this is challenging as the dimension of the posterior density in (2.7) can become very large. We present applications of bridge building and stochastic reconstruction of unobserved processes in our case studies. One also needs to decide upon the size of a virtual sampling interval for which one can safely assume that (*) holds. Since there are no analytical results we base our choice on Monte Carlo studies of simulated systems.

2.5 Case studies

2.5.1 Case study 1: Red light pulse Experiment

The Arabidopsis thaliana gene Chlorophyll A/B binding Protein 2 *CAB2* is regulated by light and the circadian clock [Millar and Kay, 1996]. The aim here is to estimate degradation rate of *CAB2* mRNA and to reconstruct the transcriptional dynamics of the *CAB2:LUC* reporter gene as a result of a 20 min red-light induction. At subjective dawn on the 6th day of the experiment (see supplementary section 2.8 for a description of experiment), the grown Arabidopsis seedlings were given a 20 min red light pulse to induce *CAB2* expression. Samples were harvested at the indicated time-points and total-RNA and -protein was extracted. Steady state levels of *LUC* mRNA were measured by Quantitative PCR (Q-PCR) and an in vitro LUC assay was used to measure LUC activity in the protein samples. Concurrently, red light pulsed seedlings were also imaged for LUC activity using light sensitive cameras [Millar et al., 1995]. This allows the measurement of LUC activity within the same seedlings throughout the entire experiment, whereas the in vitro LUC assays and Q-PCR experiments necessarily sacrificed different samples for each time-point. All data are probes from whole leaves (plots of all time series in supplementary section 2.8) representing cell populations and the activity of the clock gene can be assumed to be synchronized between cells by the light pulse. There are three replicates of each measurement variable sampled every half hour for a length of seven hours. Matching control replicates that have not been subject to light induction were sampled for the same time length albeit more sparsely for the Q-PCR and in vitro assay data. Assuming that molecular populations all scale differently with the Q-PCR, in vitro and in vivo imaging data we use (2.1) to describe the dynamics of mRNA and imaged LUC protein and add a third equation

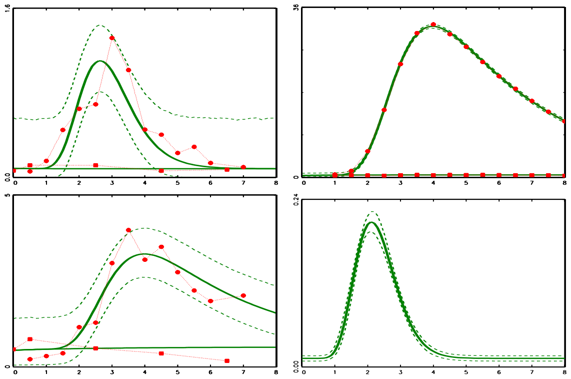

Figure 2.1: This figure shows mean ODE fit for average data (data points given by big dots) of red light pulse experiment. *LUC* mRNA (top left), LUC activity in vitro (bottom left) and imaging the luminescence from LUC protein (top right) under two experimental conditions: with and without red light pulse. Solid lines give the mean ODE fit using mean posterior estimates for the parameters. The 95 % credible intervals (dashed lines) are shown for the control experiments. The reconstructed transcription profile $\tau(t)$ is shown in the bottom right panel (the area between dashed lines gives 95 % central values of the transcription profile for 10,000 iterations of Markov chain).

$$dP_v/dt = \alpha_{P_v} M(t) - \delta_P P_v(t), \tag{2.8}$$

which represents the protein dynamics measured by the in vitro LUC protein assays (see supplementary section 2.8 for full model statement). The two protein equations are identical except for differently scaled translation rates α_P and α_{P_v}.

Furthermore a constant c_P is added to the imaging data to represent some threshold level at which the camera is able to detect a signal. To specify a form for the transcription $\tau(t)$ consider an indicator function $L(t) = 1$ for the time of the red light pulse, and $L(t) = 0$ otherwise ($L(t) = 0$ for all control experiments). The response of mRNA transcription to the stimulus can then be modeled as a convolution of $L(t)$ and $d(u)$ which is a probability density for the waiting time u between the pulse and the initiation of transcription *i.e.* ,

$$\tau(t) = \alpha_M \left(\int_0^{\infty} d(u)L(t-u)du + \overline{\tau} \right), \tag{2.9}$$

where $\overline{\tau}$ represents a baseline transcription. We take $d(u)$ to be a Gamma density with mean μ_Γ and standard deviation σ_Γ to be estimated. The specification in (2.9) is motivated by the fact that it successfully reproduced the qualitative features observed in the data in preliminary model simulations and because d is flexible. Since data are from aggregated cell populations, the imaged protein data is very smooth and successive data points of the Q-PCR and in vitro time series come from different samples of cell populations, we choose to fit the model using the mean ODE approach with independent error. To ensure all variables are strictly non-negative we used an independent Gamma distribution for g in the likelihood (2.5) for each of the three variables where parameters were specified to have mean process equal to an ODE solution and time constant variance $\sigma_M^2, \sigma_P^2, \sigma_{Pv}^2$. Applying (2.5) the likelihood of replicate $r = 1, 2, 3$ is

$$L^r(\boldsymbol{\theta}^r|\mathbf{Y}^r) = \prod_{i=1}^{T^R} g(\mathbf{y}_i^{r,R}|\mu(t_i), \theta^r) \prod_{j=1}^{T^C} g(\mathbf{y}_j^{r,C}|\mu(t_j), \theta^r), \tag{2.10}$$

where $\mathbf{y}_i^{r,R}$ is the vector of observed data points $i = 1, ..., T^R$ for variables M, P, P_v for replicate r under the red light experiment, $\mathbf{y}_j^{r,C}$ denotes observed data points $j = 1, ..., T^C$ for the corresponding control experiment and g is a product of Gamma densities. The ODE model was fitted to each of the replicates $r = 1, 2, 3$ and to the average of the replicates where prior distributions for all parameters were chosen to be uninformative. Results of posterior estimates are summarized in Table 2.1 and the model fit can be seen in Figure 2.1. The mean delay time between light induction and transcription is about 2h with almost all transcription happening between 0.8h and 3.2h after the pulse. Convergence of the Markov chains for parameters associated with the Gamma delay is relatively quick and precise. Chains for α_M and δ_M are correlated and convergence for these is slower. The half-life of *LUC* mRNA is estimated to be around 0.5 hours with some small variation between replicates. In contrast the chains for δ_L converged quickly due to the abundance and smoothness of the imaging data. Protein half-life was estimated to be around 2 to 2.5 hours. Although the control data do not seem very dynamic they are useful in inferring the base rates of transcription and translation. If the control series are omitted from the analysis these rates were estimated with considerably less precision and slower convergence due to correlations.

Parameter	average	r1	r2	r3
δ_M	1.542 (0.019)	1.726 (0.044)	1.417 (0.121)	3.526 (0.315)
(half-life)	0.45 h	0.4 h	0.49 h	0.2 h
μ_Γ	2.008 (0.011)	2.101 (0.014)	1.902 (0.045)	2.362 (0.0289)
σ_Γ	0.631 (0.013)	0.692 (0.014)	0.686 (0.039)	0.723 (0.0217)
$\overline{\tau}$	0.012 (0.001)	0.014 (0.001)	0.014 (0.002)	0.013 (0.002)
δ_P	0.305 (0.0045)	0.286 (0.0040)	0.272 (0.010)	0.365 (0.0093)
(half-life)	2.27 h	2.42 h	2.5 h	1.9 h

Table 2.1: Case 1: Posterior results for selected parameters. Posterior means and standard deviations of selected estimated parameters (see supplementary section 2.8 for all parameters), where the red light pulse model was fitted to average data and to single replicate data sets denoted by r1, r2, r3. Estimated rates are per hour. Degradation rates are translated into half-lives as follows: half-life (in hours)=ln(2)/degradation rate (per hour).

2.5.2 Case study 2: A Switch model for CCA1

The Circadian Clock Associated 1 (CCA1) gene in Arabidopsis thaliana has been identified as one of the core genes of the circadian clock [Wang and Tobin, 1998]. In this case study we show results for the reconstruction of an ON/OFF switching transcription profile from the following two experimental data sets:

(1) Native mRNA Q-PCR data: Q-PCR measurements were taken at 2 h intervals over 72 h on CCA1 mRNA entrained under a photoperiod of 18 hours before being released into constant light. The data used are an average of concentrations relative to the start of two biological replicates.

(2) Protein imaging: High resolution imaging data for a different experiment with identical conditions as for data (1) was sampled at 1.5h intervals over a length of 91.5 h on LUC protein activity resulting from *LUC* reporter constructs fused to the CCA1 promoter. Similar to case study 1 all data come from whole leaves and thus represent a population of cells where the activity of the clock gene is synchronized between cells during the exposure to dark, light cycles during the entrainment period (see supplementary section 2.8 for further details of experiment). The data used are an average of concentrations relative to the start of 20 replicates[2].

No data were available for the *CCA1:LUC* mRNA. However, if we assume that *CCA1:LUC* and *CCA1* mRNA have the same transcriptional dynamics, then the available two time series are connected in a dynamic model with 3 variables where *LUC* mRNA and LUC protein dynamics are described by (2.1) and a further equation

$$dM_g/dt = \tau(t) - \delta_{M_g} M_g(t) \tag{2.11}$$

is added for the native *CCA1* mRNA. We assume that observed variables are proportional to M_g and P populations with scaling factors s_{M_g} and s_P, while M is unobserved. To describe the oscillatory nature of the data we consider an ON/OFF switching function for the transcription $\tau(t) = \tau_{\text{on}}$ if transcription is active at time t, and $\tau(t) = \tau_{\text{off}}$ if transcription is inactive. This function has the advantage of being interpretable and parsimonious. If it produces realistic oscillations then its simple structure makes it an interesting ingredient to models of larger networks. Let $Sw = (s_1, ..., s_R)$ where $s_1 < s_2 < ... < s_R$ are the times at which a switching between an ON and OFF state occurs. They are estimated as part of the MCMC algorithm where we assume that here the number of switches and the initial state are known[3]. To set the phase of the clock both data series experienced a light-dark (LD) cycle of 18 h of L and 6 h of D at the beginning of the sampling period and this seems to generate a higher amplitude. We allow for this by setting the

[2] For computational precision we amplified the mRNA concentrations by factor 10^5 and the protein concentrations by 10^4.

[3] The number of switches and initial state are fairly obvious here. The inference algorithm can however be generalized to allow for an arbitrary number of switches and where the initial state is estimated. We will describe work on this elsewhere.

transcription on-rate to $p_d\tau_{on}$ during the first 35 hours (allowing also for some delayed effect of the dark period). For purpose of estimation, the mean ODE approach will be appropriate for similar reasons as case study 1. However, an SDE approach is a superior theoretical model that should be considered even if data do not (yet) strictly comply with its underlying assumptions. We use this case study to show the application of both approaches.

SDE approach: Consider a system of SDEs formulated analogously to (2.2). Since M is unobserved it can be imputed stochastically as realizations of the SDE but the cost of computation is high. Simulation studies suggested that the more practicable way of imputing M as solution to an ODE from an initial condition M_0 to be estimated had no discernable impact on our inference results here. In order to fit an SDE model to discrete data points for M_g and P we augment the coarse grid to a virtually fine grid (for which assumption (*) is valid) by imputing auxiliary data in the form of bridges. Let $\theta = (Sw, \tau_{on}, \tau_{off}, \delta_{M_g}, M_0, \delta_M, S_{M_g}, \alpha, \delta_P, S_P)$ denote the vector of unknown parameters and let M_g^* and P^* be the auxiliary data for M_g and P, respectively. Then according to (2.7) the posterior distribution for the unknown Θ, M_g^*, P^* is given by

$$\pi(\theta, M_g^*, P^*|M_g, P) \propto L(M_g, P, M_g^*, P^*|\theta)\pi(\theta),$$

where we approximate $L(M_g, P, M_g^*, P^*|\theta)$ with the augmented likelihood in (2.4) for small sampling intervals for all observed and auxiliary data, *i.e.* $\mathbf{y} = (M_g, P, M_g^*, P^*)$. More details of the SDE inference algorithm are provided in supplementary section 2.8.

Mean ODE approach: Here the likelihood is given by (2.5) where the unobserved variable M is reconstructed as a solution of an ODE from an initial condition M_0 to be estimated. The density g was specified to be the product of two independent normal distributions with mean equal to the joint ODE solutions for M_g and P and with variance parameters $\sigma^2_{M_g}$ and σ^2_P. We have set $\tau_{\text{off}} = 0$ for the off-time as initial estimations showed that it was not different from zero[4]. As the

[4] We could not set $\tau_{\text{off}} = 0$ in the SDE case for the practical problem that the bridge building algorithm becomes numerically unstable for values of the mRNA too close to zero. It is because according

variables are concentrations relative to initial conditions the ODE solutions are assumed to start at one. Thus, the parameter vector for the mean ODE approach is $\theta = (Sw, \tau_{\text{on}}, \tau_{\text{off}}, \delta_{M_g}, M_0, \delta_M, \alpha, \delta_P, \sigma_{M_g}, \sigma_P)$.

To ensure identifiability in both estimation approaches the prior distribution for *CCA1:LUC* mRNA degradation δ_M has to be informative. We hence used a Gamma distribution with mean 1.542 and standard deviation 0.019, corresponding to the results in Table 2.1. All other priors were taken independently uniform in an attempt to estimate all remaining parameters only from the experimental data at hand. Posterior estimates are given in Table 2.2. Fig. 2.2 shows the transcription profiles and model fits for both approaches. The plots suggest that the switch model is remarkably able at reproducing the observed oscillations. The main feature of the reconstructed profiles is that the inactive times (around 15-18) hours are at least twice as long as the active times (around 7 hours) and this produces the pronounced asymmetric cycles in the protein and mRNA time series. The estimates also suggest that there is a shorter but larger burst of transcription during the dark period. Both approaches deliver similar posterior rates for degradation. Our results for *CCA1* mRNA degradation are in remarkable agreement with the analysis in [Yakir et al., 2007] whose estimates correspond to 0.23 in darkness to 0.46 in light for δ_{M_g}. Both approaches reliably estimate the half-life of the LUC protein to be around 9.5 h. This is surprisingly long and is probably due to a lack in provision of luciferin. The most notable difference between the two approaches lies in the variance estimation. The SDE approach has to deal with the estimation of the two scaling parameters, s_P and s_{M_g}. We find that their identification from the experimental data is problematic as convergence could not achieved although this did not affect convergence of all other parameters. The two scaling parameters were thus sampled within some chosen bounded region of parameter space. In particular in order for the bridge sampling to remain numerically stable for low values of the mRNA series, the sampling of s_{M_g} had to be bounded to artificially low values. The identifiability problem of the scaling parameters leads to

to equations 2.2 molecular concentrations can take negative values. It is more likely to happen for small values of τ_{off}.

problems in realistically quantifying the volatility. The estimated intervals in Fig. 2.2 illustrate this for the mRNA series. For the mean ODE approach variability is measured by the posterior standard error of the fit similar to a regression and the graph shows that predictions can be made more precisely about the protein dynamics than about the native mRNA. This is reflecting the fact that the protein data is a more aggregated and smoother time series than the mRNA series.

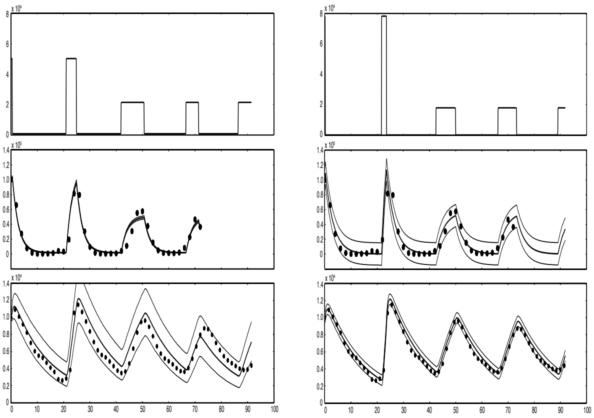

Figure 2.2: Results of fitting SDEs (left) and ODEs (right) in case study 2. Top panel shows the mean reconstructed transcription profile $\tau(t)$ using the switch approximation. Middle panel shows results for M_g. Bottom panel gives results for P. Big dots are experimental data for M_g (middle panel) and P (bottom panel). The variation is shown as follows: For SDE approach (left): solid lines in middle and bottom panel give the 5 % , mean and 95 % values computed from 10,000 simulations of the SDE (using mean posterior parameter estimates). For ODE approach (right): Solid lines corresponds to the mean ODE fit (using mean posterior parameter estimates) plus/minus twice the mean posterior standard error.

	δ_{M_g}	δ_M	δ_P
SDE	0.426 (0.0043)	1.54 (0.019)	0.072 (0.0057)
ODE	0.313 (0.0273)	1.42 (0.101)	0.075 (0.0018)

Table 2.2: Case 2: Posterior results for selected parameters. Posterior mean and standard error estimates of selected parameters of model in case 2 using the SDE and mean ODE approach. All rates are per hour. Estimates for all parameters and switch-times are provided in supplementary section 2.8.

2.5.3 Case study 3: Stochastic transcription for single cell data

In this experiment protein activity was imaged from GH3 rat pituitary cells stably transfected with a construct comprising a 5kb human prolactin gene promoter fragment linked to a destabilized EGFP reporter gene (hPRL-d2EGFP) (see supplementary section 2.8 for details of experiment). Images were taken 108 times in 15 minutes intervals giving a total of 27 hours of data for a single cell (see Figure 2.3). We assume that the dynamics are described by the SDE model in (2.2). Since M is not observed we cannot identify the degradation rates (δ_M, δ_P) and a strongly informative prior density is needed. Here we assume that they each have an independent Gamma distribution with mean 0.4 for δ_M and 0.5 for δ_P[5]. The prior variance was arbitrarily chosen to be small at 0.02 for both parameters. Since M is unobserved we can arbitrarily fix $s_M = 1$. Given the bell-shape of the time series obtained in the experiment (see Figure 2.3), where transcription is induced and subsequently returns to its initial level, we have specified $\tau(t)$ as follows

$$\tau(t) = \begin{cases} b_0 \exp(-\frac{(t-b_3)^2}{b_1}) + b_4 & t \leq b_3 \\ b_0 \exp(-\frac{(t-b_3)^2}{b_2}) + b_4 & t > b_3, \end{cases} \tag{2.12}$$

where the parameters b_i are to be estimated. This allows the on step and off step width to be different as there is no reason to assume that these two should be equal. Priors for parameters different than degradation rates were intended to be uninformative. Here we used exponential prior with means given in Table 2.3. The challenge for inference here is to integrate over a fully unobserved process M whilst sampling bridges to augment the discretely observed P. Let P^* denote the vector of bridges augmenting the P process and M^* denote the latent M variable (we chose a grid-size of 1 min for which we assume that (*) holds). The vector of unknown parameters is $\theta = (\delta_M, \delta_P, \alpha, s_P, b_0, b_1, b_2, b_3, b_4)$. The posterior distribution takes the form

$$\pi(\Theta, M^*, P^*|P) \propto L(M^*, P^*, P|\Theta)\pi(\Theta) \tag{2.13}$$

[5] These rates were motivated by preliminary estimation using a small data set from other experiments. They are used here only to demonstrate the case as their estimates may change if more data were available.

where we approximate $L(M^*, P^*, P|\Theta)$ with the likelihood (2.4) for the augmented data case, *i.e.* $\mathbf{y} = (M^*, P^*, P)$. In practice this is a challenging sampling problem as the dimension of the posterior is very large and traces were highly autocorrelated. Faster convergence is achieved by re-parameterizing the model (details of this and the algorithm are given in supplementary section 2.8). The algorithm was first tested on simulated data from the SDE model with chosen parameters (see Table 2.3). Artificial data are simulated on a fine scale of 15/51 minutes and coarse data are extracted for P at 15 min intervals. The simulated and observed time series, and the reconstructed $\tau(t)$ are shown in Fig. 2.3. Posterior inference results are given in Table 2.3. Note that since M is not scaled the transcription profile corresponds to molecular population sizes which here are about 150 mRNA molecules per hour. This case study demonstrates that for high frequency single cell data the SDE approach can be extremely powerful as it allows estimation of absolute transcription rates in terms of molecule numbers and since s_P can be estimated it is possible to calculate back to molecular levels of protein and translation rate. The need for precise prior information about degradation rates is irrespective of either SDE or ODE approach. The problem of non-identifiability of these parameters is due to not observing M as one can infer both degradation rates in either approach if both M and P are observed.

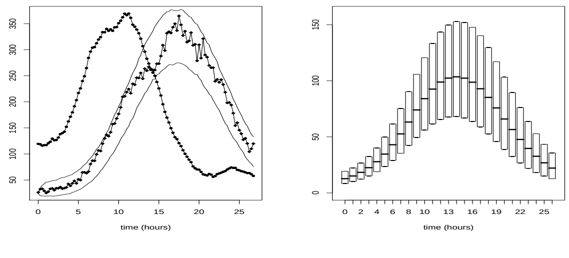

Figure 2.3: Left: Time series of fluorescence intensity used in case study 3. Solid and dashed lines represent experimental and simulated data, respectively. The variation of the SDE fit to the real data is shown by the 5 % and 95 % values computed from 1,000 simulations of the SDE (using mean posterior parameter estimates). Right: Box-plot representing transcription profile in molecules per hour inferred from experimental data presented in the top figure. Each box represents 50% credibility interval and median of posterior distribution of the reconstructed transcription rate at particular time point.

	value	prior	Simulation	Experiment
δ_M	0.44	Γ(0.44,0.02)	0.56 (0.36 - 0.92)	0.45(0.26 - 0.82)
δ_P	0.52	Γ(0.52,0.02)	0.59 (0.38 - 0.89)	0.71 (0.45 - 1.09)
α	20	Exp(100)	16.97 (6.54 - 78.98)	0.46 (0.14 - 1.51)
s_P	0.2	Exp(1)	0.17 (0.09 - 0.3)	2.11 (1.24 - 3.56)

Table 2.3: Case 3: Posterior inference results. Parameter values used in simulation study. Priors, posterior medians and 95% credibility intervals inferred from both simulated and experimental data. Rates are per hour. $\Gamma(\mu, \sigma^2)$ denotes gamma distribution with mean μ and variance σ^2. Full list of all parameter estimates is provided in supplementary section 2.8.

2.6 Discussion

In this study we suggest a dynamical model relating protein and corresponding mRNA dynamics via transcription and translation and suggest methods for model fitting. The applications here were motivated by the availability of gene reporter data but the model and methodology apply to many other scenarios where it is of interest to link protein and mRNA dynamics. While a stochastic model such as (2.2) applies to single cell data, caution needs to be exercised in formulating an ODE model such as (2.1) for multi-cell data. In order to reasonably assume such a joint mechanistic model it is essential that the individual cell activities are synchronized with respect to the gene of interest. Rate constants associated with processes of degradation, transcription and translation arise as model parameters and it is an important question whether these can be identified. In addition to a functional kind of non-identifiability of parameters in complex dynamic models as considered in [Hengl et al., 2007] here, we find that practical or statistical non-identifiability of model parameters may result from unobserved variables. Case study 1 demonstrates that one can estimate all rate constants in systems of equations of the type given in (2.1) if all model variables - albeit coarse - are observed over time. Inference precision increases with the frequency at which the processes are sampled. In contrast, Cases 2 and 3 have latent variables and model inference is only feasible with informative prior knowledge of some parameters. Simulation studies of the model (using artificial parameters) help in identifying which sets of parameters need to be informed from other experiments. In case 3 prior knowledge of both degradation rates was needed as with M unobserved, parameters can trade-off giving rise to protein dynamics that is virtually indistinguishable via likelihood from the observed protein process. The specification of the functional form for the transcription profile also plays a role in practical identification. Even if M is observed the parameter estimates associated with transcription and degradation are correlated for obvious reasons. Such correlations affect precision of estimates and convergence of the Markov chain but can be alleviated by sampling more frequently, choosing a parsimonious functional form for transcription, and by technical aids such as the construction of independence samplers and re-parameterization of the model. We believe that the functional specifications for $\tau(t)$ suggested in our case studies are useful in conjunction with gene transcription. A theoretical application of the switch function in clock modeling can be found in [Aase and Ruoff, 2008]. Although the estimation of the switch model seems too high dimensional for data sets with many switches this could be overcome by assigning probability distributions to the on- and off times in the framework of a Bayesian hierarchical model.

Our results demonstrate that MCMC methods for ODEs and SDEs provide practical algorithms for reconstruction transcription profiles whilst estimating some of the rate parameters involved. As the real population dynamics are naturally stochastic SDEs provide the superior theoretical model. However the mean ODE approach can be useful as a vehicle for estimation when the data are not fully compatible with the SDE assumptions. Whilst they usually describe the same model in the mean, their difference lies in the specification of the variance. The SDE model provides a rigid description of the volatility process which is rigorously derived for the stochastic dynamics of the molecular processes. In theory it is straightforward to allow for additive measurement error (see [Heron et al., 2007] for estimation of SDEs with measurement error). However, identification of an unknown measurement error variance is difficult and - to our knowledge - is not possible when the data are coarse and indirectly measured with unknown scaling factors. The variance process of the mean ODE approach is not rigorously derived and can be specified by the modeler in an attempt to capture anything known about the residual process and measurement error. Estimation algorithms for the mean ODE approach are straightforward to implement although for higher dimensional or less stable systems more difficulties may occur. The algorithm for SDE estimation can be challenging to implement due to bridge sampling and is computationally expensive. Case 2 shows a problem that we have also encountered in [Heron et al., 2007], namely if molecular populations are measured indirectly then the estimation of unknown scaling parameters can be difficult in practise. This may happen as a consequence of observing data that are too coarse, in the sense that too little information about the volatility process is revealed, or that are otherwise not directly compatible with the SDE assumption. However, drawbacks of the SDE approach are associated with the current quality, quantity and availability of the data. Case study 3 exemplifies that SDE estimation constitutes a very informative approach in calibrating all processes back to the molecular population levels as the scaling parameters can be identified. Under suitable assumptions the SDE model provides a theoretically well founded modeling approach for describing the dynamics of molecular populations in a single cell. Estimation of SDEs is well studied and feasible and is highly informative when relatively frequent and clean (*i.e.* with little measurement error) single cell data are available on all model variables.

2.7 Acknowledgements

This research was funded by BBSRC and EPSRC (Interdisciplinary Programme for Cellular Regulation GR/S29256/01) and EU (BIOSIM Network Contract 005137). The Millar group's work (KE and AJM) was funded by BBSRC via award E015263 and CSBE (Centre for Integrative and Systems Biology funded by BBSRC and EPSRC). DAR is funded by EPSRC Senior Research Fellowship EP/C544587/1 and MK by studentship, Dept of Statistics, University of Warwick. CVH was funded by Wellcome Trust Programme Grant (067252, to JRED and MRHW) and

now is recipient of The Prof. John Glover Memorial Postdoctoral Fellowship.

2.8 Supplementary Information

This section contains details of mathematical models and statistical methods used in the previous sections of this chapter.

2.8.1 Scaled Model

Suppose that we measure M, P indirectly through variables $\tilde{M}(t) = s_M M(t)$ for the mRNA and $\tilde{P}(t) = s_P P(t)$ for the reporter protein. Re-formulating model (2.1) in section 2.4 gives a scaled model

$$\begin{aligned} d\tilde{M}/dt &= \tilde{\tau}(t;\theta_\tau) - \delta_M \tilde{M}(t) \\ d\tilde{P}/dt &= \tilde{\alpha}\tilde{M}(t) - \delta_P \tilde{P}(t), \end{aligned} \tag{2.14}$$

where the transcription function $\tilde{\tau}(t;\theta_\tau) = s_M \tau(t;\theta_\tau)$ and the translation rate $\tilde{\alpha} = \frac{s_P}{s_M}\alpha$ are now functions of the unknown scaling coefficients s_M, s_P. Obviously, the functional forms of (2.14) and model (2.1) in section 2.4 are identical. If we set $M = \tilde{M}, P = \tilde{P}, \tau = \tilde{\tau}$ and $\alpha = \tilde{\alpha}$ then (2.1) in section 2.4 denotes the scaled model.

2.8.2 Diffusion Approximation

Let $p_t(M, P)$ denote the probability that at time t system is in the state (M, P). The evolution of the joint probability is described by the chemical master equation of the form (see [Thattai and van Oudenaarden, 2001] for derivation)

$$\begin{aligned} \frac{dp_t(M,P,t)}{dt} &= \tau(t)(p_t(M-1,P) - p_t(M,P)) \\ &+ \alpha M(p_t(M,P-1) - p_t(M,P)) \\ &+ \delta_M(p_t(M+1,P)(M+1) - p_t(M,P)M) \\ &+ \delta_P(p_t(M,P+1)(P+1) - p_t(M,P)P). \end{aligned} \tag{2.15}$$

In order to obtain a diffusion approximation of (2.15) we replace increments on the right hand of the above with their second order Taylor expansions. This gives the Fokker-Planck equation

$$\begin{aligned}\frac{dp_t(M,P)}{dt} &= -\frac{\partial}{\partial M}(\tau(t)-\delta_M M)p_t(M,P)\\ &- \frac{\partial}{\partial P}(\alpha M-\delta_P P)p_t(M,P)\\ &+ \frac{1}{2}\frac{\partial^2}{\partial M^2}(\tau(t)+\delta_M M)p_t(M,P)\\ &+ \frac{1}{2}\frac{\partial^2}{\partial P^2}(\alpha M+\delta_P P)p_t(M,P).\end{aligned}$$

This method is called Ω (size) expansion and gives valid approximation for system of large volume (see [Golightly and Wilkinson, 2005, Van Kampen, 2006] for details). The Fokker-Planck equation describes the evolution of the probability densities of the stochastic process governed by the Itô diffusion [Gardiner, 1985]

$$\begin{aligned}dM &= (\tau(t)-\delta_M M)dt+\sqrt{\tau(t)+\delta_M M}dW_r \qquad (2.16)\\ dP &= (\alpha M-\delta_P P)dt+\sqrt{\alpha M+\delta_P P}dW_p,\end{aligned}$$

where dW_r, dW_p are increments of independent Wiener processes and thus their joint distribution is bivariate normal as stated in (*) in section 2.4. [Higham, 2001] gives an accessible algorithmic introduction to stochastic differential equations as in (2.16) and Wiener processes.

2.8.3 Supplementary information for Case study 1

Experiment

LUC reporter constructs fused to the *CAB2* promoter *CAB2:LUC*, have allowed the characterization of *CAB2* expression during high resolution imaging time-courses [Millar et al., 1995]. We used the well characterized induction of *CAB2* expression by light to study the activity of *LUC* in plants. Arabidopsis seed containing the *CAB2:LUC* reporter gene were given a 6h white light pulse to induce germination and then grown in constant darkness for 4 days. Seedlings were grown in darkness to reduce the basal level of *CAB2:LUC* expression. Since induction of *CAB2* by light is gated by the clock, occurring maximally during the early part of the day [Millar and Kay, 1996], the seedlings were entrained under temperature cycles of 12h at 24 degrees C followed by 12h at 18 degrees C, allowing us to target the light pulse to the relevant time of the day. At dawn on the 5th day the temperature cycles were stopped and the plants were maintained in darkness at 22 degrees C. They were also transferred to liquid media, containing 1mM Luciferin to ensure that the substrate did not become limiting to LUC activity. At subjective dawn on the 6th day (24h after the transfer to constant temperature; referred to as time 0 in the text and figures), the seedlings were given a 20min red light pulse to induce *CAB2* expression. Samples were harvested at the indicated time-points and total-RNA and -protein was extracted. Steady state levels of *LUC* mRNA were measured by Quantitative PCR (Q-PCR) and an in vitro LUC assay (Promega, Madison,WI, USA) was used to measure LUC activity in the protein samples. Concurrently, red light pulsed seedlings were also imaged for LUC activity using light sensitive cameras ([Millar et al., 1995]). This allows the measurement of LUC activity within the same seedlings throughout the entire experiment, whereas the in vitro LUC assays and Q-PCR experiments necessarily sacrificed different samples for each time-point.

Model

Assuming that molecular populations scale differently with the Q-PCR, in vitro and in vivo imaging data we use the following equations based on model (2.1) in section 2.4

$$\frac{dM}{dt} = \tau(t) - \delta_M M(t), \tag{2.17}$$

$$\frac{dP}{dt} = \alpha_P M(t) - \delta_P P(t) + c_P, \tag{2.18}$$

$$\frac{dP_v}{dt} = \alpha_{P_v} M(t) - \delta_P P_v(t), \tag{2.19}$$

The additional variable P_v represents the protein dynamics measured via the in vitro LUC protein assays. Both protein equations have identical degradation rates δ_P and translation proportional to $M(t)$ but with differently scaled translation rates α_P and α_{P_v}. Preliminary estimations also showed that the observed near zero levels of the control imaging data are only compatible with the higher control levels of the in vitro protein and RT-PCR mRNA if a constant c_P is added to the imaging data.

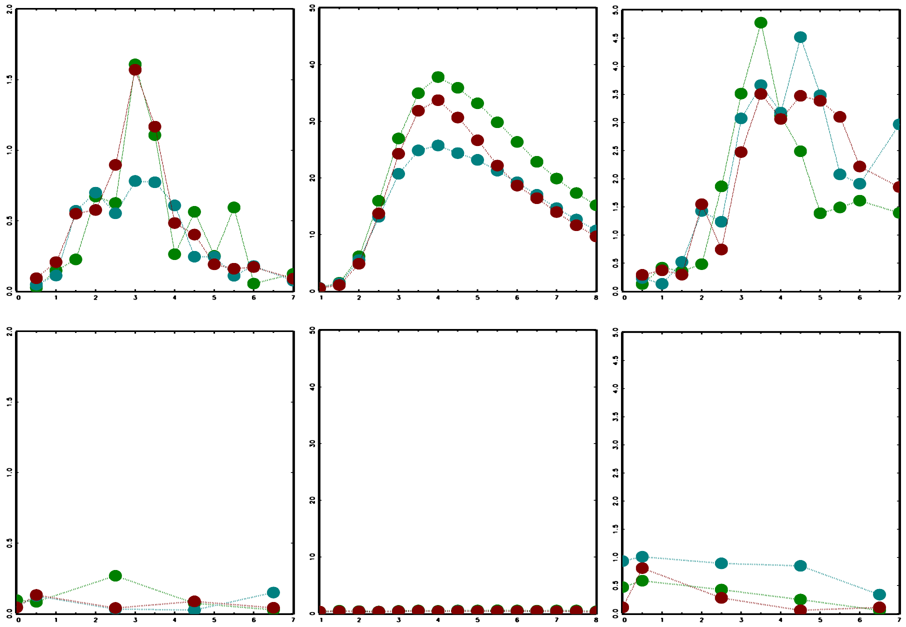

Figure 2.4: Plot of data for case study 1. Left: *LUC* mRNA Q-PCR, middle: imaging of LUC protein, right: in-vitro LUC assay, top row: experiment with red light stimulus during first 20 minutes, bottom row: un-stimulated control experiments. There are three replicates. A big dot corresponds to an observed data point.

2.8.4 Supplementary information for Case study 2

Experiment

Circadian regulation is normally tested by entraining the organism to 12h light: 12h dark cycles, then transferring the organism to constant conditions. Transgenic Arabidopsis seed were sterilised and grown as described previously [Gould et al., 2006], for 4 days at 22^o C in Sanyo MLR350 environmental test chambers (Sanyo, Osaka, Japan) under photoperiods of 75μ moles.m-2s-1 cool white fluorescent light. Seedlings were then transferred to Percival I-30BLL growth chambers (CLF Plant Climatics, Emersacker, Germany) at dawn on the 5th day and grown at 22^o C under an equal mix of Red and Blue LEDs at 20-30μ moles.m-2s-1, with 18h light: 6h dark photoperiods. In the data shown in Figure 2.2, time 0 is the time of lights-on on the 7th day of growth. *CCA1:LUC*+ plants have been described in [Doyle

Parameter	average	r1	r2	r3
α_M	10.43 (0.232)	9.62 (0.179)	8.36 (0.538)	25.38 (1.43)
δ_M	1.542 (0.019)	1.726 (0.044)	1.417 (0.121)	3.526 (0.315)
(half-life)	0.45 h	0.4 h	0.49 h	0.2 h
μ_Γ	2.008 (0.011)	2.101 (0.014)	1.902 (0.045)	2.362 (0.0289)
σ_Γ	0.631 (0.013)	0.692 (0.014)	0.686 (0.039)	0.723 (0.0217)
$\overline{\tau}$	0.012 (0.001)	0.014 (0.001)	0.014 (0.002)	0.013 (0.002)
α_P	25.07 (0.386)	34.90 (0.555)	23.02 (1.417)	24.83 (1.78)
δ_P	0.305 (0.0045)	0.286 (0.0040)	0.272 (0.010)	0.365 (0.0093)
(half-life)	2.27 h	2.42 h	2.5 h	1.9 h
α_{Pv}	2.178 (0.107)	2.141 (0.210)	2.534 (0.222)	2.152 (0.183)
c_P	-1.868 (0.198)	-2.637 (0.144)	-1.763 (0.208)	-2.121 (0.388)
σ_M	0.159 (0.0026)	0.254 (0.0018)	0.134 (0.0069)	0.171 (0.005)
σ_P	0.183 (0.0133)	0.182 (0.0125)	0.286 (0.0287)	0.331 (0.030)
σ_{Pv}	0.364 (0.0211)	0.769 (0.0222)	0.562 (0.0129)	0.442 (0.034)

Table 2.4: Posterior means and standard deviations of all estimated parameters where the red light pulse model was fitted to average data and to single replicate data sets denoted by r1, r2, r3. Estimated rates are per hour. Degradation rates are translated into half-life as follows: half-life (in hours)=ln(2) /degradation rate (per hour). $\sigma_M, \sigma_P, \sigma_{Pv}$ give estimated posterior standard error for each model equation.

et al., 2002]. Luciferase imaging was carried out as previously described [Gould et al., 2006] using Hamamatsu C4742-98 digital cameras operating at -75^o C under control of Wasabi software (Hamamatsu Photonics, Hamamatsu City, Japan). Bioluminescence levels were quantified using Metamorph software (MDS, Toronto, Canada). Experiments included 22 individuals of each genotype and were replicated 4 or more times. For Q-PCR experiments, wild type Wassilewskija (Ws) seedlings were grown for 7 days in Percival Growth chambers under experimental photoperiods of 60-65μMolesm-2s-1 cool white fluorescent light. Seedlings were harvested, RNA was extracted and reverse transcribed as described previously [Locke et al., 2005b]. Quantitative PCR was carried out in 384-well format using SYBR Green JumpStart *Taq* ReadyMix (Sigma, Gillingham, UK) in technical triplicate with a LightCycler 480 instrument (Roche, UK), using the Relative Quantification function to measure mRNA abundance. Expression values were normalised against *ACTIN 2 (ACT2). ACT2* and *CCA1.* PCR primers have previously been described [Locke et al., 2005b].

Model

The model for case study 2 is

$$\begin{aligned}\frac{dM_g}{dt} &= \tau(t) - \delta_{M_g} M_g(t), && (2.20)\\ \frac{dM}{dt} &= \tau(t) - \delta_M M(t), && (2.21)\\ \frac{dP}{dt} &= \alpha M(t) - \delta_P P(t), && (2.22)\end{aligned}$$

where (2.21) and (2.22) are as in model (2.1) in section 2.4, describing the dynamics of the luciferase reporter mRNA and protein, respectively. Equation (2.20) formulates the transcription and degradation of the native gene *CCA1* mRNA for which we have coarse Q-PCR data. Equations (2.20)-(2.22) are at the population level. We assume that the observed variables are proportional to M_g and P with scaling factors s_{M_g} and s_P, respectively, whilst M is unobserved. Equations 2.20 and 2.22 represent an equivalent parameterization of the scaled variables with $\tilde{\tau}(t) = s_{M_g}\tau(t)$ and $\tilde{\alpha} = (s_P/s_{M_g})\alpha$ replacing the transcription and translation coefficient, respectively. For ease of notation we re-use $\tau(t)$ and α.

MCMC algorithm for inference using SDE approach

1. Set iteration counter $i = 0$. Initialise parameters $\theta^{(i)}$ and all bridges M_g^* and P^*.

2. Set $i = i + 1$

3. Update $\tau_{\text{on}}^{(i)}, \tau_{\text{off}}^{(i)}$ and $\delta_{M_g}^{(i)}$ in one block. Use individual random walk Metropolis proposals and either all are accepted or all are rejected. If the proposals are accepted update $M^{(i)}$ and $\tau^{(i)}$.

4. Similarly, update $S_w^{(i)}$ in one block using a random walk Metropolis step.

5. Update $s_{M_g}^{(i)}$ using a random walk Metropolis step.

6. Update $\delta_M^{(i)}$ using a random walk Metropolis step. If the proposal is accepted update $M^{(i)}$ and $\tau^{(i)}$.

7. Update $M_0^{(i)}$ using a random walk Metropolis step. Update $M^{(i)}$.

8. Update $\alpha^{(i)}$ and $\delta_P^{(i)}$ in one block using an independence sampler.

9. Update $s_P^{(i)}$ using a random walk Metropolis step.

10. Sample M_g^* bridges (using the method in [Elerian et al., 2001]).

11. Similarly, update P^* bridges.

12. Repeat from Step 2 until a sufficient sample from the converged chains has been obtained.

Parameter	SDE	ODE
δ_{M_g}	0.426 (0.0043)	0.313 (0.0273)
δ_M	1.54 (0.019)	1.42 (0.101)
α	0.095 (0.0090)	0.113 (0.0050)
δ_P	0.072 (0.0057)	0.075 (0.0018)
$M(0)$	19508 (13880)	13487 (6950)
τon	21401 (331)	17865 (1366)
τ_{off}	651 (33.85)	0
p_d	2.35 (0.05)	4.39 (0.58)
s_{M_g} (SDE), σ_{M_g} (ODE))	24.99 (0.013)	7560 (188)
s_P (SDE), σ_P (ODE)	188.6 (8.922)	290 (11.33)
s_1	0.56 (0.030)	0.26 (0.09)
s_2	21.11 (0.040)	21.85 (0.11)
s_3	25.07 (0.046)	23.76 (0.16)
s_4	42.12 (0.055)	42.74 (0.19)
s_5	50.92 (0.052)	50.28 (0.19)
s_6	66.92 (0.063)	66.53 (0.18)
s_7	71.80 (0.133)	73.66 (0.26)
s_8	86.96 (2.10)	89.55 (0.22)

Table 2.5: Posterior mean and standard error estimates of parameters and switch-times for the experimental data using the SDE and mean ODE approach in case study 2.

2.8.5 Supplementary information for Case study 3

Experiment

GH3 rat pituitary cells stably transfected with 5kb human prolactin promoter destabilized EGFP reporter construct (hPRL-d2EGFP) were seeded onto 35 mm glass coverslip-based dishes (IWAKI, Japan) and cultured in 10 % FCS for 24 h prior to imaging. Cells were transferred to the stage of a Zeiss Axiovert 200 equipped with an XL incubator (maintained at 37C, 5 % CO2, in humid conditions) and images were obtained using a Fluar x20, 0.75 numerical aperture (Zeiss), air objective. Excitation of d2EGFP was performed using an Argon ion laser at 488nm. Emitted light was captured through a 505-550 nm bandpass filter from a 545 nm dichroic mirror. Images were captured every 15 min. 5 M forskolin and 0.5 M BayK 8644 were added directly to the dish at the start of the experiment. Data was captured and analysed using LSM510 software with consecutive autofocus. Analysis was performed using Kinetic Imaging software AQM6. Regions of interest were drawn around each single cell and mean intensity data was measured 108 times in 15 minutes intervals giving a total of 27 hours of data (see Figure 2.3).

Transformation of parameters

To reduce correlation and improve convergence of the chain we re-parameterized the model in case study 3 as follows

$$\begin{aligned} & h(\Theta, M, P) = (\hat{\Theta}, \tilde{M}, P) & (2.23) \\ = \; & (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4, \hat{\theta}_5, \hat{\theta}_6, \hat{\theta}_7, \hat{\theta}_8, \hat{\theta}_9, \tilde{M}, P) & (2.24) \\ = \; & (log(\delta_M), log(\delta_P), log(s_P \alpha b_0), log(s_P \alpha), log(\alpha), \\ & log(\alpha \ s_P b_4), log(b_3), log(b_1), log(b_2), \tilde{M}, P), \end{aligned}$$

where $\tilde{M} = s_M \ \alpha M - \delta_P P$.

Algorithm for inference

1. Set iteration counter $i = 0$. Initialise all parameters $\hat{\Theta}^{(i)}$, hidden process $\tilde{M}$ and bridges P^*.
2. Set $i = i + 1$
3. Update $(\hat{\theta}_1^{(i)}, \hat{\theta}_3^{(i)}, \hat{\theta}_6^{(i)})$ in one block using multivariate normal proposals and either all are accepted or all are rejected.

4. Separately update each remaining component of $\Theta^{(i)}$ using random walk proposals.

5. Update P^* bridges (using the method in [Elerian et al., 2001])

6. Similarly, update the latent process $\tilde{M}$.

7. Repeat from Step 2 until a sufficient sample from the converged chains has been obtained.

	value	prior	Simulation	Experiment
δ_M	0.44	Γ(0.44,0.02)	0.56 (0.36 - 0.92)	0.45(0.26 - 0.82)
δ_P	0.52	Γ(0.52,0.02)	0.59 (0.38 - 0.89)	0.71 (0.45 - 1.09)
α	20	Exp(100)	16.97 (6.54 - 78.98)	0.46 (0.14 - 1.51)
s_P	0.2	Exp(1)	0.17 (0.09 - 0.3)	2.11 (1.24 - 3.56)
b_0	23	Exp(100)	40.48 (9.11 - 136.3)	112.7 (29.52 - 364.8)
b_1	10	Exp(50)	31.52 (7.08 - 83.64)	54.84 (21.67 - 97.14)
b_2	30	Exp(50)	22.83 (5.82 - 60.13)	59.43 (19.08 - 130.6)
b_3	5	Exp(7)	7.22 (4.35 - 9.55)	13.78 (11.08 - 16.05)
b_4	5	Exp(100)	6.4 (1.55 - 23.68)	7.39 (0.25 - 30.75)
M_0	15	Exp(70)	23.11 (4.8 - 66.92)	31.79 (7.81 - 97.87)

Table 2.6: Posterior inference results for case study 3. Parameter values used in simulation study. Priors, posterior medians and 95% credibility intervals inferred from both simulated and experimental data. Rates are per hour. $\Gamma(\mu, \sigma^2)$ denotes gamma distribution with mean μ and variance σ^2.

Updating Bridges

Suppose data $Y = (Y(t_1), ..., Y(t_T))$ are provided at sampling intervals that are too coarse to allow parameter estimation in the SDE approach without bridging. For example, LUC protein imaging data may be available every 30 minutes while for artificial stochastic process data from simulated clock models we find that a small enough time interval for the normal approximation in (*) in section 2.4 to produce reasonably accurate parameter estimates is 0.1 hour. The methods applied in this study make use of established strategies developed for nonlinear stochastic differential equations [Durham G. B, 2002, Elerian et al., 2001, Eraker, 2001, Kim et al., 1998, Pedersen, 1995]. The basic idea is to augment the observed data by introducing a number of latent data points (called bridges) Y^* in-between the measurements. The bridges are constructed so that the data together with the bridges (augmented data) give a time series with interval length $\Delta t_i = 0.1$ h which we know from simulation studies allows for accurate parameter estimation.

To provide an estimate of the parameters θ from sparsely sampled data, we use MCMC to sample from the joint posterior $f(\theta, Y^*|Y)$ of the parameters θ and the auxiliary variables Y^* given the data Y, using the fact that, by Bayes' theorem,

$$f(\theta, Y^*|Y) \propto L_{\mathsf{SDE}}(Y^*, Y|\theta)\pi(\theta) \tag{2.25}$$

where, as before, $\pi(\theta)$ denotes the prior distribution on θ and $L_{\mathsf{SDE}}(Y^*, Y|\theta)$ is the approximated augmented likelihood. This is achieved by sampling in turn from the full conditional densities of $\theta|Y^*, Y$ and $Y^*|\theta, Y$ ([Tanner and Wong, 1987]). The general structure of the algorithm that we employ is thus as follows:

1. Initialise Y^* by constructing linear bridges between each of the given data points. The parameters θ are initialised as usual.

2. Sample Y_i^* from $Y_i^*|Y(t_i), Y(t_{i+1}), \theta$ for $i = 1, 2, \ldots, T-1$. The two samples constitute a full set of imputed data Y^*.

3. Sample θ from $\theta|Y, Y^*$, i.e. use the fully augmented data to update the parameter vector.

4. Repeat steps 2 and 3 until the required sample is obtained after the chain has converged.

Updating the parameter vector in step 3 is quite straightforward as for a given fully augmented time path the constant rate approximation for (*) in section 2.4 is valid and the inference problem is the same as for a finely sampled time path. To sample Y_i^* in step 2 we use the bridging methodology suggested by [Elerian et al., 2001] which has proved very satisfactory but it should be noted that there exist various other available methods for bridging (see [Durham G. B, 2002] for a survey) that may also be used for this kind of problem. We now briefly describe the bridge building part (step 2) of the algorithm using the [Elerian et al., 2001] sampler. Consider a general SDE of the form:

$$dy(t_i) = \mu(y(t_i), t_i, \theta)dt + \sigma(y(t_i), t_i, \theta)dW, \tag{2.26}$$

where y could be, for example, mRNA (M) or protein (P). We denote $y(t_i)$ by y_i and $y^*(\tau_{i,j})$ by $y^*_{i,j}$. Consider any two consecutive observations (y_i, y_{i+1}), the observed time series being given by $y = (y_1, y_2, \ldots, y_T)$. We want to impute a bridge of F auxiliary data points between the pair y_i and y_{i+1} at times $(\tau_{i,1}, \ldots, \tau_{i,F})$, where $\tau_{i,j+1} - \tau_{i,j} = \Delta$ (= 0.1 h for example) for all $j = 1, \ldots, F-1$. Let $y^*_i = (y^*_{i,1}, \ldots, y^*_{i,F})$ denote the auxiliary bridge and let $y^* = (y^*_1, \ldots, y^*_{T-1})$ denote all the auxiliary bridges.

We know that y^*_i is conditionally independent of the other bridges, given (y_i, y_{i+1}, θ). Thus

$$f(y^*|y_1, y, \theta) = \prod_{i=1}^{T-1} f(y^*_i|y_i, y_{i+1}, \theta),$$

and

$$\begin{aligned} f(y^*_i|y_i, y_{i+1}, \theta) &\propto \prod_{j=0}^{F} f(y^*_{i,j+1}|y^*_{i,j}, \theta), \\ &\propto \prod_{j=0}^{F} \Phi(y^*_{i,j+1} - y^*_{i,j}; \mu(y^*_{i,j}, \tau_{i,j}, \theta)\Delta, \sigma^2(Y^*_{i,j}, \tau_{i,j}, \theta)\Delta), \end{aligned}$$

where Φ is the Normal density.

Given two data points we construct a bridge of length F between them, but sampling a bridge of length F is not recommended because it is difficult to sample a high-dimensional y^*_i in one block. Instead we construct sub-bridges of length m. A sub-bridge of m auxiliary data points starts at $y^*_{i,k}$ and ends at $y^*_{i,k+m-1}$.

$$y^*_{i(k,m)} = (y^*_{i,k}, y^*_{i,k+1}, \ldots, y^*_{i,k+m-1}),\ k = 1, m-1, 2m-1, \ldots \tag{2.27}$$

The density conditioned on the two points at either end of this sub-bridge, $y^*_{i,k-1}, y^*_{i,k+m}$, is given by

$$f(y^*_{i(k,m)}|y^*_{i,k-1}, y^*_{i,k+m}, \theta) \propto \prod_{j=k-1}^{k+m} \Phi(y^*_{i,j+1} - y^*_{i,j}; \mu(y^*_{i,j}, \tau_{i,j}, \theta)\Delta, \sigma^2(y^*_{i,j}, \tau_{i,j}, \theta)\Delta(2.28)$$

We sample each of the sub-bridges of length m in sequence and accept or reject each of them using the Metropolis-Hastings algorithm [Chib and Greenberg, 1995]. Let $q(y^*_{i(k,m)}|y^*_{i,k-1}, y^*_{i,k+m}, \theta)$ denote the proposal density. Suppose that at each iteration n of our MCMC algorithm, the sub-bridge $y^*_{i(k,m)}$ is given by $y^{*(n)}_{i(k,m)}$. We propose a new sub-bridge $w \sim q(y^*_{i(k,m)}|y^*_{i,k-1}, y^*_{i,k+m}, \theta)$. The new sub-bridge is then accepted with probability:

$$\begin{aligned}&\alpha(y^{*(n)}_{i(k,m)}, w|y^*_{i,k-1}, y^*_{i,k+m}, \theta) = \\ &\min\left(1, \frac{f(w|y^*_{i,k-1}, y^*_{i,k+m}, \theta)q(y^{*(n)}_{i(k,m)}|y^*_{i,k-1}, y^*_{i,k+m}, \theta)}{f(y^{*(n)}_{i(k,m)}|y^*_{i,k-1}, y^*_{i,k+m}, \theta)q(w|y^*_{i,k-1}, y^*_{i,k+m}, \theta)}\right)\end{aligned}$$

The proposal density $q(.|.)$ is chosen to be a multivariate Normal approximation of the target density at the mode. The location of $q(.|.)$ is given by the mode of the target density obtained by a few Newton-Raphson steps and the dispersion is given by the negative of the inverse Hessian evaluated at the mode. This is a multi-dimensional independence sampler, as the proposal distribution $q(.|.)$ does not depend on the current value of the chain. [Elerian et al., 2001] give analytic functions for both the gradient and negative Hessian for the type of stochastic differential equation we are considering, removing the need to approximate these functions.

Step 3, i.e. updating the parameters, is carried out as usual with the augmented data (Y, Y^*) being treated in the same way as if we had fine data.

Chapter 3

Estimation of biochemical kinetic parameters using the linear noise approximation

3.1 Author contributions and chapter's structure

This chapter is a paper by Michał Komorowski, Bärbel Finkenstädt, Claire V. Harper and David A. Rand submitted to BMC Bioinformatics. Author contributions are as follows. MK proposed and implemented the algorithm. CVH performed the cycloheximide experiment. MK wrote the paper with assistance from BF and DAR, who supervised the study.

Sections 3.2 - 3.6 are followed by supplementary section 3.7 that contains details of mathematical modeling and statistical methods.

3.2 Abstract

Fluorescent and luminescent gene reporters allow us to dynamically quantify changes in molecular species concentration over time on the single cell level. The mathematical modeling of their interaction through multivariate dynamical models re-

quires the development of effective statistical methods to calibrate such models against available data. Given the prevalence of stochasticity and noise in biochemical systems inference for stochastic models is of special interest. In this chapter we present a simple and computationally efficient algorithm for the estimation of biochemical kinetic parameters from gene reporter data.

We use the linear noise approximation to model biochemical reactions through a stochastic dynamic model which essentially approximates a diffusion model by an ordinary differential equation model with an appropriately defined noise process. An explicit formula for the likelihood function can be derived allowing for computationally efficient parameter estimation. The proposed algorithm is embedded in a Bayesian framework and inference is performed using Markov chain Monte Carlo.

The major advantage of the method is that in contrast to the more established diffusion approximation based methods the computationally costly methods of data augmentation are not necessary. Our approach also allows for unobserved variables and measurement error. The application of the method to both simulated and experimental data shows that the proposed methodology provides a useful alternative to diffusion approximation based methods.

3.3 Background

The estimation of parameters in biokinetic models from experimental data is an important problem in Systems Biology. In general the aim is to calibrate the model so as to reproduce experimental results in the best possible way. The solution of this task plays a key role in interpreting experimental data in the context of dynamic mathematical models and hence in understanding the dynamics and control of complex intracellular chemical networks and the construction of synthetic regulatory circuits [Ehrenberg et al., 2003]. Among biochemical kinetic systems, the dynamics of gene expression and of gene regulatory networks are of particular interest. Recent developments of fluorescent microscopy allow us to quantify changes in protein concentration over time in single cells (e.g. [Elowitz et al., 2002a, Nelson et al., 2004]) even with single molecule precision (see [Xie et al., 2008] for review). Therefore an abundance of data is becoming available to estimate parameters of mathematical models in many important cellular systems.

Single cell imaging techniques have revealed the stochastic nature of biochemical reactions (see [Raser and O'Shea, 2005] for review) that most often occur far from thermodynamic equilibrium [Keizer, 1987] and may involve small copy numbers of reacting macromolecules [Guptasarma, 1995]. This inherent stochasticity implies that the dynamic behaviour of one cell is not exactly reproducible and that there exists stochastic heterogeneity between cells. The disparate biological systems, experimental designs and data types impose conditions on the statistical methods that should be used for inference [Finkenstadt et al., 2008, Golightly and Wilkinson, 2005, Moles et al., 2003]. From the modeling point of view the current common consensus is that the most exact stochastic description of the biochemical kinetic system is provided by the chemical master equation (CME) [Gillespie, 1992a]. Unfortunately, for many tasks such as inference the CME is not a convenient mathematical tool and hence various types of approximations have been developed. The three most commonly used approximations are [Van Kampen, 2006]:

1. The macroscopic rate equation (MRE) approach which describes the thermodynamic limit of the system with ordinary differential equations and does not take into account random fluctuations due to the stochasticity of the reactions.

2. The diffusion approximation (DA) which provides stochastic differential equation (SDE) models where the stochastic perturbation is introduced by a state dependant Gaussian noise.

3. The linear noise approximation (LNA) which can be seen as a combination because it incorporates the deterministic MRE as a model of the macroscopic system and the SDEs to approximatively describe the fluctuations around the deterministic state.

Statistical methods based on the MRE have been most widely studied [Esposito and Floudas, 2000, Mendes and Kell, 1998, Moles et al., 2003, Ramsay et al., 2007]. They require data based on large populations. The main advantages of this method are its conceptual simplicity and the existence of extensive theory for differential equations. However, single cells experiments and studies of noise in small regulatory networks created the need for statistical tools that are capable to extract information from fluctuations in molecular species. Two methods have been proposed to address this. The one by [Reinker et al., 2006] assumes availability of single molecule precision data. Another approach is based on the diffusion approximation [Golightly and Wilkinson, 2005, Heron et al., 2007]. This uses likelihood approximation methods (e.g. [Elerian et al., 2001]) that are computationally intensive and require sampling from high dimensional posterior distributions. Inference using these methods is particularly difficult for low frequency data with unobserved model variables [Finkenstadt et al., 2008, Heron et al., 2007]. The aim of this study is to investigate the use of the LNA as a method for inference about kinetic parameters of stochastic biochemical systems. We find that the LNA approximation provides an explicit Gaussian likelihood for models with hidden variables and measurement error and is therefore simpler to use and computationally efficient. To account for prior information on parameters our methodology is embedded in the Bayesian paradigm.

We first provide a description of the LNA based modeling approach and then formulate the relevant statistical framework. We then study its applicability in four examples, based on both simulated and experimental data, that clarify principles of the method.

3.4 Methods

The chemical master equation (CME) is the primary tool to model the stochastic behaviour of a reacting chemical system. It describes the evolution of the joint probability distribution of the number of different molecular species in a spatially homogeneous, well stirred and thermally equilibrated chemical system [Gillespie, 1992a]. Even though these assumptions are not necessarily satisfied in living organisms the CME is commonly regarded as the most realistic model of biochemical reactions inside living cells. Consider a general system of N chemical species inside a volume Ω and let $\mathbf{X} = (X_1, \ldots, X_N)^T$ denote the number and $\mathbf{x} = \mathbf{X}/\Omega$ the concentrations of molecules. The stoichiometry matrix $\mathbf{S} = \{S_{ij}\}_{i=1,2 \ldots N;\ j=1,2 \ldots R}$ describes changes in the population sizes due to R different chemical events, where each S_{ij} describes the change in the number of molecules of type i from X_i to $X_i + S_{ij}$ caused by an event of type j. The probability that an event of type j occurs in the time interval $[t, t+dt)$ equals $\tilde{f}_j(\mathbf{x}, \Omega, t)\Omega dt$. The functions $\tilde{f}_j(\mathbf{x}, \Omega, t)$ are called *mesoscopic transition rates*. This specification leads to a Poisson birth and death process where the probability $h(\mathbf{X}, t)$ that the system is in the state $\mathbf{X}$ at time t is described by the CME [Van Kampen, 2006] which is given in supplementary section 3.7. The first order terms of a Taylor expansian of the CME in powers of $1/\sqrt{\Omega}$ are given by the following MRE (see supplementary section 3.7)

$$\frac{d\phi_i}{dt} = \sum_{j=1}^{R} S_{ij} f_j(\varphi, t) \qquad i = 1, 2, \ldots, N; \tag{3.1}$$

where $\phi_i = \lim_{\Omega \to \infty, X \to \infty} X_i/\Omega,\ \varphi = (\phi_1, \ldots, \phi_N)^T$ and
$f_j(\varphi, t) = \lim_{\Omega \to \infty} \tilde{f}_j(\mathbf{x}, \Omega, t)$.
Including also the second order terms of this expansion produces the LNA

$$\mathbf{x}(t) = \varphi(t) + \Omega^{-\frac{1}{2}} \xi(t) \tag{3.2}$$

which decomposes the state of the system into a deterministic part φ as solution of the MRE in (3.1) and a stochastic process ξ described by an Itô diffusion equation

$$d\xi(t) = \mathbf{A}(t)\xi dt + \mathbf{E}(t) dW, \tag{3.3}$$

where dW denotes increments of a Wiener process, $[\mathbf{A}(t)]_{ik} = \sum_{j=1}^{R} S_{ij} \partial f_j / \partial \phi_k$, $[\mathbf{E}(t)]_{ij} = S_{ij} \sqrt{f_j(\varphi, t)}$ and $f_i = f_i(\varphi)$ (see supplementary section 3.7 for derivation).

The rationale behind the expansion in terms of $1/\sqrt{\Omega}$ is that for constant average concentrations relative fluctuations will decrease with the inverse of the square root of the volume [Elf and Ehrenberg, 2003]. Therefore the LNA is accurate when fluctuations are sufficiently small in relation to the mean (large Ω). Hence, the natural measure of adequacy of the LNA is the coefficient of variation i.e. ratio of the standard deviation to the mean (see supplementary section 3.7). Validity of this approximation is also discussed in details in [Elf and Ehrenberg, 2003, Ferm et al., 2007]. In addition it can be shown that the process describing the deviation from the deterministic state $\Omega^{\frac{1}{2}}(\mathbf{x} - \varphi)$ converges weakly to the diffusion (3.3) as $\Omega \to \infty$ [Kurtz, 1972]. In order to use the LNA in a likelihood based inference method we need to derive transition densities of the process $\mathbf{x}$.

3.4.1 Transition densities

The LNA provides solutions that are numerically or analytically tractable because the MRE in (3.1) can be solved numerically and the linear SDE in (3.3) for an initial condition $\xi(t_i) = \xi_{t_i}$ has a solution of the form [Arnold, 1974]

$$\xi(t) = \Phi_{t_i}(t - t_i)\left(\xi_{t_i} + \int_{t_i}^{t} \Phi_{t_i}(s - t_i)^{-1}\mathbf{E}(s)dW(s)\right), \tag{3.4}$$

where the integral is in the Itô sense and $\Phi_{t_i}(s)$ is the fundamental matrix of the non-autonomous system of ODEs

$$\frac{d\Phi_{t_i}}{ds} = \mathbf{A}(t_i + s)\Phi_{t_i}, \quad \Phi_{t_i}(0) = I. \tag{3.5}$$

Equations (3.4), (3.5) imply that the transition densities of the process ξ are Gaussian [1][Oksendal, 1992]

$$\mathbf{p}(\xi_{t_i}|\xi_{t_{i-1}}, \Theta) = \psi(\xi_{t_i}|\mu_{i-1}, \Xi_{i-1}) \tag{3.6}$$

where Θ denotes a vector of all model parameters, $\psi(\cdot|\mu_{i-1}, \Xi_{i-1})$ is the normal density with mean μ_{i-1} and variance Ξ_{i-1} specified by

$$\begin{aligned} \mu_{i-1} &= \Phi_{t_{i-1}}(\Delta_{i-1})\xi_{t_{i-1}}, \qquad \Delta_{i-1} = t_i - t_{i-1}, \\ \Xi_{i-1} &= \int_{t_{i-1}}^{t_i} (\Phi_s(t_i - s)E(s))(\Phi_s(t_i - s)E(s))^T ds. \end{aligned} \tag{3.7}$$

[1] Throughout the thesis we use 'Gaussian' or 'normal' shortly to denote either a univariate or a multivariate normal distribution depending on the context.

It follows from (3.2) and (3.6) that the transition densities of $\mathbf{x}$ are normal

$$\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta) = \psi(\mathbf{x}_{t_i}|\varphi(t_i) + \Omega^{-\frac{1}{2}}\mu_{i-1}, \Omega^{-1}\Xi_{i-1}). \tag{3.8}$$

The properties of the normal distribution allow us to construct a convenient inference framework that is reminiscent of the Kalman filtering methodology (see e.g. [Brockwell and Davis, 2002]).

3.4.2 Inference

It is rarely possible to observe the time evolution of all molecular components participating in the system of interest [Ronen et al., 2002]. Therefore, we partition the process $\mathbf{x}_t$ into those components $\mathbf{y}_t$ that are observed and those $\mathbf{z}_t$ that are unobserved.

Let $\bar{\mathbf{x}} \equiv (\mathbf{x}_{t_0},\ \ldots, \mathbf{x}_{t_n})$, $\bar{\mathbf{y}} \equiv (\mathbf{y}_{t_0},\ \ldots, \mathbf{y}_{t_n})$ and $\bar{\mathbf{z}} \equiv (\mathbf{z}_{t_0},\ \ldots, \mathbf{z}_{t_n})$ denote the time-series that comprise the values[2] of processes $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$, respectively, at times $t_0, \ldots t_n$. Our aim is to estimate the vector of unknown parameters Θ from a sequence of measurements $\bar{\mathbf{y}}$. Given the Markov property of the process $\mathbf{x}$ the augmented likelihood $P(\bar{\mathbf{y}}, \bar{\mathbf{z}}|\Theta)$ is given by

$$P(\bar{\mathbf{y}}, \bar{\mathbf{z}}|\Theta) = \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta)\mathbf{p}(\mathbf{x}_{t_0}|\Theta), \tag{3.9}$$

where $\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta)$ are Gaussian densities specified in (3.8). For mathematical convenience we assume that $\mathbf{p}(\mathbf{x}_{t_0}|\Theta)$ is also normal with mean $\varphi(t_0)$ and covariance matrix Ξ_{-1}. This assumption is justified as equations (3.2) and (3.3) imply normal distribution at any time given a fixed initial condition. We also assume that mean $\varphi(t_0)$ and covariance matrix Ξ_{-1} are parameterized as elements of Θ. It can then be shown (see supplementary section 3.7) that $\bar{\mathbf{x}}$ is Gaussian. Therefore

$$P(\bar{\mathbf{y}}, \bar{\mathbf{z}}|\Theta) = \psi(\bar{\mathbf{x}}|\varphi(t_0), \ldots, \varphi(t_n), \hat{\Sigma}), \tag{3.10}$$

where $\psi(\cdot|\varphi(t_0), \ldots, \varphi(t_n), \hat{\Sigma})$ is Gaussian density with mean vector $(\varphi(t_0), \ldots, \varphi(t_n))$ and covariance matrix $\hat{\Sigma}$ whose elements can be calculated numerically in a straightforward way (see supplementary section 3.7). Since the marginal distributions are

[2] Here and throughout the chapter we use the same letter to denote the stochastic process and its realization.

also Gaussian it follows that the likelihood function $P(\bar{\mathbf{y}}|\Theta)$ can be obtained from the augmented likelihood (3.10)

$$P(\bar{\mathbf{y}}|\Theta) = \psi(\bar{\mathbf{y}}|(\varphi_y(t_0), \ldots, \varphi_y(t_n)), \Sigma), \tag{3.11}$$

where the covariance matrix $\Sigma = \{\Sigma^{(i,j)}\}_{i,j=0,\ldots,n}$ is a sub-matrix of $\hat{\Sigma}$ such that $\Sigma^{(i,j)} = Cov(\mathbf{y}_{t_i}, \mathbf{y}_{t_j})$ and φ_y is the vector consisting of the observed components of φ.

Fluorescent reporter data are usually assumed to be proportional to the number of fluorescent molecules [Wu and Pollard, 2005] and measurements are subject to *measurement error*, i.e. errors that do not influence the stochastic dynamics of the system. We therefore assume that instead of the matrix $\bar{\mathbf{y}}$ our data have the form $\bar{\mathbf{u}} \equiv \lambda\bar{\mathbf{y}} + (\epsilon_{t_0}, \ldots, \epsilon_{t_n})$. The parameter λ is a proportionality constant[3] and ϵ_{t_i} denotes a random vector for additive measurement error. For mathematical convenience we assume that the joint distribution of the measurement error is normal with mean 0 and known covariance matrix Σ_ϵ, i.e. $(\epsilon_{t_0}, \ldots, \epsilon_{t_n}) \sim N(0, \Sigma_\epsilon)$. If measurement errors are independent with a constant variance σ_ϵ^2 then $\Sigma_\epsilon = \sigma_\epsilon^2 I$. Equation (3.11) implies that the likelihood function can be written as

$$P(\bar{\mathbf{u}}|\Theta) = \psi(\bar{\mathbf{u}}|\lambda(\varphi_y(t_0), \ldots, \varphi_y(t_n)), \lambda^2\Sigma + \Sigma_\epsilon). \tag{3.12}$$

Since for given data $\bar{\mathbf{u}}$ the likelihood function (3.12) can be numerically evaluated any likelihood based inference is straightforward to implement. Using Bayes' theorem, the posterior distribution $P(\Theta|\bar{\mathbf{u}})$ satisfies the relation [Gamerman and Lopes, 2006]

$$P(\Theta|\bar{\mathbf{u}}) \propto P(\bar{\mathbf{u}}|\Theta)\pi(\Theta). \tag{3.13}$$

We use the standard Metropolis-Hastings (MH) algorithm [Gamerman and Lopes, 2006] to sample from the posterior distribution in (3.13).

[3] It is straightforward to generalize for the case with different proportionality constants for different molecular components.

3.5 Results and Discussion

In order to study the use of the LNA method for inference we have selected four examples which are related to commonly used quantitative experimental techniques such as measurements based on reporter gene constructs and reporter assays based on Polymerase Chain Reaction (e.g. RT-PCR, Q-PCR). For expository reasons, all case studies consider a model of single gene expression.

3.5.1 Model of single gene expression

Although gene expression involves various biochemical reactions it is essentially modeled in terms of only three biochemical species (DNA, mRNA, protein) and four reaction channels (transcription, mRNA degradation, translation, protein degradation) [Chabot et al., 2007, Komorowski et al., 2009b, Thattai and van Oudenaarden, 2001]. Let $\mathbf{x} = (r, p)$ denote concentrations of mRNA and protein, respectively. The stoichiometry matrix has the form

$$S = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \tag{3.14}$$

where rows correspond to molecular species and columns to reaction channels. For the reaction rates

$$\tilde{\mathbf{f}}(\mathbf{x}) = (k_R(t), \gamma_R r, k_P r, \gamma_P p)^T \tag{3.15}$$

using (3.1) we can derive the following macroscopic rate equations

$$\dot{\phi}_R = k_R(t) - \gamma_R \phi_R, \qquad \dot{\phi}_P = k_P \phi_R - \gamma_P \phi_P. \tag{3.16}$$

For the general case it is assumed that the transcription rate $k_R(t)$ is time-dependent, reflecting changes in the regulatory environment of the gene such as the availability of transcription factors or chromatin structure.

Using (3.15) and (3.16) in (3.3) we obtain the following SDEs describing the deviation from the macroscopic state

$$\begin{aligned} d\xi_R &= -\gamma_R \xi_R dt + \sqrt{k_R(t) + \gamma_R \phi_R(t)} dW_R \qquad (3.17) \\ d\xi_P &= (k_P \xi_R - \gamma_P \xi_P) dt + \sqrt{k_P \phi_R(t) + \gamma_P \phi_P(t)} dW_P. \end{aligned}$$

We will refer to the model in (3.16) and (3.17) as the *simple model* of single gene expression.

In order to test our method on a nonlinear system we will also consider the case of an autoregulated network where the transcription rate of the gene is a function of its protein concentration as the protein interferes with production of its own mRNA. This is parameterized by a Hill function [Thattai and van Oudenaarden, 2001] $k_R(t,p) = k_R(t)/(1+(p/H)^{n_H})$ where $k_R(t)$ now describes the maximum rate of transcription, H is a dissociation constant and n_H is a Hill coefficient. Thus, the nonlinear autoregulatory model is described by the MRE

$$\dot{\phi}_R = k_R(t,\phi_P) - \gamma_R\phi_R, \qquad \dot{\phi}_P = k_P\phi_R - \gamma_P\phi_P \tag{3.18}$$

and the SDEs

$$\begin{aligned} d\xi_R &= (k'_R(t)\xi_P \;-\; \gamma_R\xi_R)dt \quad + \quad \sqrt{k_R(t)+\gamma_R\phi_R(t)}dW_R \\ d\xi_P &= (k_P\xi_R - \gamma_P\xi_P)dt + \sqrt{k_P\phi_R(t)+\gamma_P\phi_P(t)}dW_P \end{aligned} \tag{3.19}$$

where $k'_R(t) \equiv \partial k_R(t,\phi_P)/\partial\phi_P$. We refer to this model as *the autoregulatory model* of single gene expression. The two models constitute the basis of our inference studies below.

3.5.2 Inference from fluorescent reporter gene data for the simple model of single gene expression

To test the algorithm we first use the simple model of single gene expression. We generate data according to the stoichiometry matrix (3.14) and rates (3.15) using Gillespie's algorithm [Gillespie, 1977] and sample it at discrete time points. We then generate artificial data that are proportional to the simulated protein data with added normally distributed measurement error with known variance σ_ϵ^2. Furthermore we assume that mRNA levels are unobserved. Thus the data are of the form[4]

$$\bar{\mathbf{u}} = (u_{t_0},\ldots,u_{t_n})^T, \tag{3.20}$$

[4] The volume of the system Ω is unknown and we set $\Omega = 1$ so that concentration equals the number of molecules.

where $u_{t_i} = \lambda p_{t_i} + \epsilon_{t_i}$, p_{t_i} is the simulated protein concentration, λ is an unknown proportionality constant and ϵ_{t_i} is measurement error. For the purpose of our example we model the transcription function by

$$k_R(t) = \begin{cases} b_0 \exp(-b_1(t-b_3)^2) + b_4 & t \leq b_3 \\ b_0 \exp(-b_2(t-b_3)^2) + b_4 & t > b_3 \end{cases} \tag{3.21}$$

This form of transcription corresponds to an experiment, where transcription increases for $t \leq b_3$ as a result of being induced by an environmental stimulus and for $t > b_3$ decreases towards a baseline level b_4.
We assume that at time t_0 ($t_0 << b_3$) the system is in a stationary state. Therefore, the initial condition of the MRE is a function of unknown parameters $(\phi_R(t_0), \phi_P(t_0)) = (b_4/\gamma_R, b_4 k_P/\gamma_R\gamma_P)$.

To ensure identifiability of all model parameters we assume that informative prior distributions for both degradation rates are available. Priors for all other parameters were specified to be non-informative.

To infer the vector of unknown parameters

$$\Theta = (\gamma_R, \gamma_P, k_P, \lambda, b_0, b_1, b_2, b_3, b_4)$$

we sample from the posterior distribution

$$P(\Theta|\bar{\mathbf{u}}) \propto P(\bar{\mathbf{u}}|\Theta)\pi(\Theta)$$

using the standard MH algorithm. The distribution $P(\bar{\mathbf{u}}|\Theta)$ is given by (3.12).

The protein level of the simulated trajectory is sampled every 15 minutes and a sample size of 101 points obtained. We perform inference for two simulated data sets: estimate 1 is based on a single trajectory while estimate 2 represents a larger data set using 20 sampled trajectories (see Figure 3.1A). All prior specifications, parameters used for the simulations and inference results are presented in Table 3.1A.

Estimate 1 in Table 3.1A demonstrates that it is possible to infer all parameters from a single, short length time series with a realistically achievable time resolution. Estimate 2 shows that usage of the LNA does not seem to result in any significant bias. A bias has not been detected despite the very small number of mRNA molecules (5 to 35 - Figure 3.4A in supplementary section 3.7) and protein molecules (100 to 500 - Figure 3.1A). The coefficient of variation varied between approximately 0.15 and 0.4 for both molecular species (Figure 3.3 in supplementary section 3.7).

Inference for this model required sampling from the 9 dimensional posterior distribution (number of unknown parameters). If instead one used a diffusion approximation based method it would be necessary to sample from a posterior distribution of much higher dimension (see supplementary section 3.7). In addition, incorporation of the measurement error is straightforward here, whereas for other methods it involves a substantial computational cost [Heron et al., 2007].

3.5.3 Inference from fluorescent reporter gene data for the model of single gene expression with autoregulation

The following example considers the autoregulatory system with only a small number of reacting molecules. Using Gillespie's algorithm we generate artificial data from the single gene expression model with autoregulation. The protein time courses were then sampled every 15 minutes at 101 discrete points per trajectory (see Figure 3.1B). As before we assume that the mRNA time courses are not observed and that the protein data are of the form given in (3.20), i.e. proportional to the actual amount of protein with additive Gaussian measurement error. As in the previous case study we estimate parameters from two simulated data sets, a single trajectory and an ensemble of 20 independent trajectories. The inference results summarized in Table 3.1B show that despite the low number of mRNA (0-15 molecules, see Fig. 2 in supplementary section 3.7) and protein (10-250 molecules, see Fig. 3.1B) all parameters can be estimated well with appropriate precision.

(A)

Param.	Prior	Value	Estimate 1	Estimate 2
γ_R	Γ(0.44,10^{-2})	0.44	0.43 (0.27-0.60)	0.49 (0.38-0.61)
γ_P	Γ(0.52,10^{-2})	0.52	0.51 (0.35-0.67)	0.49 (0.38-0.61)
k_P	Exp(100)	10.00	21.09 (3.91-67.17)	11.41 (7.64-16.00)
λ	Exp(100)	1.00	1.42 (0.60-2.57)	1.08 (0.76-1.36)
b_0	Exp(100)	15.00	6.80 (0.97-18.43)	12.78 (9.80-15.33)
b_1	Exp(1)	0.40	0.79 (0.05-3.02)	0.29 (0.18-0.43)
b_2	Exp(1)	0.40	0.77 (0.08-2.79)	0.77 (0.32-1.59)
b_3	Exp(10)	7.00	6.13 (4.41-7.85)	7.25 (6.79-7.55)
b_4	Exp(100)	3.00	0.94 (0.11-2.88)	2.87 (2.11-3.52)

(B)

Param.	Prior	Value	Estimate 1	Estimate 2
γ_R	Γ(0.44,10^{-2})	0.44	0.44 (0.27-0.60)	0.42 (0.32-0.54)
γ_P	Γ(0.52,10^{-2})	0.52	0.49 (0.33-0.65)	0.49 (0.36-0.61)
k_P	Exp(100)	10.00	14.86 (3.18-47.97)	9.35 (5.87-13.15)
λ	Exp(100)	1.00	1.26 (0.48-2.30)	1.15 (0.81-1.50)
b_0	Exp(100)	15.00	5.99 (0.20-23.06)	13.47 (9.24-17.13)
b_1	Exp(1)	0.40	0.59 (0.01-2.75)	0.27 (0.14-0.53)
b_2	Exp(1)	0.40	0.92 (0.05-2.92)	0.83 (0.21-3.52)
b_3	Exp(10)	7.00	6.53(0.74-14.69)	7.27 (6.44-7.79)
b_4	Exp(100)	3.00	2.85 (0.35-7.19)	2.64 (1.82-3.32)

Table 3.1: Inference results for **(A)** the simple model and **(B)** autoregulatory model of single gene expression Parameter values used in simulation, prior distribution, posterior medians and 95% credibility intervals. Estimate 1 corresponds to inference from single time series, Estimate 2 relates to estimates obtained from 20 independent time series. Data used for inference are plotted in Figure 3.1A for case **A** and Figure 3.1B for case **B**. Variance of the measurement error was assumed to be known $\sigma_\epsilon = 9$. Rates are per hour. Parameters are $n_H = 1$, $H = 61.98$ in case **B**.

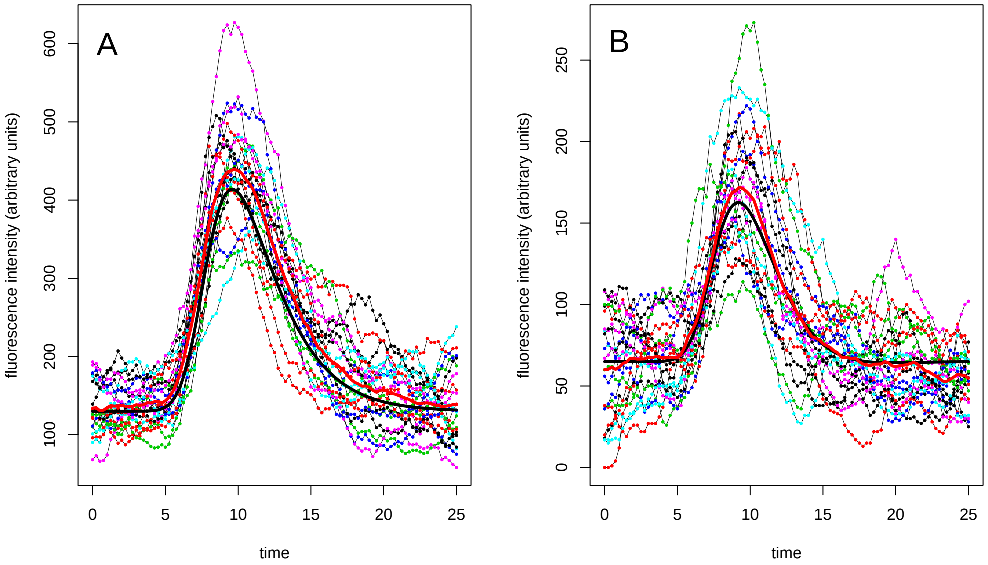

Figure 3.1: Protein timeseries generated using Gillespie's algorithm for the simple **A** and autoregulatory **B** models of single gene expression with added measurement error ($\sigma_\epsilon^2 = 9$). Initial conditions for mRNA and protein were sampled from Poisson distributions with means equal to the stationary means of the system with equal constant transcription rate b_4. In the autoregulatory case we set $H = \frac{b_4 k_P}{2\gamma_R \gamma_P}$. In each panel 20 time series are presented. The deterministic and average trajectories are plotted in bold black and red lines respectively. Corresponding mRNA trajectories (not used for inference) are presented in supplementary section 3.7.

3.5.4 Inference for PCR based reporter data

In the case of reporter assays based on Polymerase Chain Reaction (e.g. RT-PCR, Q-PCR) measurements are obtained from the extraction of the molecular contents from the inside of cells. Since the sample is sacrificed, the sequence of measurements are not strictly associated with a stochastic process describing the same evolving unit. Assume that at each time point t_i $(i = 0, ..n)$ we observe l measurements that are proportional to the number of RNA molecules either from a single cell or from a population of l cells. This gives a $(n+1) \times l$ matrix of data points

$$\bar{\mathbf{u}} \equiv \{u_{t_i,j}\}_{i=0,\ldots n;j=1,\ldots,l} \tag{3.22}$$

where $u_{t_i,j} = \lambda r_{t_i,j} + \epsilon_{t_i,j}$, $r_{t_i,j}$ is the actual RNA level, λ is the proportionality constant, $\epsilon_{t_i,j}$ is a Gaussian independent measurement error indexed by time t_i. $j = 1, \dots, l$ indexes the l measurements that are taken at time t_i. Note that $r_{t_i,j}$ and $r_{t_{i+1},j}$ are independent random variables as they refer to different cells. We assume that the dynamics of RNA is described by the simple model of single gene expression with LNA equations (3.16) and (3.17). Let Υ_t denote the distribution of measured RNA at time t ($u_t \sim \Upsilon_t$). In order to accommodate for the different form of data we modify the estimation procedure as follows. For analytical convenience we assumed that the initial distribution is normal $\Upsilon_{t_0} = N(\mu_{t_0}, \sigma^2_{t_0})$. This together with eq. (3.8) and normality of measurement error implies that $\Upsilon_t = N(\mu_t, \sigma^2_t)$. Simple explicit formulae for μ_t and σ^2_t are derived in supplementary section 3.7. Since all observations $u_{t_i,\cdot}$ are independent we can write the posterior distribution as

$$\pi(\Theta|\bar{\mathbf{u}}) \propto \prod_{i=0}^{n} \prod_{j=1}^{l} \psi(u_{t_i,j}|\mu_{t_i}, \sigma^2_{t_i})\ \pi(\Theta), \tag{3.23}$$

where $\psi(\cdot|\mu_{t_i}, \sigma^2_{t_i})$ is the normal density with parameters $\mu_{t_i}, \sigma^2_{t_i}$. In order to infer the vector of the unknown parameters $\Theta = (\gamma_R, \lambda, b_0, b_1, b_2, b_3, b_4, \mu_{t_0}, \sigma^2_{t_0})$ we sample from the posterior using a standard MH algorithm. To test the algorithm we have simulated a small ($l = 10$, $n = 50$, plotted in Figure 3.2) and a large ($l = 100$, $n = 50$) data set using Gillespie's algorithm with parameter values given in Table 3.2. The data were sampled discretely every 30 minutes and a standard normal error was added. Initial conditions were sampled from the Poisson distribution with mean b_4/γ_R. The estimation results in Table 3.2 show that parameters can be inferred well in both cases even though the number of RNA molecules in the generated data is very small (about 5-35 molecules). Since subsequent measurements do not belong to the same stochastic trajectory, estimation for the model presented here is not straightforward with the diffusion approximation based methods.

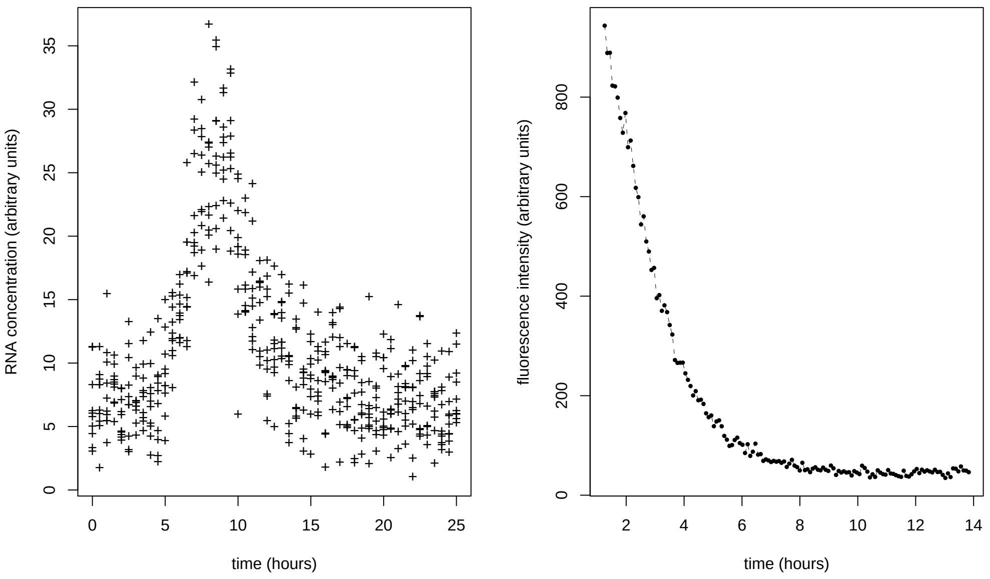

Figure 3.2: **Left:**PCR based reporter assay data simulated with Gillespie's algorithm using parameters presented in Table 3.2 and extracted 51 times (n=50), every 30 minutes with an independently and normally distributed error ($\sigma_\epsilon^2 = 9$). Each cross correspond to the end of simulated trajectory, so that the data drawn are of form (3.22). Since number of RNA molecules is know upto proportionality constant y-axis is in arbitrary units. Time on x-axis is expressed in hours. Estimates inferred form this data are shown in column *Estimate 1* in Table 3.2. **Right:** Fluorescence level from cycloheximide experiment is plotted against time (in hours). Subsequent dots correspond to measurements taken every 6 minutes.

3.5.5 Estimation of gfp protein degradation rate from cycloheximide experiment

In this section the method is applied to experimental data. After a period of transcriptional induction, translation of gfp was blocked by the addition of cycloheximide (CHX). Details of the experiment are presented in supplementary section 3.7. Fluorescence was imaged every 6 minutes for 12.5h (see Figure 3.2). Since inhibition may not be fully efficient we assume that translation may be occurring at a (possibly small) positive rate k_P. The model with the LNA is

$$\begin{aligned} \dot{\phi}_P &= k_P - \gamma_P \phi_P, \qquad (3.24) \\ d\xi_P &= -\gamma_P \xi_P dt + \sqrt{k_P + \gamma_P \phi_P} dW_P. \end{aligned}$$

Parameter	Prior	Value	Estimate 1	Estimate 2
γ_R	Exp(1)	0.44	0.45 (0.35-0.60)	0.46 (0.42-0.50)
λ	Exp(100)	1.00	1.07 (0.90-1.22)	1.01 (0.95-1.05)
b_0	Exp(100)	15.00	13.13 (10.20-15.87)	14.91 (13.86-15.77)
b_1	Exp(1)	0.40	0.29 (0.14-0.51)	0.43 (0.32-0.54)
b_2	Exp(1)	0.40	0.32 (0.12-0.93)	0.32 (0.21-0.43)
b_3	Exp(10)	7.00	7.05 (6.39-7.63)	6.99 (6.76-7.15)
b_4	Exp(100)	3.00	2.97 (2.00-4.18)	3.10 (2.76-3.43)
μ_0	Exp(100)	6.76	6.90 (5.79-7.69)	6.55 (6.14-6.85)
σ_0^2	Exp(100)	6.76	3.52 (0.01-8.99)	7.59 (5.44-9.49)

Table 3.2: Inference results for PCR based reporter assay simulated data Parameter values used to generate data, prior distributions used for estimation, posterior median estimates together with 95% credibility intervals. Estimate 1, Estimate 2 columns relate to small (l=5, n=50) and large (l=100, n=50) sample sizes. Variance of the measurement was assumed to be known $\sigma_\epsilon^2 = 4$. Estimated rates are per hour.

Param.	Prior	Estimate LNA	Estimate DA
γ_P	Exp(1)	0.53 (0.39-0.67)	0.45 (0.31-0.62)
k_P	Exp(50)	0.42 (0.15-1.04)	0.32(0.10-1.75)
λ	Exp(50)	24.07(16.57-37.05)	22.79(13.79-36.92)

Table 3.3: Inference results for CHX experimental data . Priors, posterior mean and 95% credibility intervals obtained from CHX experimental data using the LNA approach and diffusion approximation approach. Estimation with the LNA assumed $u_{t_0} = \lambda\phi_P(0)$. Estimated rates are per hour.

The observed fluorescence is assumed to be proportional to the signal with proportionality constant λ. For comparison we also consider the diffusion approximation for which an exact transition density is analytically available (see supplementary section 3.7)

$$dp = (k_P - \gamma_P p)dt + \sqrt{k_P + \gamma_P p} dW_P. \tag{3.25}$$

Since incorporation of measurement error for the diffusion approximation based model is not straightforward, we assume that measurements were taken without any error to ensure fair comparison between the two approaches. Table 3.3 shows that estimates obtained with both methods are not very different.

3.6 Conclusions

The aim of this chapter is to suggest the LNA as a useful and novel approach to the inference of biochemical kinetics parameters. Its major advantage is that an explicit formula for the likelihood can be derived even for systems with unobserved variables and data with additional measurement error. In contrast to the more established diffusion approximation based methods [Golightly and Wilkinson, 2005, Heron et al., 2007] the computationally costly methods of data augmentation to approximate transition densities and to integrate out unobserved model variables are not necessary. Furthermore, this method can also accommodate measurement error in a straightforward way. The suggested procedure here is implemented in a Bayesian framework using MCMC simulation to generate posterior distributions. The LNA has previously been studied in the context of approximating Poisson birth and death processes [Elf and Ehrenberg, 2003, Ferm et al., 2007, Kurtz, 1972, Tomioka et al., 2004] and it was shown that for a large class of models the LNA provides an excellent approximation. Furthermore, in [Tomioka et al., 2004] it is shown that for the systems with linear reaction rates the first two moments of the transition densities resulting from the CME and the LNA are equal. Here we propose using the LNA directly for inference and provide evidence that the resulting method can give very good results even if the number of reacting molecules is very small. Our experience from previous works with diffusion approximation based methods [Finkenstadt et al., 2008, Heron et al., 2007] is that their implementation is challenging especially for data that are sparsely sampled in time because the need for imputation of unobserved time points leads to a very high dimensionality of the posterior distribution. This usually results in highly autocorrelated traces affecting the speed of convergence of the Markov chain. Our method considerably reduces the dimension of the posterior distribution to the number of unknown parameters of a model only and is independent of the number of unobserved components. Nevertheless it can only be applied to the systems with sufficiently large volume, where fluctuations around a deterministic state are relatively close to the mean.

3.7 Supplementary Information

This section contains details of mathematical models and statistical methods used in the previous sections of this chapter.

3.7.1 Modelling of Chemical Kinetics

In this section we derive the macroscopic rate equation (MRE), the diffusion approximation (DA) and the linear noise approximation (LNA) for the chemical system described in the section 3.4. Our derivations here (subsection 3.7) follow [Van Kampen, 2006] and [Elf and Ehrenberg, 2003] . The chemical master equation (CME) describes the time evolution of the probability h that at time t the system is in the state $\mathbf{X}$

$$\frac{dh(\mathbf{X},t)}{dt} = \Omega \sum_{j=1}^{R} \left(\prod_{i=1}^{N} E^{-S_{ij}} - 1 \right) \tilde{f}_j(\mathbf{x},\Omega,t) h(\mathbf{X},t). \tag{3.26}$$

Here, $E^{-S_{ij}}$ is a step operator defined by

$$E^{-S_{ij}} f(...,X_i,...) = f(...,X_i - S_{ij},...).$$

Macroscopic rate equation

As the system's volume Ω increases, relative fluctuations become negligible and in the limit of infinitely large Ω the system becomes deterministic. To derive the macroscopic rate equation we write the operator $\prod_{i=1}^{N} E^{-S_{ij}}$ in the form of a first order multivariate Taylor expansion

$$\prod_{i=1}^{N} E^{-S_{ij}} = 1 - \sum_{i=1}^{N} \frac{S_{ij}}{\Omega} \frac{\partial}{\partial x_i} + O(\Omega^{-2}).$$

After substitution into the CME (3.26), in the limit of infinitely large Ω we obtain

$$\frac{dh(\varphi,t)}{dt} = - \sum_{j=1}^{R} \left(\sum_{i=1}^{N} S_{ij} \frac{\partial}{\partial \phi_i} \right) f_j(\varphi,t) h(\varphi,t). \tag{3.27}$$

This partial differential equation can be solved by the method of characteristics that reduces a partial differential equation to a family of ordinary differential equations along which the solution can be integrated [Evans, 1998].

The solution is called the *macroscopoic rate equation* and has the form

$$\frac{d\phi_i}{dt} = \sum_{j=1}^{R} S_{ij} f_j(\varphi, t) \qquad i = 1, 2, ..., N. \tag{3.28}$$

Diffusion approximation

Similarly, one may write the second order Taylor approximation of the step operator in the following way

$$\prod_{i=1}^{N} E^{-S_{ij}} = 1 - \sum_{i=1}^{N} \frac{S_{ij}}{\Omega} \frac{\partial}{\partial x_i} + \frac{1}{2} \frac{1}{\Omega^2} \sum_i \sum_k S_{ij} S_{ik} \frac{\partial^2}{\partial x_i \partial x_k} + O(\Omega^{-3}).$$

Again, if the volume is large enough the terms of order $O(\Omega^{-3})$ can be neglected and substitution of the expanded operator into (3.26) implies the Fokker-Planck equation of the form

$$\frac{dh(\mathbf{x}, t)}{dt} = -\sum_{i=1}^{N} \sum_{k=1}^{R} \frac{\partial}{\partial x_i} [\mathbf{A}]_{ik} h(\mathbf{x}, t) + \frac{1}{2} \sum_{i,k=1}^{N} \frac{\partial}{\partial x_i} \frac{\partial}{\partial x_k} \left[\mathbf{E}\mathbf{E}^T\right]_{ik} h(\mathbf{x}, t), \tag{3.29}$$

where

$$\begin{aligned} [\mathbf{A}]_{ik} &= S_{ik} \tilde{f}_k(\mathbf{x}, \Omega, t), \ \mathbf{E} = \frac{1}{\sqrt{\Omega}} S \sqrt{diag(\tilde{\mathbf{f}}(\mathbf{x}, \Omega, t))}, \\ \left[\mathbf{E}\mathbf{E}^T\right]_{ik} &= \sum_{j=1}^{R} \frac{1}{\Omega} S_{ij} S_{kj} \tilde{f}_j(\mathbf{x}, \Omega, t) \\ \tilde{\mathbf{f}}(\mathbf{x}, \Omega, t) &= (\tilde{f}_1(\mathbf{x}, \Omega, t), ..., \tilde{f}_R(\mathbf{x}, \Omega, t))^T. \end{aligned}$$

The above Fokker-Planck equation describes the time evolution of the transition densities of the Itô diffusion equation [Gardiner, 1985]

$$d\mathbf{x} = \mathbf{A}(\mathbf{x}, \mathbf{t})dt + \mathbf{E}(\mathbf{x}, \mathbf{t})dW, \tag{3.30}$$

where dW denotes increments of the Wiener process.

Linear noise approximation

In order to obtain the linear noise approximation transition rates, $\tilde{f}_j(\mathbf{x}, t)$ and the step operator $E^{\cdot}$ are Taylor expanded around the deterministic state φ in powers of $1/\sqrt{\Omega}$. To obtain such an expansion process, X_i is decomposed into the deterministic φ and stochastic $\xi = (\xi_1, ..., \xi_N)^T$ components according to the relation

$$X_i \equiv \Omega\phi_i + \Omega^{1/2}\xi_i. \tag{3.31}$$

Transition rates are expanded as follows

$$\tilde{f}_j(\mathbf{x}, t) = f_j(\varphi, t) + \frac{1}{\sqrt{\Omega}} \sum_{i=1}^{N} \frac{\partial f_i(\varphi, t)}{\partial \phi_i} \xi_i + O(\Omega^{-1}). \tag{3.32}$$

Similarly, we have an expansion of the step operator

$$\prod_{i=1}^{N} E^{-S_{ij}} = 1 - \Omega^{-1/2} \sum_{i=1}^{N} S_{ij} \frac{\partial}{\partial \xi_i} + \frac{1}{2\Omega} \sum_{i=1}^{N} \sum_{k=1}^{N} S_{ij} S_{kj} \frac{\partial^2}{\partial \xi_i \partial \xi_k} + O(\Omega^{-\frac{3}{2}}). \tag{3.33}$$

Let us denote the probability distribution of ξ at time t by $\Pi(\xi, t)$. The distribution $h(\mathbf{X}, t)$ is related to $\Pi(\xi, t)$ through the relation

$$h(\mathbf{X}, t) = h(\Omega\varphi + \Omega^{1/2}\xi, t) = \Pi(\xi, t). \tag{3.34}$$

If $\frac{\partial X}{\partial t} = 0$ then $\frac{\partial \xi_i}{\partial t} = \Omega^{-\frac{1}{2}} \frac{\partial \phi_i}{\partial t}$ and therefore differentiating $\Pi(\xi, t)$ with respect to time at constant molecule numbers gives

$$\frac{\partial h(\mathbf{X}, t)}{\partial t} = \frac{\partial \Pi(\xi, t)}{\partial t} + \sum_{i=1}^{N} \frac{\partial \xi_i}{\partial t} \frac{\Pi(\xi, t)}{\partial \xi_i} = \frac{\partial \Pi(\xi, t)}{\partial t} - \Omega^{\frac{1}{2}} \sum_{i=1}^{N} \frac{\partial \phi_i}{\partial t} \frac{\Pi(\xi, t)}{\partial \xi_i}. \tag{3.35}$$

Putting (3.32), (3.33) and (3.34) into (3.26) and identifying terms of order Ω^0 we obtain the Fokker-Planck equation describing the evolution of Π [Elf and Ehrenberg, 2003]

$$\frac{d\Pi(\xi, t)}{dt} = -\sum_{i,k=1}^{N} [\mathbf{A}]_{ik} \frac{\partial}{\partial \xi_i} \xi_k \Pi + \frac{1}{2} \sum_{i,k=1}^{N} \left[\mathbf{E}\mathbf{E}^T\right]_{ik} \frac{\partial^2 \Pi}{\partial \xi_i \partial \xi_k}, \tag{3.36}$$

where

$$\begin{aligned} f_i &= f_i(\varphi, t), \; [\mathbf{A}]_{ik} = \sum_{j=1}^{R} S_{ij} \frac{\partial f_j}{\partial \phi_k}, \\ \mathbf{E} &= S\sqrt{diag(\mathbf{f}(\varphi, t))}, \text{ and } \left[\mathbf{E}\mathbf{E}^T\right]_{ik} = \sum_{j=1}^{R} S_{ij} S_{kj} f_j. \end{aligned} \tag{3.37}$$

The related Itô diffusion equation has the form

$$d\xi(t) = \mathbf{A}(t)\xi dt + \mathbf{E}(t) dW. \tag{3.38}$$

It is a linear SDE with time inhomogeneous coefficients and its explicit solution has form 3.4.

To obtain MRE (3.28) we expanded CME (3.26) in terms of Ω^{-1}. Nevertheless MRE can also be seen as first order expansion in term of $\Omega^{-\frac{1}{2}}$. It is because neglecting the terms of order higher than $\Omega^{-\frac{1}{2}}$ in expansion (3.33) leads to the equation

$$d\xi(t) = \mathbf{A}(t)\xi dt$$

that for an initial condition $\xi = 0$ has zero solution. In such a case the system is described solely by the MRE.

3.7.2 Derivation of the likelihood function

In this section we derive the likelihood function 3.11. We use the notation introduced in the section 3.4.

Recall that in section 3.4 we partitioned the process $\mathbf{x}_t$ into observed variables $\mathbf{y}_t$ and unobserved latent variables $\mathbf{z}_t$. The Markov property of the process $\mathbf{x}_t$ implies that the augmented likelihood function $P(\bar{\mathbf{y}}, \bar{\mathbf{z}}|\Theta)$ can be written as

$$P(\bar{\mathbf{y}}, \bar{\mathbf{z}}|\Theta) = \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta)\mathbf{p}(\mathbf{x}_{t_0}|\Theta), \tag{3.39}$$

where

$$\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta) = \psi(\mathbf{x}_{t_i}|\varphi(t_i) + \Omega^{-\frac{1}{2}}\mu_{i-1}, \Omega^{-1}\Xi_{i-1})$$

and

$$\mathbf{p}(\mathbf{x}_{t_0}|\Theta) = \psi(\mathbf{x}_{t_o}|\varphi(t_o), \Omega^{-1}\Xi_{-1}).$$

From now on to simplify notation we write μ_{i-1} instead of $\Omega^{-\frac{1}{2}}\mu_{i-1}$ and Ξ_{i-1} instead of $\Omega^{-1}\Xi_{i-1}$.

In order to write an explicit form of distribution (3.39) we use equations (3.2), (3.6) and (3.7) that imply that $\mathbf{x}_{t_i}$ can be represented as

$$\mathbf{x}_{t_i} = \phi(t_i) + \sum_{j=0}^{i} \Phi_{t_j}(t_i - t_j)\zeta_{t_j}, \tag{3.40}$$

where ζ_{t_j} are independently normally distributed random variables with mean 0 and covariance matrix Ξ_{j-1}. This implies that

$$P(\bar{\mathbf{y}}, \bar{\mathbf{z}}|\Theta) = \psi(\bar{\mathbf{x}}|(\varphi(t_0), \ldots, \varphi(t_n)), \hat{\Sigma}), \tag{3.41}$$

where the covariance matrix $\hat{\Sigma} = \{\hat{\Sigma}^{(i,j)}\}_{i,j=0,\ldots,n}$, is the $(n+1)N \times (n+1)N$ block matrix that is composed of $N \times N$ submatrices $\hat{\Sigma}^{(i,j)} = Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_j})$. Covariances $Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_j})$ can be computed using the following relations $(j \geq i)$

$$\begin{aligned} Cov(\mathbf{x}_{t_0}, \mathbf{x}_{t_0}) &= \Xi_{-1}, & (3.42)\\ Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_i}) &= \Xi_{i-1} + \Phi_{t_{i-1}}(\Delta_{i-1}) Cov(\mathbf{x}_{t_{i-1}}, \mathbf{x}_{t_{i-1}}) \Phi_{t_{i-1}}(\Delta_{i-1})^T & (3.43)\\ Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_{j+1}}) &= Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) \Phi_{t_j}(\Delta_j)^T. & (3.44) \end{aligned}$$

In general the initial covariance matrix Ξ_{-1} can be treated as a model parameter. Sometimes, however, it can be expressed in term of other model parameters (see further examples).

In order to find the likelihood function $P(\bar{\mathbf{y}}|\Theta)$ from the augmented likelihood (3.41) we use the fact that marginal distributions of the normal distribution are normal. Thus, we obtain

$$P(\bar{\mathbf{y}}|\Theta) = \psi(\bar{\mathbf{y}}|(\varphi_y(t_0), \ldots, \varphi_y(t_n)), \Sigma), \tag{3.45}$$

where the covariance matrix Σ is a block matrix $\Sigma = \{\Sigma^{(i,j)}\}_{i,j=0,\ldots,n}$ and $\Sigma^{(i,j)} = Cov(\mathbf{y}_{t_i}, \mathbf{y}_{t_j})$. Therefore $\Sigma^{(i,j)}$ is the lower right square submatrix of $\hat{\Sigma}^{(i,j)}$ which corresponds to the observed part of the process.

3.7.3 Examples

The simple model of single gene expression

The simple model of single gene expression can be summarised by the following stoichiometric equations [Thattai and van Oudenaarden, 2001]

$$\begin{array}{rcl} \mathrm{R}_1 : DNA & \xrightarrow{k_R(t)} & DNA + R \\ \mathrm{R}_2 : R & \xrightarrow{\gamma_R R/\Omega} & \varnothing \\ \mathrm{R}_3 : R & \xrightarrow{k_P R/\Omega} & R + P \\ \mathrm{R}_4 : P & \xrightarrow{\gamma_P P/\Omega} & \varnothing \end{array}$$

Vectors of molecular copy numbers ($\mathbf{X}$), concentrations ($\mathbf{x}$), and macroscopic counterparts are

$$\mathbf{X} = (R, P), \quad \mathbf{x} = (r, p), \quad \varphi = (\phi_R, \phi_P).$$

The mesoscopic and macroscopic transition rate vectors and stoichiometric matrix have the form

$$\tilde{\mathbf{f}}(\mathbf{x}, t) = \begin{pmatrix} k_R(t) \\ \gamma_R r \\ k_P r \\ \gamma_P p \end{pmatrix}, \quad \mathbf{f}(\varphi, t) = \begin{pmatrix} k_R(t) \\ \gamma_R \phi_R \\ k_P \phi_R \\ \gamma_P \phi_P \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}. \tag{3.46}$$

Chemical master equation

To obtain the CME for the system we substitute (3.46) into (3.26) to obtain [Komorowski et al., 2009b]

$$\begin{aligned} &\frac{dh(R, P, t)}{dt} = \\ &\Omega k_R(t)(h(R-1, P, t) - h(R, P, t)) + k_P R(h(R, P-1, t) - h(R, P, t)) \\ &+\gamma_R(h(R+1, P, t)(R+1) - h(R, P, t)R) + \gamma_P(h(R, P+1, t)(P+1) - h(R, P, t)P). \end{aligned} \tag{3.47}$$

Macroscopic rate equations

Similarly (3.28) and (3.46) results in

$$\begin{aligned} \dot{\phi}_R &= k_R(t) - \gamma_R \phi_R, \\ \dot{\phi}_P &= k_P \phi_R - \gamma_P \phi_P. \end{aligned} \tag{3.48}$$

Diffusion approximation

In order to apply diffusion approximation to CME (3.47) we need the drift $\mathbf{A}(\mathbf{x}, t)$ and diffusion matrices $\mathbf{E}\mathbf{E}^T$. By substitution of (3.46) into (3.29) we get

$$\mathbf{A}(\mathbf{x}, t) = \begin{pmatrix} k_R(t) & -\gamma_R r \\ k_P r & -\gamma_P p \end{pmatrix}, \quad (\mathbf{E}\mathbf{E}^T)(\mathbf{x}, t) = \frac{1}{\Omega} \begin{pmatrix} k_R(t) + \gamma_R r & 0 \\ 0 & k_P \phi_R + \gamma_P p \end{pmatrix}, \tag{3.49}$$

where $\mathbf{E} = S\sqrt{diag(\tilde{\mathbf{f}}(\mathbf{x}, t))}$. The above matrices imply the Fokker-Planck equation:

$$\begin{aligned} \frac{dh(r,p,t)}{dt} &= -\frac{\partial}{\partial r}(k_R(t) - \gamma_R r)h(r,p,t) \\ &- \frac{\partial}{\partial p}(k_P r - \gamma_P p)h(r,p,t) \\ &+ \frac{1}{2\Omega}\frac{\partial^2}{\partial r}(k_R(t) + \gamma_R r)h(r,p,t) \\ &+ \frac{1}{2\Omega}\frac{\partial^2}{\partial p}(k_P r + \gamma_P p)h(r,p,t). \end{aligned} \tag{3.50}$$

This corresponds to the Itô diffusion

$$\begin{aligned} dr &= (k_R(t) - \gamma_R r)dt + \sqrt{1/\Omega}\sqrt{k_R(t) + \gamma_R r dW_r}, \\ dp &= (k_P r - \gamma_P p)dt + \sqrt{1/\Omega}\sqrt{k_P p + \gamma_P P dW_p}. \end{aligned} \tag{3.51}$$

Linear noise approximation

In the LNA the deterministic and stochastic part are separated according to (3.31) so that $r(t) = \phi_R + \Omega^{-1/2}\xi_R$, $p(t) = \phi_P + \Omega^{-1/2}\xi_P$. Given formulae (3.46) and

(3.37) we have the following drift and diffusion matrices

$$\mathbf{A} = \begin{pmatrix} -\gamma_R & 0 \\ k_P & -\gamma_P \end{pmatrix}, \qquad (\mathbf{E}\mathbf{E}^T)(t) = \begin{pmatrix} k_R(t) + \gamma_R\phi_R & 0 \\ 0 & k_P r + \gamma_P\phi_P \end{pmatrix}. \tag{3.52}$$

Hence, the Fokker-Planck equation has the form

$$\begin{aligned} \frac{dh(\xi_R, \xi_P, t)}{dt} &= -\frac{\partial}{\partial \xi_R}(-\gamma_R\xi_R)h(\xi_R, \xi_P, t) \\ &- \frac{\partial}{\partial \xi_P}(k_P\xi_R - \gamma_P\xi_P)h(\xi_R, \xi_P, t) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial \xi_R}(k_R(t) + \gamma_R\phi_R(t))h(\xi_R, \xi_P, t) \\ &+ \frac{1}{2}\frac{\partial^2}{\partial \xi_P}(k_P\xi_R + \gamma_P\phi_P(t))h(\xi_R, \xi_P, t) \end{aligned}$$

and implies the Itô diffusion

$$\begin{aligned} d\xi_R &= (-\gamma_R\xi_R)dt + \sqrt{\phi_R(t) + \gamma_R\phi_R}dW_{\xi_R}, \\ d\xi_P &= (k_P\xi_R - \gamma_P\xi_P)dt + \sqrt{k_P\phi_P + \gamma_P\phi_P}dW_{\xi_P}. \end{aligned} \tag{3.53}$$

We assume that before time t_0 the transcription rate was constant and equal $k_R(t_0)$ ($k_R(t) = k_R(t_0)$ for $t \leq t_0$) and that the system is in the stationary state at time t_0. Therefore as the initial covariance matrix Ξ_{-1} we use the covariance matrix of the stationary distribution of the process (3.53) that by the fluctuation-dissipation theorem [Van Kampen, 2006] can be found as the solution of the following equation

$$\mathbf{A}\Xi_{-1} + \Xi_{-1}\mathbf{A}^T + \mathbf{E}\mathbf{E}^T(t_0) = 0. \tag{3.54}$$

Single gene expression with autoregulation

For the model of single gene expression with autoregulation the stoichiometric equation remain unchanged. The mesoscopic and macroscopic transition rates vector are as follows

$$\tilde{\mathbf{f}}(\mathbf{x}, t) = \begin{pmatrix} k_R(t, \phi_P) \\ \gamma_R r \\ k_P r \\ \gamma_P p \end{pmatrix}, \qquad \mathbf{f}(\varphi, t) = \begin{pmatrix} k_R(t, \phi_P) \\ \gamma_R\phi_R \\ k_P\phi_R \\ \gamma_P\phi_P \end{pmatrix}. \tag{3.55}$$

where $k_R(t,p) = k_R(t)/(1+(p/H)^{n_H})$.

To derive the LNA equations for this model we use formulae (3.55) and eq. (3.37) and write the drift and diffusion matrices as

$$\mathbf{A}(t) = \begin{pmatrix} -\gamma_R & k_R'(t) \\ k_P & -\gamma_P \end{pmatrix}, \quad (\mathbf{E}\mathbf{E}^T)(t) = \begin{pmatrix} k_R(t,\phi_P) + \gamma_R\phi_R & 0 \\ 0 & k_P\phi_R + \gamma_P\phi_P \end{pmatrix}, \tag{3.56}$$

where $k_R'(t) = \partial k_R/\partial\phi_P(t,\phi_P)$. Therefore, the equations given by the LNA are as follows

$$\begin{aligned} \dot{\phi}_R &= k_R(t,\phi_P) - \gamma_R\phi_R, & (3.57)\\ \dot{\phi}_P &= k_P\phi_R - \gamma_P\phi_P, \\ d\xi_R &= (k_R'(t)\xi_P - \gamma_R\xi_R)dt + \sqrt{k_R(t) + \gamma_R\phi_R(t)}dW_R, \\ d\xi_P &= (k_P\xi_R - \gamma_P\xi_P)dt + \sqrt{k_P\phi_R(t) + \gamma_P\phi_P(t)}dW_P. & (3.58) \end{aligned}$$

Using the same argument as in the previous example we find the initial covariance matrix Ξ_{-1} as the solution of the following equation

$$\mathbf{A}(t_0)\Xi_{-1} + \Xi_{-1}\mathbf{A}(t_0)^T + \mathbf{E}(t_0)\mathbf{E}(t_0)^T = 0. \tag{3.59}$$

Derivation of likelihood for PCR based reporter data.

In this section we derive formula (3.23). The data for the PCR based reporter case has the form

$$\bar{\mathbf{u}} = \begin{pmatrix} u_{t_0,1} & u_{t_0,2} & ,\dots, & u_{t_0,l-1} & u_{t_0,l} \\ u_{t_1,1} & u_{t_1,2} & ,\dots, & u_{t_1,l-1} & u_{t_1,l} \\ \vdots & & & & \vdots \\ u_{t_{n-1},1} & u_{t_{n-1},2} & ,\dots, & u_{t_{n-1},l-1} & u_{t_{n-1},l} \\ u_{t_n,1} & u_{t_n,2} & ,\dots, & u_{t_n,l-1} & u_{t_n,l} \end{pmatrix}, \tag{3.60}$$

where $u_{t_i,j} = \lambda r_{t_i,j} + \epsilon_{t_i,j}$, $r_{t_i,j}$ is the actual RNA concentration, λ is the proportionality constant, $\epsilon_{t_i,j}$ is the normally and independently distributed measurement error with variance σ_ϵ^2. The first of the lower indices t_i denotes the time of observation and the second index j refers to the measurement. The random variables $u_{t_i,j}$ and $u_{t_{i+1},j'}$ are independent since they belong to different cells.

We set $\Omega = 1$ and assume that the RNA levels in all cells are described by independent processes $r(t) = \phi_R(t) + \xi_R(t)$, where

$$\begin{aligned}
\dot{\phi}_R &= k_R(t) - \gamma_R \phi_R, \\
d\xi_R &= (-\gamma_R \xi_R)dt + \sqrt{k_R(t) + \gamma_R \phi_R} dW_{\xi_R}.
\end{aligned}$$

We assume that $r(t_0)$ is normally distributed with mean $\tilde{\mu}_{t_0} = \phi_R(t_0)$ and variance $\tilde{\sigma}^2_{t_0}$. Using equations 3.7 and 3.8 we obtain that

$$\mathbf{p}(r(t)|\Theta) = \psi(r(t)|\phi_R(t), \tilde{\sigma}^2_t), \tag{3.61}$$

where

$$\tilde{\sigma}^2_t = \int_{t_0}^{t} (exp(-2\gamma_R(t-s))((k_R(s) + \gamma_R \phi_R(s))))ds + \tilde{\sigma}^2_{t_0} exp(-2(\gamma_R(t - t_0))). \tag{3.62}$$

Taking into account that $u_{t_i,j} = \lambda r_{t_i,j} + \epsilon_{t_i,j}$ we obtain that

$$\mathbf{p}(u_{t_i,j}|\Theta) = \psi\left(u_{t_i,j}|\mu_{t_i}, \sigma^2_{t_i}\right), \tag{3.63}$$

where

$$\mu_{t_i} = \lambda \phi_R(t_i), \qquad \sigma^2_{t_i} = \lambda^2 \tilde{\sigma}^2_{t_i} + \sigma^2_{\epsilon}. \tag{3.64}$$

Since all observations are independent the likelihood function $P(\bar{\mathbf{u}}|\theta)$ has the form

$$P(\bar{\mathbf{u}}|\theta) = \prod_{i=0}^{n} \prod_{j=1}^{l} \psi(u_{t_i,j}|\mu_{t_i}, \sigma^2_{t_i}).$$

Cycloheximide experiment

Cycloheximide is an inhibitor of protein biosynthesis in eukaryotic organisms. It is widely used to determine degradation rates of proteins. In the experiment GH3 rat pituitary cells stably transfected with 5kb human prolactin promoter destabilised EGFP reporter construct (hPRL-d2EGFP) were seeded onto 35 mm glass coverslip-based dishes (IWAKI, Japan) and cultured in 10% FCS for 24h prior to imaging. Cells were transferred to the stage of a Zeiss Axiovert 200 equipped with an XL incubator (maintained at 37C, 5% CO2, in humid conditions) and images were obtained using a Fluar x20, 0.75 numerical aperture (Zeiss), air objective.

Excitation of d2EGFP was performed using an Argon ion laser at 488nm. Emitted light was captured through a 505-550 nm bandpass filter from a 545 nm dichroic mirror. Images were captured every 6 min. 5 μM forskolin and 0.5 μM BayK 8644 was added directly to the dish for 6h followed by the addition of 10μg/ml cyclohexamide to inhibit translation. Data was captured and analysed using LSM510 software with consecutive autofocus. Analysis was performed using Kinetic Imaging software AQM6. Regions of interest were drawn around each single cell and mean intensity data was collected over 14h.

We assume that action of cyclohexomide does not fully block translation but reduces the translation rate significantly. If the amount of mRNA is assumed constant then translation events can be treated as occurring at a small constant rate k_P. Then the model of single gene expression reduces to equations describing the variation in the amount of protein. From (3.48),(3.53) these are given by

$$\begin{aligned} \dot{\phi}_P &= k_P - \gamma_P \phi_P, \\ d\xi_P &= -\gamma_P \xi_P + \sqrt{k_P + \gamma_P \phi_P}\, dW_P. \end{aligned} \tag{3.65}$$

The DA can be used to obtain an analogous model. Again, neglecting fluctuation of mRNA concentration, assuming constant translation and setting $\Omega = 1$ from (3.51) we have

$$dp = (k_P - \gamma_P p)dt + \sqrt{k_P + \gamma_P p}\, dW_P. \tag{3.66}$$

By multiplication of the above equation with the scaling factor λ we obtain an equation for data $q = \lambda p$ proportional to the number of molecules

$$dq = (\lambda k_P - \gamma_P q)dt + \sqrt{\lambda}\sqrt{\lambda k_P + \gamma_P q}\, dW_P. \tag{3.67}$$

This equation is equivalent to Cox, Ingersoll and Ross model and has known transition densities [Durham G. B, 2002] given by

$$\mathbf{p}(q_{t_{i+1}}|q_{t_i}) = \gamma_P c \exp(-u - v)(\frac{v}{u})^{\frac{w}{2}} I_w(2\sqrt{uv}), \tag{3.68}$$

where $c = 2(\lambda\gamma_P(1 - exp(-\gamma_P \Delta_{t_i})))^{-1}$, $u = c(\lambda k_P + \gamma_P p^M_{t_i}) \exp(-\gamma_P \Delta_{t_i})$, $v = c(\lambda k_P + \gamma_P p^M_{t_i})$, $w = \frac{4k_P}{\gamma_P} - 1$, $\Delta_{t_i} = t_{i+1} - t_i$ and $I_w(\cdot)$ is the modified Bessel function of the first kind of order w.

3.7.4 Validity of the LNA

In this section we provide some guidelines for decisions about whether our method can be used to obtain reliable estimates of kinetic rates or whether a more accurate method (e.g. DA) should be used.

The linear noise approximation has been obtained by a Taylor expansion of the CME and reaction rates around deterministic system trajectories in terms of $1/\sqrt{\Omega}$. The rationale behind this expansion is that for constant average concentrations relative fluctuations will decrease with the inverse of the square root of the volume. Therefore the LNA is accurate when fluctuations are sufficiently small in relation to the mean (indication of large Ω) [Elf and Ehrenberg, 2003]. Hence, a natural measure of adequacy of the LNA is the ratio of the standard deviation to the mean, i.e. the coefficient of variation (CV). To clarify this principle consider again the simple model of single gene expression given by the CME (3.47). For simplicity assume that the transcription rate $k_R(t) = k_R$ is time-independent. It can be shown [Thattai and van Oudenaarden, 2001] that CVs for mRNA and protein concentrations have the form

$$CV(r) = \frac{1}{\sqrt{\Omega}}\frac{1}{\sqrt{k_R/\gamma_R}}, \qquad CV(p) = \frac{1}{\sqrt{\Omega}}\frac{\sqrt{1 + k_P/(\gamma_R + \gamma_P)}}{\sqrt{k_R k_P/\gamma_R\gamma_P}}. \tag{3.69}$$

The CV decreases with the $\sqrt{\Omega}$. Since Ω is not identifiable with k_R it can not be estimated from the data. Nevertheless, the CV can be easily calculated during the estimation procedure, since variances and means at all times t_i are computed to evaluate the likelihood function (3.45). Figure 3.3 presents the CV for mRNA and protein for the simple model of single gene expression and the model of single gene expression with autoregulation. The CV is always smaller than approximately 0.5 and decreases during times when the number of molecules is high. Our simulations show (data not presented) that for higher values of CV estimates may start to exhibit bias. Therefore for large values of the CV the LNA is likely to be a less reliable inference method.

There are two additional arguments that justify the usage of the LNA in a more precise way. If X is a Poisson birth and death process governed by the CME

(3.26), φ is a solution of the MRE (3.28) and ξ is described by (3.38) then

1. the process $\Omega^{\frac{1}{2}}(\mathbf{X}-\Omega\varphi)$ weakly converges to the diffusion (3.38) as $\Omega \to \infty$ [Kurtz, 1972]; and

2. for the systems with linear reaction rates the mean and variance of transition densities of the process X and of the process $\Omega\varphi + \Omega^{1/2}\xi$ are equal [Tomioka et al., 2004].

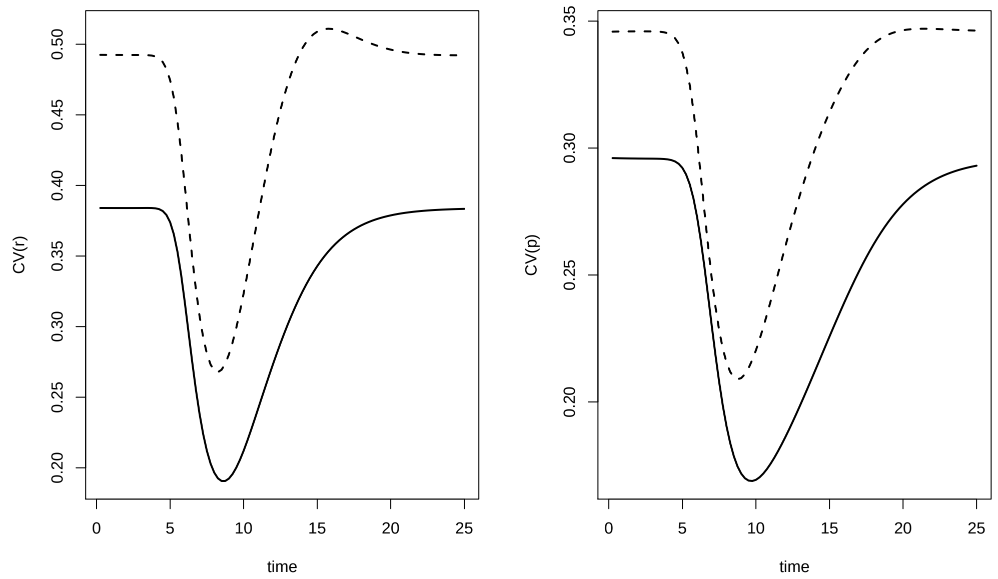

Figure 3.3: Coefficient of variation of RNA (left panel) and protein (right panel) for the models of simple gene expression (solid line) and gene expression with autoregulation (dashed line). The coefficient is calculated numerically for parameters presented in Table 3.1.

3.7.5 Notes on the practical implementation of the algorithm

Computation of the likelihood

Computation of the likelihood function 3.12 can be summarised by the following steps

```
1 Numerically find $\varphi(t)$ for $t \in [t_0, t_n]$ ;
```

2 For $i = 0, ..., n-1$ numerically find fundamental matrices $\Phi_{t_i}(s)$ for $s \in [t_i, t_{i+1}]$;

3 Use results of steps 1 and 2 to compute covariance matrices Ξ_{i-1} for $i = 0, ...n$;

4 Use matrices computed in steps 2 and 3 to construct covariance matrix $\hat{\Sigma}$ according to the procedure from section 3.7.2;

5 Extract covariance matrix Σ from $\hat{\Sigma}$ (according to section 3.7.2);

6 For given data $\bar{\mathbf{u}}$ evaluate multivariate normal density with mean vector
$\lambda(\varphi(t_0), ..., \varphi(t_{n-1}))$ and covariance matrix $\lambda^2 \Sigma + \Sigma_\epsilon$, where λ and Σ_ϵ are defined the in section 3.4;

Updating Θ

Since biochemical rates are positive it is convenient to parametrise the model in terms of logarithms of the original parameters. We denote the new parameterization by $\bar{\Theta} = (\bar{\theta}_1, ..., \bar{\theta}_k) = (log(\theta_1), ..., log(\theta_k))$. The posterior distributions $\hat{P}(\bar{\Theta}|\bar{\mathbf{u}})$ can be obtained from $P(\bar{\Theta}|\bar{\mathbf{u}})$ according to the reparameterization rule [Gamerman and Lopes, 2006] given here by the equation

$$\hat{P}(\bar{\Theta}|\bar{\mathbf{u}}) \propto P(\Theta|\bar{\mathbf{u}}) \prod_{j=1}^{k} \theta_j \tag{3.70}$$

where $\prod_{j=1}^{k} \theta_j$ is a determinant of the inverse Jacobian matrix of the parameterization $\bar{\Theta}$. Each parameter $\bar{\theta}_j$ is updated individually using a random-walk Metropolis algorithm. Let $\bar{\theta}_j^{(i)}$ be the value of $\bar{\theta}_j$ at iteration i of the MCMC algorithm. A new value $\bar{\theta}_j^{(new)}$ is proposed from the symmetric proposal distribution

$$\bar{\theta}_j^{(new)} \sim N(\bar{\theta}_j^{(i)}, \sigma^2_{\theta_j})).$$

The new value $\bar{\theta}_j^{(new)}$ is then accepted with probability given by the following

$$\min\left\{1, \frac{\hat{P}(\bar{\Theta}_j^{(new)}, \bar{\Theta}_{-j,i}|\bar{\mathbf{u}})}{\hat{P}(\bar{\Theta}_j^{i}, \bar{\Theta}_{-j,i}|\bar{\mathbf{u}}))}\right\},$$

where $\bar{\Theta}_{-j,i}$ indicates all other parameters excluding $\bar{\theta}_j$ at iteration i and $\hat{P}(\bar{\Theta}, \bar{\mathbf{z}}|\bar{\mathbf{y}})$ is the posterior distribution of parameters $\bar{\Theta}$. If $\bar{\theta}_j^{(new)}$ is not accepted then $\bar{\theta}_j^{(i+1)} = \bar{\theta}_j^{(i)}$. The variance parameter of the proposal distribution, $\sigma^2_{\theta_j}$, is carefully chosen to ensure that the proposed moves are not too small (in this case there is very high acceptance of the proposed values and the chains take a long time to explore the parameter space) or too large (in this case the chains can get 'stuck' as the proposed parameter values are not often accepted which also leads to a slow exploration of the parameter space).

Numerical approximation of fundamental matrices

Consider the linear ODE

$$\frac{d\Phi_s}{dt} = \mathbf{A}(s+t)\Phi_s, \tag{3.71}$$

where $\mathbf{A}(s+t)$ and Φ_s is an $N \times N$ matrix. Let $\Phi_s(t)$ be the solution of this with initial condition the identity matrix i.e. $\Phi_s(0) = I$. In order to compute the transition density covariances Ξ_{i-1} (eq. (3.7)), it is necessary to find these matrices. This can be done either by solving the equation directly (which gives $\Phi_s(t)$ as t varies) or by solving the adjoint equation (which gives $\Phi_s(t)$ as s varies). More detailed explanation can be found in [Zwillinger, 1989].

3.7.6 Inference using diffusion approximation

In this section we briefly describe inference methods based on the diffusion approximation. We also use the example of the simple model of gene expression to demonstrate advantages of using our method instead.

Similarly as in section 3.4. suppose we observe a discretely sampled multivariate time series $\bar{\mathbf{x}} = (x_{t_0}, ..., x_{t_n})$ that is assumed to be a realisation of the process (3.30). For simplicity we assume that all components of $\mathbf{x}$ are observed and are measured without error. The aim is to estimate the unknown parameters θ given the data $\bar{\mathbf{x}}$ through the posterior distribution $P(\theta|\bar{\mathbf{x}}) \propto P(\bar{\mathbf{x}}|\theta)\pi(\theta)$, where $\pi(\theta)$ denotes the prior distribution. In order to perform inference the likelihood $P(\bar{\mathbf{x}}|\theta)$

must be derived. Through the Markov property of the process (3.30) we have that

$$P(\bar{\mathbf{x}}|\theta) = \prod_{i=1}^{n} \mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta). \tag{3.72}$$

Exact transition densities of the diffusion (3.30) are unknown and an approximation has to be considered. If the time increment between observations $\Delta_{t_{i-1}} = t_i - t_{i-1}$ is small then a good approximation is given by the normal density [Kloeden and E., 1999]

$$\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}) = \psi(\mathbf{x}_{t_i}|\mu_{t_{i-1}}, \Xi_{t_{i-1}}), \tag{3.73}$$

The mean $\mu_{t_{i-1}}$ and covariance matrix $\Xi_{t_{i-1}}$ are given by

$$\begin{aligned} \mu_{t_{i-1}} &= \mathbf{x}_{t_{i-1}} + \mathbf{A}(\mathbf{x_{t_{i-1}}}, \mathbf{t_{i-1}})\Delta_{t_{i-1}}, && (3.74) \\ \Xi_{t_{i-1}} &= \Delta_{t_{i-1}}\ \mathbf{E}(\mathbf{x_{t_{i-1}}}, \mathbf{t})\mathbf{E}(\mathbf{x_{t_{i-1}}}, \mathbf{t})^T, && (3.75) \end{aligned}$$

where $\Delta_{t_{i-1}} = t_i - t_{i-1}$. Justification for this approximations follow from the Euler-Maruyama approximation of equation (3.30) and is discussed in details in [Kloeden and E., 1999].

In practical applications the Δ_{t_i} are usually not small. There exist various approaches in the literature to deal with such a situation (e.g. [Elerian et al., 2001],[Durham G. B, 2002],[Beskos et al., 2006]). One simple idea leading to MCMC based inference is to augment the data by introducing a finer set of times $\tau_{i,j}$ so that each interval $[t_i, t_{i+1}]$ is partitioned into $M+1$ subintervals $[t_i = \tau_{i,0}, \tau_{i,1}, ..., , \tau_{i,M+1} = t_{i+1}]$. Data is imputed at the new times $\tau_{i,j}$ which we will denote by $x^*_{\tau_{i,j}}$, $j = 1, ..., M$. Let denote $\bar{\mathbf{x}}^*$ the set of all imputed points.

The new times are chosen so that the Euler approximation can be safely assumed to be accurate on each subinterval $[\tau_{i,j}, \tau_{i,j+1}]$. We can then use equation (3.72) to obtain an augmented approximate likelihood $P(\bar{\mathbf{x}}, \bar{\mathbf{x}}^*|\theta)$ and write densities $\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}})$ in terms of imputed variables $\mathbf{x}^*_{\tau_{i-1,1}}, ..., \mathbf{x}^*_{\tau_{i-1,M}}$

$$\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}) = \prod_{j=0}^{M+1} \mathbf{p}(\mathbf{x}^*_{\tau_{(i-1)j}}|\mathbf{x}^*_{t_{(i-1)j}}) \tag{3.76}$$

and $\mathbf{p}(\mathbf{x}_{\tau_{(i-1)j}}|\mathbf{x}_{\tau_{(i-1)j}})$ are calculated according to the formula (3.73). Monte Carlo methods provide a feasible way to integrate out auxiliary variables.

By Bayes' theorem, $P(\theta, \bar{\mathbf{x}}^*|\bar{\mathbf{x}}) \propto P(\bar{\mathbf{x}}^*, \bar{\mathbf{x}}|\theta)\pi(\theta)$. Thus, to provide an estimate of θ from sparsely sampled data, MCMC can be used to sample from the joint posterior $P(\theta, \bar{\mathbf{x}}^*|\bar{\mathbf{x}})$ of the parameters θ and the auxiliary variables $\bar{\mathbf{x}}^*$ given the data $\bar{\mathbf{x}}$. The main problem of this approach is that it increases the dimension of posterior distribution by nMN (number of imputed points). The high dimension of a posterior distribution leads to highly correlated Markov Chains. Therefore long chains must be generated to provide a reliable sample from the posterior distribution. It may become practically unfeasible or extremely difficult if the data frequency is low (large M needed) or if the process $\mathbf{x}$ is high-dimensional.

If some components of the process $\mathbf{x}$ are unobserved then the same data augmentation procedure may be used to integrate out unobserved variables.

Inference for single gene expression model using the diffusion approximation

To illustrate problems related to inference using the diffusion approximation method we use the simple model of single gene expression (More detailed explanation can be found in [Finkenstadt et al., 2008]). Suppose we have a sequence of measurements

$$\bar{\mathbf{x}} = (p_{t_0}, p_{t_1}, ..., p_{t_n}),$$

that can be treated as a realisation of the p component of the process (3.51). Assume that $\Omega = 1$. To perform inference between each pair of subsequent observations $(p_{t_i}, p_{t_{i+1}})$ M additional points are introduced. In addition the r process in unobserved. Therefore the augmented data matrix (matrix composed of both $\bar{\mathbf{x}}^*$ and $\bar{\mathbf{x}}$) has the form

$$\begin{pmatrix} r^*_{t_0} & r^*_{\tau_{0,1}} & \cdots & r^*_{\tau_{0,M}} & r^*_{t_1} & \cdots & r^*_{t_{n-1}} & r^*_{\tau_{n-1,1}} & \cdots & r^*_{\tau_{n-1,M}} & r^*_{t_n} \\ p_{t_0} & p^*_{\tau_{0,1}} & \cdots & p^*_{\tau_{0,M}} & p_{t_1} & \cdots & p_{t_{n-1}} & p^*_{\tau_{n-1,1}} & \cdots & p^*_{\tau_{n-1,M}} & p_{t_n} \end{pmatrix}.$$

There are $2n(M+1)+2$ elements of the augmented data matrix (of which $2nM+n+1$ are unknown and $n+1$ are known). Therefore, the posterior distribution $P(\bar{\mathbf{x}}^*, \theta|\bar{\mathbf{x}}^*)$ has dimension $2n(M)+n+1+\dim(\Theta)$.

For comparison, if we use the approach based on the LNA the number of unknowns is equal to the dimension of Θ.

Practical adjustment of the parameter M depends mostly on the time distance between observations. For instance, if we assume that in an experiment fluorescence is measured every 17 minutes, 101 times in total (n=100) and if we set $M = 15$, postulating that RNA and protein changes in one minute intervals are normal, then we obtain that the dimension of the posterior equals 3101 plus number of elements of the vector Θ. If we use the LNA based approach, proposed in this chapter, the analogous posterior has the dimension equal to the dimension of Θ.

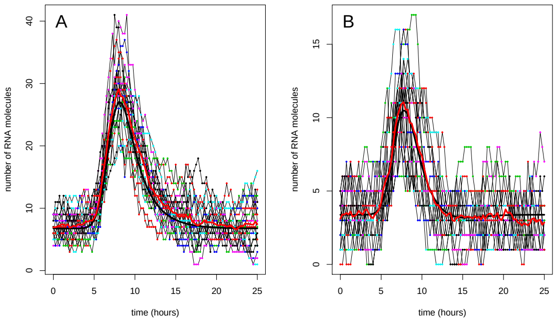

Figure 3.4: Timeseries of mRNA generated using Gillespie algorithm for models of single gene expression without autoregulation **A** and with autoregulation **B**. Parameters used for simulation and estimates inferred from the timeseries are presented in Tables 3.1A and 3.1B. In each panel 20 time series are presented. The deterministic and the average trajectory are plotted in bold black and red, respectively. Corresponding protein trajectories used for inference are presented in Figure 3.1.

Chapter 4

Using single fluorescent reporter gene to infer half-life of extrinsic noise and other parameters of gene expression

4.1 Author contributions and chapter's structure

This chapter is a paper by Michal Komorowski, Bärbel Finkenstädt, and David A. Rand submitted to Biophysical Journal. Author contributions are as follows. MK proposed and implemented the algorithm. MK wrote the paper with assistance from BF and DAR, who supervised the study.

Sections 4.2 - 4.7 are followed by supplementary section 4.8 that contains details about mathematical modeling and statistical methods.

4.2 Abstract

Fluorescent proteins are often used as reporters of transcriptional activity. Given the prevalence of noise in biochemical systems the yielded data is of significant

interests in efforts to calibrate stochastic models of gene expression and obtain information about sources of non-genetic variability.

Here we present a statistical inference framework that can be used to estimate kinetic parameters of gene expression, strength and half-life of extrinsic noise from single fluorescent reporter gene time series data. The method takes into account stochastic variability in a fluorescent signal resulting from intrinsic noise of gene expression, kinetics of fluorescent protein maturation and extrinsic noise. We use the linear noise approximation and derive an explicit formula for the likelihood of observed fluorescent data. The method is embedded in a Bayesian paradigm, so that certain parameters can be informed from other experiments allowing portability of results across different studies. Inference is performed using Markov chain Monte Carlo.

Fluorescent reporters are primary tools to observe dynamics of gene expression and correct interpretation of fluorescent data is crucial to investigate this fundamental processes of cellular live. As both magnitude and frequency of the noise may have a dramatic effect on the cell fitness, quantification of stochastic fluctuation is essential to understand how genes are regulated. Our method provides a framework that addresses this important questions.

4.3 Introduction

Since their discovery [Shimomura et al., 1962] fluorescent proteins have become one of the most commonly used markers of gene expression [Chalfie et al., 1994] in intact cells and organisms. In particular they are used to quantify changes in protein concentration over time [Nelson et al., 2004] and as reporters of transcriptional activity [Rosenfeld et al., 2005] in tissue and at the single cell level. Hence an abundance of data is becoming available, that is of interest to the estimation of kinetic parameters of expression of many different genes.

The significance of single gene expression dynamics has resulted in numerous theoretical models [Friedman et al., 2006, Kepler and Elston, 2001, Paulsson, 2006, Thattai and van Oudenaarden, 2001] and experimental studies ([Chabot et al., 2007, Elowitz et al., 2002a, Ozbudak et al., 2002, Xie et al., 2008]) that revealed aspects of the stochastic nature of this process (see [Raj and van Oudenaarden, 2008, Raser and O'Shea, 2005] for reviews). Most often occurs these systems are far from thermodynamic equilibrium [Keizer, 1987] and they may involve small copy numbers of reacting macromolecules [Guptasarma, 1995]. Determining the origins and the magnitude of the stochastic effects is of interest because of their implications for cell fate decisions, development and nongenetic individuality (see [Martinez Arias and Hayward, 2006, Raj and van Oudenaarden, 2008, Raser and O'Shea, 2005] for reviews). One of the important advances in the studies of noise in gene expression is the development of experimental methods based on using two equivalent reporters in the same cell because this allows the determination of extrinsic and intrinsic components of the total gene expression noise [Elowitz et al., 2002a, Swain et al., 2002a]. The intrinsic noise is defined as a source of variability creating differences between expression of two identical genes placed in the same cell. By contrast, extrinsic noise refers to the sources that affect the two genes equally in any given cell.

A basic assumption behind using fluorescent proteins as reporters of dynamical gene expression, particularly in the experiments investigating noise in gene expression, is that the observed fluorescence intensity is proportional to the number of proteins being expressed in the cell [Chabot et al., 2007, Elowitz et al., 2002a, Finkenstadt et al., 2008, Ozbudak et al., 2002]. There is a reasonable basis to assume that such a proportionality exists for molecules that are actively fluorescent [Wu and Pollard, 2005]. Nevertheless before the expressed protein becomes visible to fluorescent detection techniques it must undergo a maturation process that can last from few minutes to over a day [Dong and McMillen, 2008, Nagai et al., 2002] and comprises three major steps: folding, cyclization of tripeptide motif and oxidation of the cyclized motif [Tsien, 1998]. The dynamics of this process significantly contributes to the observed variability of a fluorescent signal and has the potential to impact both estimates of the number of proteins present and estimates of the variability in gene expression [Dong and McMillen, 2008, Wang et al., 2008]. Even though the maturation process has been recognised it is most often neglected in the quantitative analysis of fluorescent data (e.g. [Blake et al., 2003, Chabot et al., 2007, Elerian et al., 2001, Elowitz et al., 2002a, Finkenstadt et al., 2008, Pedraza and van Oudenaarden, 2005]).

The presence of extrinsic and intrinsic noises and stochastic effects of protein maturation indicate that extracting information from the fluorescent signal is not straightforward. Stochastic fluctuations arising at each level of gene expression are masked by subsequent steps of this process, so that the observed variability is a filtered mixture of multiple noise sources. In particular, the fluctuations in transcription rate, which is of great importance to the understanding of gene regulation, are masked by random events that occur between the release of mRNA molecules and the occurrence of fluorescent proteins. Therefore a precise interpretation of the fluorescent signal requires a mathematical model and a statistical method for its calibration. Various approaches have been proposed to address this problem [Finkenstadt et al., 2008, Friedman et al., 2006, Golightly and Wilkinson, 2005, Heron et al., 2007, Komorowski et al., 2009a, Reinker et al., 2006]. Nevertheless none of the currently available inference methods takes into account the stochasticity of the protein maturation kinetics or infers strength of extrinsic fluctuations from commonly used single reporter gene data. The currently established methods to quantify extrinsic noise require a reporter assay with two copies of a specific promoter [Chabot et al., 2007, Elowitz et al., 2002a, Swain et al., 2002a].

In this chapter we provide both a model and an efficient inference method that account for the variability originating from the fluorescent protein maturation. The method allows for the inference on the decay rate (half-life) and magnitude of extrinsic fluctuations from data of a single reporter gene experiment. Quantification of fluctuations in protein abundance is important to the understanding of how genes are regulated. For example, it has been demonstrated that both magnitude and frequency of the noise may determine cell fitness [Rosenfeld et al., 2005]. Small changes in protein concentration may have a significant effect if they last for long enough, whereas large fluctuations in concentration may not have any effect if they occur too frequently to influence cellular processes [Raser and O'Shea, 2005]. This observation stimulated studies of protein level dynamics [Rausenberger and Kollmann, 2008, Sigal et al., 2006] and reveals the need for a method to quantify the stochastic characteristics of the expression of different genes.

Our approach simultaneously solves two important problems. It infers the strength of the extrinsic variability from single fluorescent reporter gene and accounts for stochasticity of the fluorescent protein maturation. Therefore the method constitutes a general framework for the interpretation of fluorescent time-lapse data.

First we introduce the mathematical model of gene expression that incorporates stochasticity of protein maturation kinetics and extrinsic noise. We briefly analyse the influence of kinetic parameters on stochastic properties of the fluorescent signal, in order to demonstrate how filtering effects influence the identifiability of model parameters. Finally we present the statistical method to quantify observed stochasticity in fluorescent signal and demonstrate its applicability using examples of a gene that is expressed both in a steady state and out-of-steady-state. We demonstrate why all the model components are necessary to reliably interpret the fluorescent signal.

4.4 Methods

In this section we extend the standard model of single gene expression by adding the protein maturation process and a model for extrinsic noise. Subsequently we analyse stationary fluorescence fluctuations predicted by the model using the autocorrelation function and the power spectral density. Finally we use the linear noise approximation [Elf and Ehrenberg, 2003, Komorowski et al., 2009a, Van Kampen, 2006] to construct a statistical method for estimation of model parameters from fluorescent reporter gene time series.

4.4.1 Model of fluorescent gene expression

Although gene expression involves numerous biochemical reactions the current common consensus is to model it in terms of only three biochemical species (DNA, mRNA, protein) and four reaction channels (transcription, mRNA degradation, translation, protein degradation) [Friedman et al., 2006, Komorowski et al., 2009b, Thattai and van Oudenaarden, 2001]. Such a simple model has been successfully used in a variety of applications and can generate data with the same statistical behaviour as more complicated models [Dong et al., 2006, Iafolla and McMillen, 2006].

We assume what are now standard simplifications employed in this model. We assume that the process begins with production of mRNA molecules (R) at time dependent rate $k_r(t)$. Each mRNA molecule may be independently translated into protein molecules (P) at rate k_p. Both mRNA and protein molecules are degraded at rates γ_r, γ_p respectively. In order to model the expression of a fluorescent proteins we extend the standard model in a similar way to [Dong and McMillen, 2008, Wang et al., 2008]. After translation proteins are folded at a rate k_f and subsequently matured (oxidated) at a rate k_m. The number of unmatured folded proteins and matured proteins are denoted by P_f and P_m. When illuminated matured proteins are capable of emitting a fluorescent signal. Here, we neglect the cyclization reaction since it is much faster than the other two constituting the maturation process [Tsien, 1998]. We also assume that both folded and matured proteins degrade at rate γ_p. The reactions in the this model can thus be summarised as the following stoichiometric equations

$$
\begin{array}{llllll}
\mathrm{R}_1 : DNA & \xrightarrow{k_r(t)} & DNA + R & \mathrm{R}_5 : P & \xrightarrow{k_f} & P_f \\
\mathrm{R}_2 : R & \xrightarrow{\gamma_r} & \varnothing & \mathrm{R}_6 : P_f & \xrightarrow{\gamma_p} & \varnothing \\
\mathrm{R}_3 : R & \xrightarrow{k_p} & R + P & \mathrm{R}_7 : P_f & \xrightarrow{k_m} & P_m \\
\mathrm{R}_4 : P & \xrightarrow{\gamma_p} & \varnothing & \mathrm{R}_8 : P_m & \xrightarrow{\gamma_p} & \varnothing
\end{array}
$$

We model biochemical reactions as Poisson birth and death processes. Precisely, we assume that the probability for each reaction to occur in a small time interval is proportional to the product of the length of that interval, the rate of the reaction and the number of molecules which may undergo the reaction. The probability that more than one event will take place in a small time interval is of the higher order with respect to the length of the interval. Finally, we assume that events taking place in disjoint time intervals are independent, when conditioned on events in the previous interval. This specification leads to the Chemical Master Equation (see supplementary section 4.8). Unfortunately, for many tasks such as inference the CME is not a convenient mathematical tool and hence various types of approximations have been developed. As shown in [Komorowski et al., 2009a] the linear noise approximation provides a useful and reliable inference framework. The linear noise approximation models biochemical reactions through a stochastic dynamic model which essentially approximates a Poisson process by an ordinary differential equation model with an appropriately defined noise process. Using the linear noise approximation our model equations are (see supplementary section 4.8 for derivation)

$$
\begin{aligned}
dr &= (k_r(t) - \gamma_r r)dt + \sqrt{\tau(t) + \gamma_r \phi_r(t)}dW_1, && (4.1)\\
dp &= (k_p r - (\gamma_p + k_f)p)dt + \sqrt{k_p \phi_r(t) + \gamma_p \phi_p(t)}dW_2 - \sqrt{k_f \phi_p(t)}dW_3, && (4.2)\\
dp_f &= (k_f p - (\gamma_p + k_m)p_f)dt + \sqrt{k_f \phi_p(t)}dW_3 + \sqrt{\gamma_p \phi_{p_f}(t)}dW_4 - \sqrt{k_m \phi_{p_f}(t)}dW_5, && (4.3)\\
dp_m &= (k_m p_f - \gamma_p p_m)dt + \sqrt{k_m \phi_{p_f}(t)}dW_5 + \sqrt{\gamma_p \phi_{p_m}(t)}dW_6, && (4.4)
\end{aligned}
$$

where r, p, p_f, p_m are concentrations of mRNA, unfolded protein, folded protein and mature protein respectively; $\{dW_i\}_{(i=1,\dots,6)}$ denote increments of independent Wiener processes; $\tau(t)$ is a macroscopic transcription process; variables $\phi_r, \phi_p, \phi_{p_f}, \phi_{p_m}$ are macroscopic concentrations of mRNA, unfolded protein, folded protein and mature protein respectively, described by the following ordinary differential equations (see supplementary section 4.8 for derivation):

$$
\begin{aligned}
\dot{\phi}_r &= \tau(t) - \gamma_r \phi_r, && (4.5)\\
\dot{\phi}_p &= k_p \phi_p - (\gamma_p + k_f)\phi_p, && (4.6)\\
\dot{\phi}_{p_f} &= k_f \phi_p - (\gamma_p + k_m)\phi_{p_f}, && (4.7)\\
\dot{\phi}_{p_m} &= k_m \phi_{p_f} - \gamma_p \phi_{p_m}. && (4.8)
\end{aligned}
$$

The macroscopic variables describe the behaviour of the system in the thermodynamic limit. This is the limit of an infinitely large number of reacting molecules, where fluctuations average out leading to a deterministic behaviour [Van Kampen, 2006].

4.4.2 Extending the standard model by extrinsic noise

Genetically identical cells exhibit significant diversity even when exposed to the same environmental conditions. Recent studies concluded that this noise has intrinsic and extrinsic sources that could be distinguished by placing two independent gene reporters in the same cell in order to partition observed variability into these two categories [Elowitz et al., 2002a, Swain et al., 2002a]. Noise sources that create differences between the two reporters within the same cell are called intrinsic noise. Extrinsic noise, on the other hand, refers to sources that affect the two reporters equally in any given cell but create differences between two cells. Noise arising from the stochastic events of births and deaths of mRNA and proteins molecules can be identified as intrinsic. Differences between cells, either in environment or in the concentration of any factor that affects gene expression, will result in extrinsic noise (see [Raser and O'Shea, 2005] for more details).

This definition of the two sources of variability implies that in the model (4.1-4.8) intrinsic noise due to the birth and death events is modelled by diffusion terms

(terms that include dW_i).

The sources of extrinsic variability are defined less clearly. Here we focus on the stochasticity arising from fluctuations in the overall transcription rate, as it is argued in [Blake et al., 2003, Chabot et al., 2007, Shahrezaei et al., 2008], that it dominates over other sources of extrinsic noise. As proposed by [Chabot et al., 2007] and [Shahrezaei et al., 2008] extrinsic noise can arise from a multiplicative factor in the transcription rate. In this case

$$k_r(t) = \tau(t)(1 - \zeta(t)), \tag{4.9}$$

where $\tau(t)$ is a macroscopic transcription (deterministic function) and $\zeta(t)$ is a stochastic perturbation representing the extrinsic noise. In order to allow for a potential memory of the extrinsic factor, $\zeta(t)$ is modelled as an Ornstein-Uhlenbeck (OU) process

$$d\zeta = (-\gamma_\zeta \zeta)dt + \sigma_\zeta dW_7. \tag{4.10}$$

This form of transcriptional extrinsic noise has been indicated by experimental data [Chabot et al., 2007]. The OU process has an exponentially decaying autocorrelation function (ACF) of the form [Gardiner, 1985]

$$ACF_\zeta(t) = \frac{\sigma_\zeta^2}{2\gamma_\zeta} \exp(-\gamma_\zeta t). \tag{4.11}$$

The parameter γ_ζ can be thus interpreted as a decay rate of the extrinsic fluctuations and $log(2)/\gamma_\zeta$ constitutes the half-life of the extrinsic noise . Small values of γ_ζ correspond to slow transcriptional fluctuations and a slowly decaying ACF. In this case we say that transcription has long memory. The stationary variance of the OU process is given by $\frac{\sigma_\zeta^2}{2\gamma_\zeta}$ [Gardiner, 1985] and this quantity describes the strength of the extrinsic fluctuations. The model that incorporates protein maturation dynamics and extrinsic noise and for which we construct an inference method is given by equations (4.1-4.10).

4.4.3 Analysis of the fluorescent protein fluctuations

Before we present our inference method we examine how the model parameters determine the memory of fluorescence fluctuations and how they affect the filtering of the stochasticity arising from the different reactions constituting the expression process. We are particularly interested in how transcriptional memory and the strength of transcriptional fluctuations are masked by translation and protein maturation processes.

To understand how memory is determined by model parameters we analytically calculate the autocorrelation function (ACF) for the fluctuations of matured proteins p_m in the stationary state. We assure existence of the steady state by assuming that the macroscopic component of transcription is constant $\tau(t) = b$ and obtain (see supplementary section 4.8 for derivation)

$$\begin{aligned} ACF_{p_f}(t) &= a_1 \exp(-\gamma_\zeta t) + a_2 \exp(-\gamma_r t) \\ &+ a_3 \exp(-\gamma_p t) + a_4 \exp(-(\gamma_p + k_f)t) + a_5 \exp(-(\gamma_p + k_m)t), \end{aligned} \tag{4.12}$$

where $a_1, ..., a_5$ are time independent functions of model parameters. We say that the observed fluctuations have long memory (are slow) if the ACF is a slowly decreasing function of time. Formula (4.12) shows that there are five main parameters that determine how the ACF depends on time and therefore jointly determine the total memory of the observed fluctuations. These parameters are: decay rate of transcriptional fluctuations γ_ζ, mRNA and protein degradation rates γ_r, γ_p and kinetic parameters of maturation k_f, k_m. Therefore estimates of all these parameters are necessary in order to understand the origins of the observed fluorescence fluctuations.

The Fourier transform of the ACF (4.12) gives the power spectrum of the fluorescent protein fluctuations. Analysis of the spectrum (see supplementary section 4.8) reveals that the variability generated at the transcriptional level undergoes low pass filtering. Therefore fast transcriptional fluctuations (large γ_ξ) will be filtered out. The strength of the filtering depends on γ_r, γ_p, k_f, k_m. For large values of these parameters high frequencies have a smaller contribution to the observed variability.

The above analysis, similar to the more insightful studies [Austin et al., 2006, Dong and McMillen, 2008, Wang et al., 2008], is important from the point of view of inference. It shows that the filtering effect influences the identifiability of model parameters. Fast transcriptional fluctuations will not be present in the fluorescent signal and therefore the precision of estimates for γ_ζ, σ_ζ^2 will be limited. In further sections we demonstrate that our inference framework can detect this effect and account for it so that estimates of other model parameters are not affected.

4.5 Inference from fluorescent microscopy experimental data

In this section we present a method for estimating parameters of the equations (4.1-4.10) from sequences of single cell fluorescent microscopy measurements

$$\mathbf{u} = (u_{t_0}, \ldots, u_{t_n}). \tag{4.13}$$

Let $\mathbf{y}$ denote values of the process p_m evaluated at times $t_0, ..., t_n$

$$\mathbf{y} = (p_{m t_0}, \cdot, p_{m t_n}). \tag{4.14}$$

It can be shown (see supplementary section 4.8) that $\mathbf{y}$ has a Gaussian distribution

$$P(\mathbf{y}|\Theta) = \psi(\mathbf{y}|\mu(\Theta), \Sigma(\Theta)), \tag{4.15}$$

where Θ is a vector of all unknown parameters of equations (4.1-4.10), $\psi(\cdot|\mu(\Theta), \Sigma(\Theta))$ is a multivariate Gaussian density with mean vector $\mu(\Theta)$ and covariance matrix $\Sigma(\Theta)$ whose elements can be calculated numerically in a straightforward way (see supplementary section 4.8).

In order to find the distribution of the measurements $\mathbf{u}$ we define the relation between the time series of protein concentration $\mathbf{y}$ and the measurements $\mathbf{u}$, assuming that the fluorescent signal is proportional to the number of fluorescent molecules with additional measurement error

$$u_{t_i} = \lambda p_{m t_i} + \epsilon_{t_i}, \tag{4.16}$$

where λ is an unknown proportionality constant and ϵ_{t_i} is a measurement error. For mathematical convenience we assume that the joint distribution of the measurement error is normal with mean 0 and known covariance matrix Σ_ϵ, i.e. $(\epsilon_{t_0}, ..., \epsilon_{t_n}) \sim N(0, \Sigma_\epsilon)$. If measurement errors are independent with a constant variance σ_ϵ^2 then $\Sigma_\epsilon = \sigma_\epsilon^2 I$.

Equations (4.15, 4.16) and normality of the measurement error imply that the distribution of the vector $\mathbf{u}$ is also Gaussian

$$P(\mathbf{u}|\Theta) = \psi(\mathbf{u}|\lambda\mu(\Theta), \lambda^2\Sigma(\Theta) + \Sigma_\epsilon). \tag{4.17}$$

Henceforth λ is an element of vector Θ and will be estimated from experimental data. Equation (4.17) provides the joint distribution of a single time series. Often not only single but many isogenic cells are simultaneously observed under a fluorescent microscope. In this case the data matrix comprises l time series

$$\mathbf{U} = (\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(l)}). \tag{4.18}$$

As the time series corresponding to different cells are independent the likelihood function takes the form

$$P(\mathbf{U}|\Theta) = \prod_{i=1}^{l} \psi(\mathbf{u}^{(i)}|\lambda\mu(\Theta), \lambda^2\Sigma(\Theta) + \Sigma_\epsilon). \tag{4.19}$$

Since the likelihood is given explicitly, both maximum likelihood and a Bayesian approach can be used in a straightforward way. To account for prior information on parameters our methodology is embedded in the Bayesian paradigm where the posterior distribution $P(\Theta|\mathbf{U})$ satisfies [Gamerman and Lopes, 2006]

$$P(\Theta|\mathbf{U}) \propto P(\mathbf{U}|\Theta)\pi(\Theta). \tag{4.20}$$

The formulae (4.19) and (4.20) allow us to use the standard Metropolis-Hastings (MH) algorithm [Gamerman and Lopes, 2006] to generate samples from the posterior $P(\Theta|\mathbf{U})$.

4.6 Results

In this section we show that parameters of extrinsic noise can be inferred from single reporter fluorescent microscopy time series, in contrast to currently available methods that require double reporter gene experimental data [Chabot et al., 2007, Rosenfeld et al., 2005]. In addition we estimate the kinetic parameters of gene expression such as the transcription profile and the translation rate. Also the scaling factor λ which relates the fluorescent signal to the number of matured fluorescent proteins can be inferred from data.

The estimation of the model parameters is possible under the assumption that informative prior distributions for degradation rates γ_r, γ_p are obtained in additional experiments. These experiments are often not difficult to conduct [Gordon et al., 2007]. Similarly, we use informative priors for the parameters of the protein maturation process. These values are not gene or promoter dependent but characterise the fluorescent reporter. They can either be found in literature [Tsien, 1998] or estimated in experiments similar to those used to obtain degradation rates [Gordon et al., 2007] .

As the transcription and translation rates and the parameters of extrinsic noise (decay rate and variance) provide the insightful explanation of the observed fluorescent variability, our method can be seen as a quantification of different types of stochastic behaviours. To demonstrate its applicability we consider two examples. The first is an inference from steady state fluctuations, while the second is based on oscillatory, out-of-steady-state expression.

4.6.1 Stationary fluctuations

In this section, we consider a gene that is expressed in a steady state by assuming that the deterministic component of the transcription rate is constant (i.e. $\tau(t) = b$). Using a modified version of Gillespie's algorithm [Shahrezaei et al., 2008], that allows for fluctuation in reaction rates (see supplementary section 4.8 for details) we generated 50 time series for parameter values that give rise to four different types of stochastic fluctuations. The parameters values are presented in

Table 4.1 and the corresponding fluorescence signal is plotted in Figure 4.1.

Type A represents fast transcriptional fluctuations (half-life 8 minutes), that due

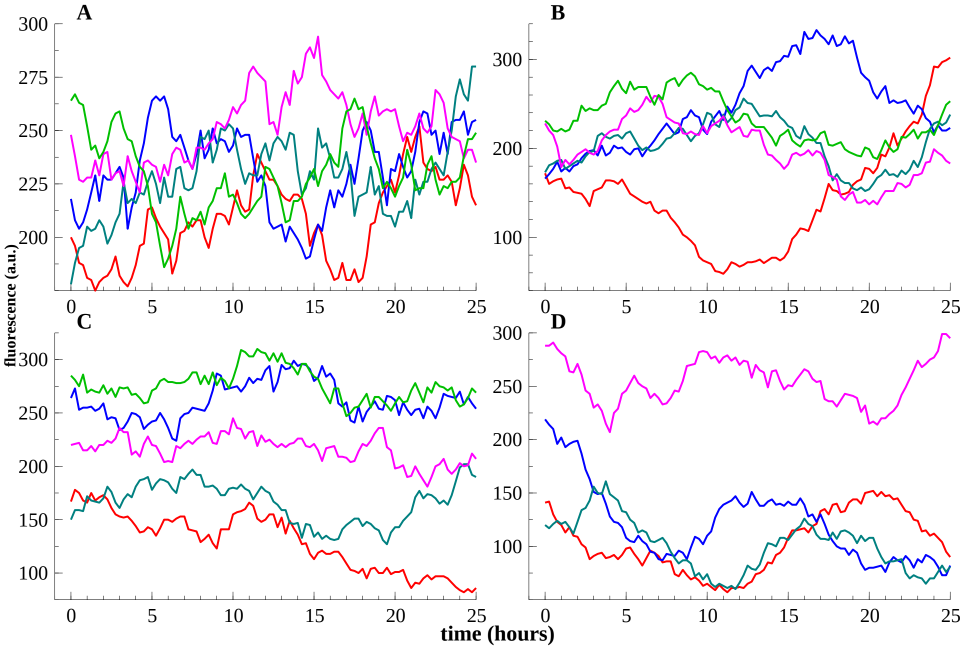

Figure 4.1: Detection of extrinsic noise in steady state data. Prior distributions (red line) and posterior distributions (black line) of parameters γ_ζ (top row), σ_ζ^2 (bottom row). Posterior distributions correspond to estimates given in Table 4.2. For fast extrinsic fluctuations (A,D) prior and posterior distribution are similar demonstrating that extrinsic fluctuations have been filtered out. In contrast, posterior distributions for slow extrinsic fluctuation (B,C) are significantly different from prior distributions and represent information about extrinsic fluctuations contained in the data.

.

to the low pass filtering effect have relatively small impact on the observed signal. In addition, the mRNA and protein degradation rates γ_r, γ_p are relatively large so that the observed variability demonstrates rather homogeneous, short memory behaviour.

Types B and C demonstrate the effect of long (half-life 83 minutes) and very long (half-life 69 hours) transcriptional memory. The degradation rates of mRNA and protein γ_r, γ_p are large (similarly to type A) so that the observed long-term memory behaviour at the fluorescent protein level is a result of the slow transcriptional fluctuations.

As the ACF in (4.12) indicates, slow fluorescence fluctuations may appear which are not necessarily due to long memory in transcription but are, for instance, due to a low mRNA degradation rate. This regime of behaviour is reflected in type D where long-term memory of fluorescence appears despite short-term memory of the transcriptional fluctuations (half-life 8 minutes).

The prior distributions and results of the inference are presented in Table 4.2 and in Figure 4.2. All kinetic parameters of gene expression, particularly the transcription and translation $(b,\ k_p)$ rates as well as the proportionality constant λ can be estimated with reasonable precision. For the cases with slow extrinsic fluctuations (B and C) the parameters of extrinsic noise γ_ζ, σ^2_ζ have been estimated from data. In cases A and D where extrinsic fluctuations are fast the obtained posterior distribution are not much different from the uninformative priors (Figure 4.2). This is due to a lack of information about these parameters in the data, that results from low pass filtering predicted by the analysis of the power spectral density (see supplementary section 4.8). Although we cannot precisely estimate the values of γ_ζ, σ^2_ζ we can detect the filtering effect that is demonstrated by the similarity of the prior and posterior distributions. This is presented in Figure 4.2, where prior and posterior distributions for these parameters are plotted. We used uninformative exponential priors (see Table 4.1). In contrast to cases A and D the posteriors and priors are significantly different for cases B and C as the slow extrinsic fluctuations are presented in the data.

This example demonstrates that our method can detect the influence of extrinsic fluctuations on the observed variability; if enough information is present in the data, the half-life and variance of the extrinsic fluctuations can be accurately estimated. Although this is an advantage in comparison to existing methods of inferring extrinsic noise our approach is valid only under the assumption that extrinsic noise arises from fluctuations in transcription rate. However, the origin of extrinsic fluctuations has never been confirmed experimentaly, therefore it would be very valuable to use biological data and compare inference results provided by our approach with that of other methods Elowitz et al. [2002a], Swain et al. [2002a].

The separation of slow and fast fluctuations can be achieved by fitting a two component autocorrelation function as shown in [Rosenfeld et al., 2005]. Nevertheless, such an ad hoc procedure will not provide information about the kinetic parameters of gene expression and cannot distinguish between the sources of fast and slow fluctuations. Moreover equation (4.12) shows that fluorescent fluctuations can contain more than two time scales. Therefore, our method provides a more insightful quantification method. Nevertheless its application requires prior knowledge about degradation and maturation rates.

Parameter	A	B	C	D	Prior
γ_r	0.44	0.44	0.44	0.1	$\Gamma(0.44, 0.01)$
γ_p	0.52	0.52	0.52	0.52	$\Gamma(0.52, 0.01)$
b	100	200	200	0.5	$Exp(100)$
k_p	1	0.5	0.5	30	$Exp(100)$
γ_ζ	5	0.5	0.01	5	$Exp(10)$
σ_ζ	1	0.1	0.002	1.25	$Exp(10)$
λ	1	1	1	1	$Exp(10)$
k_m	4.16	4.16	4.16	4.16	$\Gamma(4.16, 0.01)$
k_f	0.74	0.74	0.74	0.74	$\Gamma(0.74, 0.01)$

Table 4.1: Parameter values that correspond to the four different noise characteristics. All rates are per hour. These values give raise to the four different types of stochastic behaviour (Figure 4.1) and has been used to generate data to obtain estimates presented in Table 4.2. Last column contains prior distributions used for estimation.

4.6.2 An oscillatory gene

Most often experimental data exhibit non-equilibrium behaviour [Chabot et al., 2007, Sigal et al., 2006]. Theoretical models of gene expression have focused on analysis of steady state distributions [Friedman et al., 2006, Paulsson, 2006, Thattai and van Oudenaarden, 2001] with relatively little work done to analyse

Par.	Estimate A	Estimate B	Estimate C	Estimate D
γ_r	0.457(0.326-0.587)	0.32(0.216-0.451)	0.401(0.248-0.563)	0.091(0.063-0.119)
γ_p	0.512(0.366-0.657)	0.439(0.277-0.618)	0.546(0.397-0.683)	0.492(0.35-0.613)
b	45.175(25.89-129.98)	45.484(4.22-199.33)	132.091(50.06-278.01)	0.399(0.21-0.65)
k_p	2.089(0.443-4.597)	1.362(0.162-6.743)	0.743(0.168-1.833)	27.046(11.355-48.284)
γ_ζ	15.677(4.207-34.209)	0.965(0.322-13.471)	0.01(0.003-0.016)	7.307(0.921-25.855)
σ_ζ^2	3.658(0.067-15.792)	0.355(0.046-13.4)	0.002(0.001-0.003)	6.275(0.192-23.809)
λ	1(0.722-1.282)	1.183(0.773-1.648)	1.006(0.753-1.263)	1.091(0.801-1.407)
k_f	0.741(0.563-0.892)	0.685(0.507-0.833)	0.731(0.553-0.883)	0.721(0.551-0.869)
k_m	4.161(3.97-4.315)	4.16(3.964-4.306)	4.162(3.962-4.315)	4.161(3.972-4.311)

Table 4.2: Posterior medians and 95% credibility intervals. Each of the estimates A,B,C and D corresponds to inference from 50 independent time series generated using Gillespie's algorithm with parameters given in Table 4.1. Data were extracted every 15 minutes and 101 point per trajectory were collected. Independent and normally distributed error with variance $\sigma_\epsilon^2 = 1$ was added to each data point. For estimation variance of the measurement error was assumed to be known. Rates are per hour. The estimates are based on the final 20,000 iterations of a run of 30,000 MCMC iterations. To ensure identifiability of all model parameters we assumed that for both degradation rates and protein maturation parameters k_f, k_m informative prior distributions are available (see Table 4.1). Priors for all other parameters were specified to be non-informative.

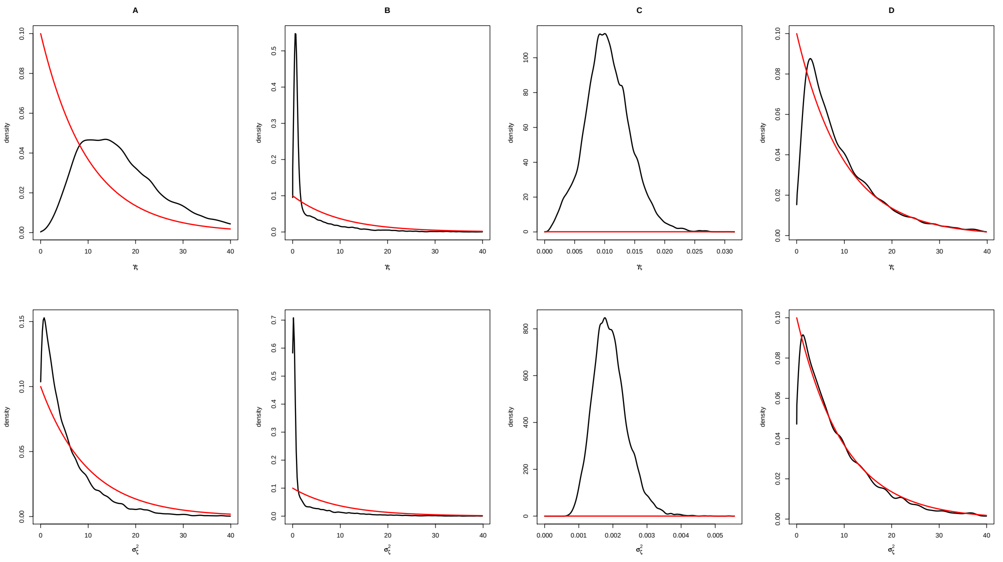

Figure 4.2: Detection of extrinsic noise in steady state data. Prior distributions (red line) and posterior distributions (black line) of parameters γ_ζ (top row), σ_ζ^2 (bottom row). Posterior distributions correspond to estimates given in Table 4.2. For fast extrinsic fluctuations (A,D) prior and posterior distribution are similar demonstrating that extrinsic fluctuations have been filtered out. In contrast, posterior distributions for slow extrinsic fluctuation (B,C) are significantly different from prior distributions and represent information about extrinsic fluctuations contained in the data.

nonequilibrium protein fluorescent trajectories [Chabot et al., 2007, Rausenberger and Kollmann, 2008]. In this section we demonstrate that our method can be applied to a system that never reaches a steady state. Although we draw similar conclusions to those in the previous section, this study demonstrates that the method can be applied to a variety of biologically relevant experiments [Chabot et al., 2007, Sigal et al., 2006]. We use oscillatory dynamics (similarly as in [Chabot et al., 2007]) as an example of non-equilibrium expression. In this case the deterministic component $\tau(t)$ of the transcription process $k_r(t)$ is modelled as

$$\tau(t) = b_0 \sin(\frac{2\pi}{24}(b_1 t + b_2)) + b_3. \tag{4.21}$$

Both slow (half-life 3.5h) and fast (half-life 21 minutes) regimes of transcriptional fluctuations are considered (see Table 4.3 for all parameter values). Figure 4.3 presents data generated using Gillespie's algorithm (see supplementary section 4.8).

As presented in Table 4.3 and Figure 4.4 the parameters of transcription and translation processes are estimated with accurate precision. For the case of slow extrinsic fluctuations $\gamma_\zeta, \sigma_\zeta^2$ are inferred precisely. In the case of fast extrinsic fluctuations the inferred posterior distributions of the parameters $\gamma_\zeta, \sigma_\zeta^2$ are not much different from uninformative priors, which demonstrates the detection of the filtering effect.

4.6.3 Necessity of all model components

In this section we demonstrate that all the components of the model (4.1-4.10,4.21) are necessary to ensure reliable interpretation of the fluorescent signal. To do so we consider two submodels of model (4.1-4.10,4.21). The first submodel assumes immediate maturation i.e. we assume that we observe $u_{t_i} = \lambda p_{t_i} + \epsilon_{t_i}$ and $k_f = k_m = 0$. The second submodel assumes immediate maturation and lack of extrinsic noise i.e. $\gamma_\zeta = \sigma_\zeta^2 = 0$. We have generated 400 independent trajectories from the full model using Gillespie's algorithm (see supplementary section 4.8) assuming that the deterministic part of transcription is oscillatory as given by

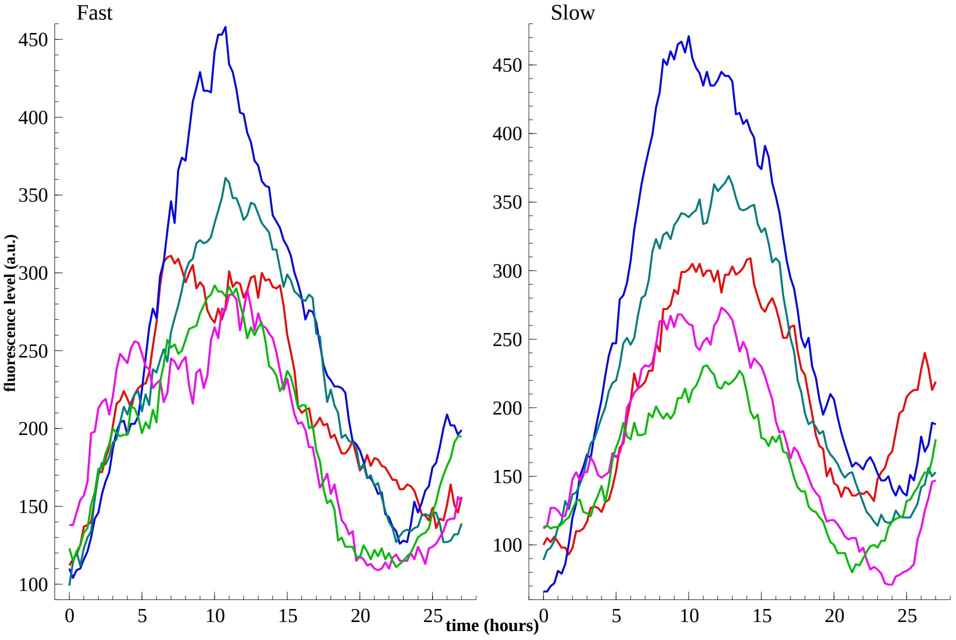

Figure 4.3: Different noise characteristics exhibited in the fluctuations of the fluorescence level for out-of-steady-state expression. **Top** Fast extrinsic fluctuations, **Bottom** Slow extrinsic fluctuations. Data generated using Gillespie's algorithm using parameters presented in Table 4.3.

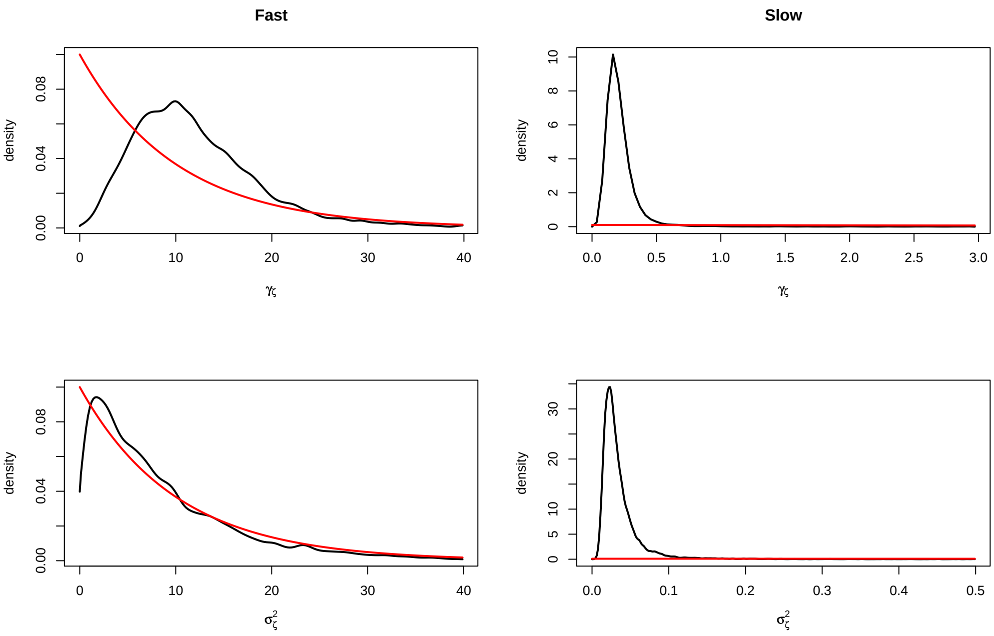

Figure 4.4: Detection of extrinsic noise in out-of-steady-state data. Prior distributions (red line) and posterior distributions (black line) of parameters γ_ζ (top row), σ^2_ζ (bottom row). Distributions correspond to the estimates for an oscillatory gene given in Table 4.3. Fast extrinsic fluctuations are not exhibited in the data therefore prior and posterior distributions are similar. In case of slow extrinsic fluctuations posterior distribution is significantly distinct from prior and contains information about extrinsic noise present in the data.

Parameter	Prior	Value F	Estimate F	Value S	Estimate S
γ_r	$\Gamma(0.44, 0.01)$	0.44	0.428(0.308-0.559)	0.44	0.392(0.251-0.524)
γ_p	$\Gamma(0.52, 0.01)$	0.52	0.502(0.357-0.635)	0.55	0.502(0.326-0.656)
b_0	Exp(100)	30	20.16(9.012-57.691)	30	44.287(9.735-119.943)
b_1	Exp(1)	1	0.967(0.927-0.995)	1	1.018(0.983-1.044)
b_2	N(0,9)	0	0.532(-0.258-1.218)	0	-0.105(-0.956-0.727)
b_3	Exp(100)	50	35.405(15.283-100.669)	50	70.447(12.701-194.011)
k_p	Exp(100)	2	2.276(0.336-6.403)	2	1.315(0.223-4.193)
γ_ζ	Exp(10)	2	11.001(2.712-23.422)	0.2	0.235(0.128-0.492)
σ_ζ^2	Exp(10)	0.4	6.124(0.364-22.249)	0.02	0.029(0.012-0.094)
λ	Exp(10)	0.1	0.117(0.063-0.172)	0.1	0.097(0.053-0.142)
k_f	$\Gamma(0.74, 0.01)$	0.74	0.724(0.546-0.879)	0.81	0.722(0.555-0.875)
k_m	$\Gamma(4.16, 0.01)$	4.16	4.165(3.975-4.308)	4.16	4.162(3.972-4.307)

Table 4.3: Parameter values used for the simulation of the oscillatory gene data, prior distribution, posterior medians and 95% credibility intervals. Value F and Estimate F corresponds to fast extrinsic fluctuations, whereas Value S and Estimate S to slow extrinsic fluctuations. Each of the estimates has been obtained from a data set of 50 independent time series sampled every 15 minutes for 25 hours (101 data points per trajectory). Independent and normally distributed error with variance $\sigma_\epsilon^2 = 9$ was added to each data point. For estimation the variance of the measurement error was assumed to be known. Rates are per hour. The estimates are based on the final 20,000 iterations of a run of 30,000 MCMC iterations. To ensure identifiability of all model parameters we assume that for both degradation rates, k_f, k_m informative prior distributions are available . Priors for all other parameters were specified to be non-informative.

equation (4.21). We simulated a large data set in this example to minimise uncertainty about the model parameters arising from any shortage of data. Then we used the full model (4.1-4.10,4.21) and both submodels to perform inference from the generated data. The results are presented in Table 4.4. As already demonstrated estimation using model (4.1-4.10,4.21) provides accurate values. Since a large data set has been used this demonstrates that application of the linear noise approximation does not result in any significant estimation bias. Inference using submodel 1 results in substantial bias in the estimates of the translation rate k_p and of the phase shift parameter b_2. This demonstrates that the incorporation of the protein maturation process is necessary to obtain the underlying transcription profile.

Estimates of all model parameters were subject to substantial bias if submodel 2 was used. As intuitively expected, this bias decreases as both protein maturation process and extrinsic fluctuations become fast enough (data not shown). Nevertheless, fast maturation and fast extrinsic fluctuations are not common [Nagai et al., 2002, Rosenfeld et al., 2005, Shahrezaei et al., 2008, Tsien, 1998] and therefore our method provides a much needed and convenient tool to interpret a fluorescent signal in the presence of slow extrinsic noise and slow maturation.

Param.	Value	Prior	Estimate 1	Estimate 2	Estimate 3
γ_r	0.440	$\Gamma(0.44, 0.01)$	0.431(0.326-0.533)	0.429(0.383-0.468)	0.302(0.241-0.375)
γ_p	0.550	$\Gamma(0.52, 0.01)$	0.575(0.436-0.707)	0.516(0.456-0.565)	0.261(0.217-0.305)
b_0	30.000	$Exp(100)$	36.160(20.150-65.999)	33.966(21.362-57.365)	6.222(5.575-6.766)
b_1	1.000	$Exp(100)$	0.997(0.986-1.005)	0.994(0.983-1.001)	0.995(0.979-1.006)
b_2	0	$N(0, 9)$	-0.156(-0.767-0.288)	-0.964(-1.229 - -0.713)	(0.867(0.499-1.142)
b_3	50.000	$Exp(100)$	61.889(32.173-116.100)	57.694(36.344-99.245)	7.425(6.520-8.169)
k_p	2.000	$Exp(100)$	1.914(0.716-3.720)	0.928(0.431-1.501)	1.494(1.082-1.937)
γ_ζ	0.200	$Exp(10)$	0.190(0.146-0.235)	0.177(0.137-0.213)	-
σ_ζ	0.012	$Exp(10)$	0.012(0.008-0.016)	0.011(0.007-0.014)	-
λ	0.100	$Exp(10)$	0.096(0.071-0.118)	0.094(0.073-0.109)	0.164(0.126-0.195)
k_f	0.74	$\Gamma(0.74, 0.01)$	0.792(0.617-0.946)	-	-
k_m	4.160	$\Gamma(4.16, 0.01)$	4.203(4.012-4.347)	-	-

Table 4.4: Parameter values used in simulation, prior distribution, posterior medians and 95% credibility intervals. Estimate 1 corresponds to the inference using model (4.1-4.10). To obtain Estimate 2 we used a model that assumes immediate protein maturation. The model used to obtain Estimate 3 assumes immediate maturation and lack of extrinsic noise. The same data set composed of 400 independent trajectories generated using Gillespie's algorithm was used for inference. Time series were sampled every 15 minutes for 25 hours (101 data points per trajectory). Rates are per hour. The estimates are based on the final 20,000 iterations of a run of 30,000 MCMC iterations. Variance of the measurement error was assumed to be known $\sigma_\epsilon^2 = 9$. To ensure identifiability of all model parameters we assumed that informative prior distributions for both degradation rates, k_f and k_m are available. Priors for all other parameters were specified to be non-informative.

4.7 Discussion

The aim of this chapter is to suggest a reliable framework for the interpretation of fluorescent reporter gene, single-cell, steady state and out-of-steady-state data. We have developed a model that shows how the observed variability depends on the kinetic parameters of a fluorescent reporter expression. The model is combined with a statistical inference framework that allows us to explain the behaviour observed in an experiment in terms of the underlying parameter values. Apart from stochasticity resulting from randomness of transcription, translation and degradation events our approach accounts for variability arising from the kinetics of fluorescent protein maturation as well as extrinsic noise represented as fluctuations in transcription rate. The proposed method allow us to infer properties of extrinsic noise such as strength and half-life from single reporter gene time-lapse data, whereas other established methods require double reporter gene experiments.

To perform parameter inference we used the linear noise approximation to derive an explicit formula for the likelihood of fluorescent reporter gene data measured with error. The suggested procedure here is implemented in a Bayesian framework using MCMC simulation to generate posterior distributions. We assure identifiability of model parameters by assuming that informative priors for mRNA and protein degradation rates as well as maturation parameters of fluorescent reporter are available and also that the variance of measurement error is known. Therefore the disadvantage of this approach is that it requires additional prior experiments to determine these parameters, nevertheless they can be measured in a relatively straightforward way described in [Gordon et al., 2007]. For some fluorescent proteins such as GFP maturation rates can be found in the literature [Tsien, 1998]. We have successfully tested our approach using data simulated with Gillespie's algorithm and demonstrated that protein maturation and extrinsic noise must be taken into account in order to reliably interpret the fluorescent signal.

We also investigated how the maturation process and transcriptional extrinsic noise influence the dynamic properties of the fluorescence fluctuations as characterized by the ACF and the power spectral density. These investigations revealed that both processes significantly affect the rate at which the ACF decays. Furthermore, they showed that the maturation process works as a low pass filter that filters out fast fluctuations in the transcription rate.

In the field of quantitative gene expression promoter-fluorescent-protein fusions are commonly used as reporters of transcriptional activity. This technique is used to address many important questions, particularly to investigate the ability of living cells to grow, divide, sense and respond to its environment in the presence of spontaneous fluctuations in their biochemical machinery. Experiments focused on establishing the origins of variability in gene expression observed from isogenic cell populations have influenced the view of how genes are regulated and how variability between cells arises [Elowitz et al., 2002a, Pedraza and van Oudenaarden, 2005, Rosenfeld et al., 2005]. Recent investigations draw attention to the assumption in the current studies that the fluorescent protein expression reflects the endogenous protein expression [Chubb, Dong and McMillen, 2008, Wang et al., 2008], potentially leading to errors in interpretation. Here we confirm this findings indicating that in order to accurately explain the magnitude, origins and temporal dynamics of variability in gene expression from fluorescence measurements a mathematical model is required, that accounts for the properties of the reporter protein. Our novel inference framework accounts for this factor and therefore allows us to reliably obtain a dynamical, detailed picture of the noise in terms of the model parameters.

4.8 Supplementary Information

This section contains details of mathematical models and statistical methods used in the previous sections of this chapter.

4.8.1 Derivation of the model equations

In this section we derive model equations (4.1-4.8) described in the section 4.4. As the mathematical theory of modelling chemical reactions is well established [Van Kampen, 2006] we start with brief review of mathematical methods and derive our model as a particular case of a general system.

General framework for modelling of chemical kinetics

Our derivations in this subsection follow [Van Kampen, 2006] and [Elf and Ehrenberg, 2003].

The chemical master equation (CME) is the primary tool to model the stochastic behaviour of a reacting chemical system. It describes the evolution of the joint probability distribution of the number of different molecular species in a spatially homogeneous, well stirred and thermally equilibrated chemical system [Gillespie, 1992a]. Even though these assumptions are not necessarily satisfied in living organisms the CME is commonly regarded as the most realistic model of biochemical reactions inside living cells. Consider a general system of N chemical species inside a volume Ω and let $\mathbf{X} = (X_1, \ldots, X_N)^T$ denote the number and $\mathbf{x} = \mathbf{X}/\Omega$ the concentrations of molecules. The stoichiometry matrix $\mathbf{S} = \{S_{ij}\}_{i=1,2\ldots N;\ j=1,2\ldots R}$ describes changes in the population sizes due to R different chemical events, where each S_{ij} describes the change in the number of molecules of type i from X_i to $X_i + S_{ij}$ caused by an event of type j. The probability that an event of type j occurs in the time interval $[t, t+dt)$ equals $\tilde{f}_j(\mathbf{x}, \Omega, t)\Omega dt$. The functions $\tilde{f}_j(\mathbf{x}, \Omega, t)$ are called *mesoscopic transition rates*. The probability that more than one event will take place in a small time interval is of the higher order with respect to the length of the interval. Finally, we assume that events taking place in disjoint time intervals are independent, when conditioned on the events in the previous interval.

This specification leads to a Poisson birth and death process where the probability $h(\mathbf{X}, t)$ that the system is in the state $\mathbf{X}$ at time t is described by the CME [Van Kampen, 2006]

$$\frac{dh(\mathbf{X}, t)}{dt} = \Omega \sum_{j=1}^{R} \left(\prod_{i=1}^{N} E^{-S_{ij}} - 1 \right) \tilde{f}_j(\mathbf{x}, \Omega, t) h(\mathbf{X}, t). \tag{4.22}$$

Here, $E^{-S_{ij}}$ is a step operator defined by

$$E^{-S_{ij}} f(\ldots, X_i, \ldots) = f(\ldots, X_i - S_{ij}, \ldots).$$

Macroscopic rate equation

As the system's volume Ω increases, relative fluctuations become negligible and in the limit of infinitely large Ω the system becomes deterministic. To derive the macroscopic rate equation we write the operator $\prod_{i=1}^{N} E^{-S_{ij}}$ in the form of a first order multivariate Taylor expansion

$$\prod_{i=1}^{N} E^{-S_{ij}} = 1 - \sum_{i=1}^{N} \frac{S_{ij}}{\Omega} \frac{\partial}{\partial x_i} + O(\Omega^{-2}).$$

After substitution into the CME (4.22), in the limit of infinitely large Ω we obtain

$$\frac{dh(\varphi,t)}{dt} = -\sum_{j=1}^{R} \left(\sum_{i=1}^{N} S_{ij} \frac{\partial}{\partial \phi_i} \right) f_j(\varphi,t) h(\varphi,t), \tag{4.23}$$

where $\phi_i = \lim_{\Omega\to\infty, X\to\infty} X_i/\Omega,\ \varphi = (\phi_1,\dots,\phi_N)^T$ and $f_j(\varphi,t) = \lim_{\Omega\to\infty} \tilde{f}_j(\mathbf{x},\Omega,t)$. The functions $f_j(\varphi,t)$ are called *macroscopic transition rates*.

This partial differential equation can be solved by the method of characteristics [Evans, 1998]. The solution is called the *macroscopoic rate equation* and has the form [Gardiner, 1985]

$$\frac{d\phi_i}{dt} = \sum_{j=1}^{R} S_{ij} f_j(\varphi,t) \qquad i = 1,2,...,N. \tag{4.24}$$

The Linear noise approximation

In order to obtain the linear noise approximation the transition rates, $\tilde{f}_j(\mathbf{x},t)$ and the step operator $E^{\cdot}$ are Taylor expanded around the deterministic state φ in powers of $1/\sqrt{\Omega}$. To obtain such an expansion process X_i is decomposed into the deterministic φ and stochastic $\xi = (\xi_1,...,\xi_N)^T$ components according to the relation

$$X_i \equiv \Omega\phi_i + \Omega^{1/2}\xi_i. \tag{4.25}$$

The transition rates are expanded as follows

$$\tilde{f}_j(\mathbf{x},t) = f_j(\varphi,t) + \frac{1}{\sqrt{\Omega}} \sum_{i=1}^{N} \frac{\partial f_i(\varphi,t)}{\partial \phi_i} \xi_i + O(\Omega^{-1}). \tag{4.26}$$

Similarly, we have an expansion of the step operator

$$\prod_{i=1}^{N} E^{-S_{ij}} = 1 - \Omega^{-1/2}\sum_{i=1}^{N} S_{ij}\frac{\partial}{\partial \xi_i} + \frac{1}{2\Omega}\sum_{i=1}^{N}\sum_{k=1}^{N} S_{ij}S_{kj}\frac{\partial^2}{\partial\xi_i\partial\xi_k} + O(\Omega^{-\frac{3}{2}}). \tag{4.27}$$

Let $\Pi(\xi, t)$ denote the probability distribution of ξ at time t. Using the fact that the distribution $h(\mathbf{X}, t)$ is related to $\Pi(\xi, t)$ through the relation

$$h(\mathbf{X}, t) = h(\Omega\varphi + \Omega^{1/2}\xi, t) = \Pi(\xi, t) \tag{4.28}$$

and substituting (4.26), (4.27) and (4.28) into (4.22), we obtain the Fokker-Planck equation describing the evolution of $\Pi(\xi, t)$ [Elf and Ehrenberg, 2003]

$$\frac{d\Pi(\xi, t)}{dt} = -\sum_{i,k=1}^{N}[\mathbf{A}]_{ik}\frac{\partial}{\partial\xi_i}\xi_k\Pi + \frac{1}{2}\sum_{i,k=1}^{N}\left[\mathbf{E}\mathbf{E}^T\right]_{ik}\frac{\partial^2\Pi}{\partial\xi_i\partial\xi_k}, \tag{4.29}$$

where

$$\begin{array}{rcl} f_i &=& f_i(\varphi, t),\ [\mathbf{A}]_{ik} = \displaystyle\sum_{j=1}^{R} S_{ij}\frac{\partial f_j}{\partial\phi_k}, \qquad\qquad (4.30) \\ \mathbf{E} &=& S\sqrt{diag(\mathbf{f}(\varphi, t))}, \quad \left[\mathbf{E}\mathbf{E}^T\right]_{ik} = \displaystyle\sum_{j=1}^{R} S_{ij}S_{kj}f_j. \end{array}$$

The related Itô diffusion equation has the form [Gardiner, 1985]

$$d\xi(t) = \mathbf{A}(t)\xi dt + \mathbf{E}(t)dW. \tag{4.31}$$

Model of fluorescent protein expression

The model of fluorescent gene expression introduced in section 4.4 is summarised by the following stoichiometric equations

$$\begin{array}{llllllll} \mathrm{R}_1 : DNA & \xrightarrow{k_r(t)} & DNA + R & & \mathrm{R}_5 : P & \xrightarrow{k_f} & P_f \\ \mathrm{R}_2 : R & \xrightarrow{\gamma_r} & \varnothing & & \mathrm{R}_6 : P_f & \xrightarrow{\gamma_p} & \varnothing \\ \mathrm{R}_3 : R & \xrightarrow{k_p} & R + P & & \mathrm{R}_7 : P_f & \xrightarrow{k_m} & P_m \\ \mathrm{R}_4 : P & \xrightarrow{\gamma_p} & \varnothing & & \mathrm{R}_8 : P_m & \xrightarrow{\gamma_p} & \varnothing, \end{array}$$

where R, P, P_f, P_m denote mRNA, protein, folded protein and matured protein respectively. Vectors of molecular copy numbers ($\mathbf{X}$), concentrations ($\mathbf{x}$), and macroscopic counterparts are

$$\mathbf{X} = (R, P, P_f, P_m), \qquad \mathbf{x} = (r, p, p_f, p_m), \qquad \varphi = (\phi_r, \phi_p, \phi_{p_f}, \phi_{p_m}).$$

Subsequent elements of the above vectors refer to mRNA, protein, folded protein and matured protein respectively. The mesoscopic and macroscopic transition rate vectors and stoichiometric matrix have the form

$$\tilde{\mathbf{f}}(\mathbf{x}, t) \;=\; \begin{pmatrix} k_r(t) \\ \gamma_r r \\ k_p r \\ \gamma_p p \\ k_f p \\ \gamma_p p_f \\ k_m p_f \\ \gamma_p p_m \end{pmatrix}, \qquad \mathbf{f}(\varphi, t) = \begin{pmatrix} \tau(t) \\ \gamma_r \phi_r \\ k_p \phi_r \\ \gamma_p \phi_p \\ k_f \phi_p \\ \gamma_p \phi_{p_f} \\ k_m \phi_{p_f} \\ \gamma_p \phi_{p_m} \end{pmatrix}, \tag{4.32}$$

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

As we want to model transcription rate as a stochastic process we distinguish between the mesoscopic transcription rate $k_r(t)$ and the macroscopic transcription rate $\tau(t)$.

Macroscopic rate equations

To obtain the macroscopic description of our model we put formulae (4.32) into eq. (4.24)

$$\begin{aligned} \dot{\phi}_r &= \tau(t) - \gamma_r \phi_r, \\ \dot{\phi}_p &= k_p \phi_r - (\gamma_p + k_f)\phi_p, \\ \dot{\phi}_{p_f} &= k_f \phi_p - (\gamma_p + k_m)\phi_{p_f}, \\ \dot{\phi}_{p_m} &= k_m \phi_{p_f} - \gamma_p \phi_{p_m}. \end{aligned} \tag{4.33}$$

Linear noise approximation

In the LNA the deterministic and stochastic part are separated according to (4.25). The deterministic part of our model is described by MRE (4.33). To describe the stochastic part we write drift and diffusion matrices $\mathbf{A}$, $\mathbf{E}$ according to (4.30) and (4.32)

$$\mathbf{A} = \begin{pmatrix} -\gamma_r & 0 & 0 & 0 \\ k_p & -(\gamma_p + k_f) & 0 & 0 \\ 0 & k_f & -(\gamma_p + k_m) & 0 \\ 0 & 0 & k_m & -\gamma_p \end{pmatrix}, \tag{4.34}$$

$$\mathbf{E}(t) = \begin{pmatrix} \sqrt{\tau(t)} & -\sqrt{\gamma_r\phi_r} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{k_p\phi_r} & -\sqrt{\gamma_p\phi_p} & -\sqrt{k_f\phi_p} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{k_f\phi_p} & -\sqrt{\gamma_p\phi_{p_f}} & -\sqrt{k_m\phi_{p_f}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{k_m\phi_{p_f}} & -\sqrt{\gamma_p\phi_{p_m}} \end{pmatrix}. \tag{4.35}$$

Therefore equation (4.31) that describes the stochastic part of the system has the form

$$\begin{aligned} d\xi_r &= (k_r(t) - \gamma_r\xi_r)dt + \sqrt{\tau(t) + \gamma_r\phi_r(t)}dW_1, \\ d\xi_p &= (k_p\xi_r - (\gamma_p + k_f)\xi_p)dt + \sqrt{k_p\phi_r(t) + \gamma_p\phi_p(t)}dW_2 - \sqrt{k_f\phi_p(t)}dW_3, \\ d\xi_{p_f} &= (k_f\xi_p - (\gamma_p + k_m)\xi_{p_f})dt + \sqrt{k_f\phi_p(t)}dW_3 + \sqrt{\gamma_p\phi_{p_f}(t)}dW_4 - \sqrt{k_m\phi_{p_f}(t)}dW_5, \\ d\xi_{p_m} &= (k_m\xi_{p_f} - \gamma_p\xi_{p_m})dt + \sqrt{k_m\phi_{p_f}(t)}dW_5 + \sqrt{\gamma_p\phi_{p_m}(t)}dW_6, \end{aligned} \tag{4.36}$$

The volume of the system is unknown and we set $\Omega = 1$ so that the concentration equals the number of molecules. Using the equations (4.25), (4.33) and (4.36) we obtain

$$\begin{aligned} dr &= (k_r(t) - \gamma_r r)dt + \sqrt{\tau(t) + \gamma_r\phi_r(t)}dW_1, \\ dp &= (k_p r - (\gamma_p + k_f)p)dt + \sqrt{k_p\phi_r(t) + \gamma_p\phi_p(t)}dW_2 - \sqrt{k_f\phi_p(t)}dW_3, \\ dp_f &= (k_f p - (\gamma_p + k_m)p_f)dt + \sqrt{k_f\phi_p(t)}dW_3 + \sqrt{\gamma_p\phi_{p_f}(t)}dW_4 - \sqrt{k_m\phi_{p_f}(t)}dW_5, \\ dp_m &= (k_m p_f - \gamma_p p_m)dt + \sqrt{k_m\phi_{p_f}(t)}dW_5 + \sqrt{\gamma_p\phi_{p_m}(t)}dW_6. \end{aligned} \tag{4.37}$$

Extension by extrinsic noise

As described in section 4.4 we model the extrinsic noise as a stochastic transcription process

$$k_r(t) = \tau(t)(1 + \zeta(t)) \tag{4.38}$$

where ζ is an Ornstein-Uhlenbeck process

$$d\zeta = (-\gamma_\zeta \zeta)dt + \sigma_\zeta dW_7. \tag{4.39}$$

Therefore the final system composed of equations (4.33) and (4.39) can be written in the form

$$d\mathbf{x} = (\mathbf{A}(t)\mathbf{x} + \mathbf{F}(t))dt + \mathbf{E}(t)dW, \tag{4.40}$$

where $\mathbf{x} = (\xi, r, p, p_f, p_m)$,

$$\mathbf{A}(t) = \begin{pmatrix} -\gamma_\zeta & 0 & 0 & 0 & 0 \\ \tau(t) & -\gamma_r & 0 & 0 & 0 \\ 0 & k_p & -(\gamma_p + k_f) & 0 & 0 \\ 0 & 0 & k_f & -(\gamma_p + k_m) & 0 \\ 0 & 0 & 0 & k_m & -\gamma_p \end{pmatrix}, \tag{4.41}$$

$$\mathbf{E}(t) = \begin{pmatrix} \sigma_\xi & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\tau(t)} & -\sqrt{\gamma_r \phi_r} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{k_p \phi_r} & -\sqrt{\gamma_p \phi_p} & -\sqrt{k_f \phi_p} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{k_f \phi_p} & -\sqrt{\gamma_p \phi_{p_f}} & -\sqrt{k_m \phi_{p_f}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{k_m \phi_{p_f}} & -\sqrt{\gamma_p \phi_{p_m}} \end{pmatrix} \tag{4.42}$$

and

$$\mathbf{F}(t) = (0, \tau(t), 0, 0, 0)^T. \tag{4.43}$$

4.8.2 Derivation of the autocorrelation function and power spectral density

In this section we derive the autocorrelation function (ACF) and power spectral density for the stationary state of the stochastic process given by equation (4.40). If the matrix $\mathbf{A}$ is time independent and all its eigenvalues have negative real parts

then the stationary state exists [Gardiner, 1985]. Therefore, we assure existence of the stationary state setting $\tau(t) = b$. As the degradation, folding and oxidation rates are positive it is straightforward to verify that all eigenvalues are negative.
It can be shown [Gardiner, 1985] that for an equation of type (4.40) the autocorrelation function has the form

$$ACF(t) = \langle(\mathbf{x}(s+t) - \langle\mathbf{x}(s+t)\rangle)(\mathbf{x}(s) - \langle\mathbf{x}(s)\rangle)^T\rangle = \exp(\mathbf{A}t)\Xi \quad \text{for } t \geq 0, \tag{4.44}$$

where matrix Ξ is a stationary covariance matrix arising by the fluctuation-dissipation theorem [Gardiner, 1985] as a solution of the equation

$$\mathbf{A}\Xi + \Xi\mathbf{A}^T + \mathbf{E}\mathbf{E}^T = 0. \tag{4.45}$$

Element $(5,5)$ of the matrix $ACF(t)$ gives the autocorrelation function of the process p_m $ACF_{p_m}(t)$. Matrix $\mathbf{A}$ has five different eigenvalues: $\lambda_1 = -\gamma_\zeta$, $\lambda_2 = -\gamma_r$, $\lambda_3 = -(\gamma_p + k_f)$, $\lambda_4 = -(\gamma_p + k_m)$, $\lambda_5 = -\gamma_p$. This and eq. (4.44) imply that for $t \geq 0$ $ACF_{p_m}(t)$ can be represented as

$$\begin{aligned} ACF_{p_m}(t) &= a_1\exp(-\gamma_\zeta t) + a_2\exp(-\gamma_r t) \qquad (4.46)\\ &+ a_3\exp(-(\gamma_p + k_f)t) + a_4\exp(-(\gamma_p + k_m)t) + a_5\exp(-\gamma_p t). \end{aligned}$$

Constants a_1, a_2, a_3, a_4, a_5 are functions of the model parameters and do not depend on time. They have to complicated form to be presented here. As discusses in section 4.4 this form of ACF_{p_m} demonstrates that there are 5 parameters that jointly determine memory of the observed fluctuations.
The Fourier transform of the autocorrelation function (4.44) gives a power spectral density $S(\omega)$ of the stationary state of equation (4.40). It can be shown [Gardiner, 1985] that for equations of type (4.40) the power spectral density has the form

$$S(\omega) = \frac{1}{2\pi}(\mathbf{A} + i\omega)\mathbf{E}\mathbf{E}^T(\mathbf{A}^T - i\omega)^{-1}. \tag{4.47}$$

The spectral density $S(\omega)$ is a 5×5 matrix where the element $(5,5)$ corresponds to the spectral density of fluctuations of the matured proteins $S_{p_m}(\omega)$. Using matrix manipulation software [Maple] we obtained

$$
\begin{aligned}
S_{p_m}(\omega) \;=\;& \frac{1}{2\pi}\frac{{k_m}^2{k_f}^2{k_p}^2{k_r}^2{\sigma_\zeta}^2}{\left((\gamma_p+k_m)^2+\omega^2\right)\left((\gamma_p+k_f)^2+\omega^2\right)({\gamma_r}^2+\omega^2)({\gamma_\zeta}^2+\omega^2)({\gamma_p}^2+\omega^2)}\\
+\;& \frac{{k_m}^2{k_f}^2{k_p}^2{q_r}^2}{\left((\gamma_p+k_m)^2+\omega^2\right)\left((\gamma_p+k_f)^2+\omega^2\right)({\gamma_r}^2+\omega^2)({\gamma_p}^2+\omega^2)}\\
+\;& \frac{{k_m}^2{k_f}^2\left({q_p}^2+{q_{pf}}^2\right)}{\left((\gamma_p+k_m)^2+\omega^2\right)({\gamma_p}^2+\omega^2)\left((\gamma_p+k_f)^2+\omega^2\right)}\\
+\;& \frac{{k_m}^2\left({q_{pf}}^2+{q_f}^2+{q_{fm}}^2\right)}{\left((\gamma_p+k_m)^2+\omega^2\right)({\gamma_p}^2+\omega^2)}\\
+\;& \frac{{q_{fm}}^2+{q_m}^2}{{\gamma_p}^2+\omega^2}\\
-\;& 2\,\frac{k_m^2k_fq_{pf}^2(\gamma_p+k_f)}{\left((\gamma_p+k_m)^2+\omega^2\right)({\gamma_p}^2+\omega^2)\left((\gamma_p+k_f)^2+\omega^2\right)}\\
-\;& 2\,\frac{k_mq_{fm}^2}{\left((\gamma_p+k_m)^2+\omega^2\right)({\gamma_p}^2+\omega^2)}
\end{aligned}
\tag{4.48}
$$

where

$$
\begin{aligned}
q_r &= \sqrt{2b},\\
q_p &= \sqrt{\frac{k_pb}{\gamma_r}+\frac{k_pb\gamma_p}{\gamma_r\left(\gamma_p+k_f\right)}},\\
q_f &= \sqrt{\frac{\gamma_pk_fk_pb}{\left(\gamma_p+k_m\right)\gamma_r\left(\gamma_p+k_f\right)}},\\
q_m &= \sqrt{\frac{k_mk_fk_pb}{\left(\gamma_p+k_m\right)\gamma_r\left(\gamma_p+k_f\right)}},\\
q_{pf} &= \sqrt{\frac{k_fk_pb}{\gamma_r\left(\gamma_p+k_f\right)}},\\
q_{fm} &= \sqrt{\frac{k_mk_fk_pb}{\left(\gamma_p+k_m\right)\gamma_r\left(\gamma_p+k_f\right)}}.
\end{aligned}
\tag{4.49}
$$

This representation shows how the stochasticity arising at each step of expression contributes to the observed variability in terms of frequencies. For instance the first element of the sum in (4.48) shows that variability generated at the transcriptional level (term containing σ_ζ^2) undergoes a low pass filtering (first term of the spectrum is a quickly decreasing function of ω). Therefore fast transcriptional fluctuation (large γ_ξ) will be filtered out. The strength of the filtering depends on γ_r, γ_p, k_f, k_m. For large values of these parameters high frequencies have a smaller contribution to the observed variability.

4.8.3 Derivation of the likelihood function

In this section we derive the likelihood function 4.15. First, we find the transition densities of the process $\mathbf{x}(t)$ defined by eq. (4.40). For an initial condition $\mathbf{x}(t_i) = \mathbf{x}_{t_i}$ and $t > t_i$ eq. (4.40) has a solution of the form [Arnold, 1974]

$$\mathbf{x}(t) = \Phi_{t_i}(t{-}t_i)\left(\mathbf{x}_{t_i} + \int_{t_i}^{t} \Phi_{t_i}(s-t_i)^{-1}\mathbf{F}(s)ds + \int_{t_i}^{t} \Phi_{t_i}(s-t_i)^{-1}\mathbf{E}(s)dW(s)\right), \tag{4.50}$$

where the integral $dW(s)$ is in Itô sense and $\Phi_{t_i}(s)$ is the fundamental matrix of the non-autonomous system of ordinary differential equations (ODEs)

$$\frac{d\Phi_{t_i}}{ds} = \mathbf{A}(t_i+s)\Phi_{t_i}, \quad \Phi_{t_i}(0) = I. \tag{4.51}$$

Equations (4.50) and (4.51) imply that the transition densities of the process $\mathbf{x}$ are Gaussian [Oksendal, 1992]

$$\mathbf{p}(\mathbf{x}_{t_i}|\mathbf{x}_{t_{i-1}}, \Theta) = \psi(\mathbf{x}_{t_i}|\mu_{i-1}, \Xi_{i-1}) \tag{4.52}$$

where Θ denotes the vector of all model parameters, $\psi(\cdot|\mu_{i-1}, \Xi_{i-1})$ is the normal density with mean μ_{i-1} and covariance Ξ_{i-1} specified by

$$\begin{aligned} \mu_{i-1} &= \Phi_{t_{i-1}}(t_i - t_{i-1})\left(x_{t_{i-1}} + \int_{t_{i-1}}^{t_i} \Phi_{t_{i-1}}(s-t_{i-1})^{-1}\mathbf{F}(s)ds\right), \quad (4.53)\\ \Xi_{i-1} &= \int_{t_{i-1}}^{t_i} (\Phi_s(t_i-s)\mathbf{E}(s))(\Phi_s(t_i-s)\mathbf{E}(s))^T ds. \end{aligned}$$

We use the transition densities of the process $\mathbf{x}$ to find the probability distribution of the vector $\mathbf{y} = (p_{m t_0}, ..., p_{m t_n})$. To do so we assume that the distribution of $\mathbf{x}_{t_0}$ is also Gaussian with mean $\varphi(t_0)$ and covariance matrix Ξ_{-1}. This assumption is natural, as equation (4.40) implies a Gaussian distribution of $\mathbf{x}_t$ for a fixed initial condition $\mathbf{x}_{t_0}$. Using this assumption and eq. (4.52), (4.53) it is straightforward to write $\mathbf{x}_{t_i}$ as

$$\mathbf{x}_{t_i} = \varphi_{t_i} + \sum_{j=0}^{i} \Phi_{t_j}(t_i - t_j)\varsigma_{t_j}, \tag{4.54}$$

where ς_{t_j} are independently normally distributed random variables with mean 0 and covariance matrix Ξ_{j-1}. Let $\bar{\mathbf{x}} = (\mathbf{x}_{t_0}, ..., \mathbf{x}_{t_n})$.

The representation (4.54) implies that

$$P(\bar{\mathbf{x}}|\Theta) = \psi(\bar{\mathbf{x}}|(\varphi_{t_0}, \ldots, \varphi_{t_n}), \hat{\Sigma}), \tag{4.55}$$

where the covariance matrix $\hat{\Sigma} = \{\hat{\Sigma}^{(i,j)}\}_{i,j=0,\ldots,n}$, is a $5(n+1) \times 5(n+1)$ block matrix that is composed of the 5×5 submatrices $\hat{\Sigma}^{(i,j)} = Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_j})$. It follows from representation (4.54) that covariances $Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_j})$ can be computed using the following relations $(j \geq i)$

$$\begin{aligned} Cov(\mathbf{x}_{t_0}, \mathbf{x}_{t_0}) &= \Xi_{-1}, \qquad (4.56)\\ Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_i}) &= \Xi_{i-1} + \Phi_{t_{i-1}}(t_i - t_{i-1}) Cov(\mathbf{x}_{t_{i-1}}, \mathbf{x}_{t_{i-1}}) \Phi_{t_{i-1}}(t_i - t_{i-1})^T, \\ Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_{j+1}}) &= Cov(\mathbf{x}_{t_i}, \mathbf{x}_{t_j}) \Phi_{t_j}(t_{j+1} - t_j)^T. \end{aligned}$$

In order to find the likelihood function $P(\bar{\mathbf{y}}|\Theta)$ for the data vector $\mathbf{y} = (p_{mt_0}, ..., p_{mt_n})$ from the augmented likelihood (4.55) we use the fact that the marginal distributions of the normal distribution are normal. Thus, we obtain

$$P(\mathbf{y}|\Theta) = \psi(\mathbf{y}|(\phi_{p_m}(t_0), \ldots, \phi_{p_m}(t_n)), \Sigma), \tag{4.57}$$

where the covariance matrix $\Sigma = \{\sigma^2_{(i,j)}\}_{i,j=0,\ldots,n}$ and $\sigma^2_{(i,j)} = Cov(p_{mt_i}, p_{mt_j})$. Therefore the covariances $\sigma^2_{(i,j)}$ are given by the elements $(5, 5)$ of matrices $\hat{\Sigma}^{(i,j)}$.

Estimation of the initial mean and the covariance matrix

Calculation of the likelihood function (4.57) requires the specification of the initial mean vector $\varphi(t_0)$ and the initial covariance matrix Ξ_{-1}. A natural solution to this problem is to estimate them, in the way all the other parameters of the model are estimated. Nevertheless, sometimes we know more about $\varphi(t_0)$ and Ξ_{-1} from the conditions under which the experimental data are obtained. For instance, for some experiments, it may be reasonable to assume that at time t_0 the system is in the steady state. In this case $\varphi(t_0)$ and Ξ_{-1} can be expressed as the stationary mean and the stationary variance that are functions of the parameters of eq. (4.40). For equation (4.40) the steady state exists for a constant macroscopic transcription rate $\tau(t) = b$ as the values of degradation, folding and oxidation rates are positive and imply negative eigenvalues of matrix $\mathbf{A}$. Therefore in the examples of the *stationary gene expression* in section 4.6 we calculate the initial mean as

$$\varphi(t_0) = -\mathbf{A}^{-1}\mathbf{F}. \tag{4.58}$$

The initial covariance matrix can be obtained using fluctuation-dissipation theorem as the solution of the equation

$$\mathbf{A}\Xi_{-1} + \Xi_{-1}\mathbf{A}^T + \mathbf{E}\mathbf{E}^T = 0. \tag{4.59}$$

In many situations the system of interest is not in the stationary state at time t_0. Nevertheless if we can predict its behaviour before t_0 we can also calculate the initial mean vector and initial covariance matrix. We demonstrate this idea using the example of the *oscillatory gene expression* from section 4.6. For a given parameter vector Θ we can calculate the transcription rate $\tau(t)$ for $t \leq t_0$. Therefore we choose $t_{00} \ll t_0$ and assume that $\mathbf{x}_{t_{00}} \sim N(\mu_{00}, \Xi_{00})$, where μ_{00}, Ξ_{00} are set arbitrarily. Using equation (4.53) we calculate $\varphi(t_0)$ and Ξ_{-1}. If $t_0 - t_{00}$ is large enough then the system "forgets" influence of μ_{00}, Ξ_{00} as the eigenvalues of $\mathbf{A}(t)$ are negative for all $t \in [t_{00}, t_0]$. In this way we can obtain a good approximation of $\varphi(t_0)$ and Ξ_{-1}.

4.8.4 Data generation

The standard Gillespie algorithm [Gillespie, 1977] allows to simulate data for systems with constant reaction rates. In our model transcription rate $k_r(t)$ is a stochastic process. Hence, to generate data we use a modified version of Gillespie's algorithm proposed in [Shahrezaei et al., 2008]. We briefly summarise this approach. If the transcription rate is time dependent, then for a given sample path of $k_r(t)$ the probability density of the transcription reaction occurring at time τ is

$$P(\tau) = k_r(\tau) \exp\left(\int_0^\tau k_r(t)dt\right). \tag{4.60}$$

During the simulation we need to sample τ from $P(\tau)$. Therefore we generate a sample path of $k_r(t)$ defined by eq. (4.38) using the Euler method [Kloeden and E., 1999]. We sample a random number υ from uniform distribution on $[0,1]$ and find τ by solving

$$\int_0^\tau P(s)ds = \upsilon. \tag{4.61}$$

Equivalently τ can be found as a solution to the simpler equation

$$\int_0^\tau k_r(s)ds = \log(\frac{1}{\upsilon}) \tag{4.62}$$

which has been obtained by inserting (4.60) into (4.61) and performing the integration. To solve (4.62) for τ we follow the procedure proposed in [Shahrezaei et al., 2008].

4.8.5 Notes on the practical implementation of the algorithm

In this section we discuss the details of the Metropolis-Hastings (MH) algorithm used to sample from posterior distribution 4.20.

Model parameterization

First we describe reparameterization of our model that allows us to reduce autocorrelation of the chains generated using the MH algorithm.
We focus on the case of the *oscillatory gene expression*, where the vector of the model parameters Θ has the form

$$\Theta = (\gamma_r, \gamma_p, b_0, b_1, b_2, b_3, k_p, \gamma_\zeta, \sigma_\zeta^2, \lambda, k_f, k_m). \tag{4.63}$$

Parameterization of the *stationary gene expression* is done analogously. Instead of using parameters Θ we parametrize the model in terms of $\bar{\Theta} = (\bar{\theta}_0, ..., \bar{\theta}_{11})$ such that $\Theta = \nu(\bar{\Theta})$, where function ν is defined as follows

$$\begin{aligned}
\gamma_r &= \exp(\bar{\theta}_0), \\
\gamma_p &= \exp(\bar{\theta}_1), \\
\lambda &= \exp(\bar{\theta}_4), \\
k_m &= \frac{\exp(\bar{\theta}_{11}))}{\exp(\bar{\theta}_4)}, \\
k_f &= \frac{\exp(\bar{\theta}_3))}{\exp(\bar{\theta}_{11})}, \\
k_p &= \frac{\exp(\bar{\theta}_8))}{\exp(\bar{\theta}_3)}, \\
b_3 &= \frac{\exp(\bar{\theta}_5))}{\exp(\bar{\theta}_8)}, \\
b_0 &= \frac{\exp(\bar{\theta}_2))}{\exp(\bar{\theta}_8)}, \\
b_1 &= \bar{\theta}_6, \\
b_2 &= \bar{\theta}_7, \\
\gamma_\zeta &= \exp(\bar{\theta}_9), \\
\sigma_\zeta^2 &= \exp(\bar{\theta}_{10}).
\end{aligned} \tag{4.64}$$

Therefore we have the following relation between the probability distributions expressed in terms of Θ and $\bar{\Theta}$ [Gamerman and Lopes, 2006]

$$P(\mathbf{U}|\bar{\Theta}) \propto P(\mathbf{U}|\nu(\bar{\Theta}))\pi(\nu(\bar{\Theta}))|J(\nu(\bar{\Theta}))|, \tag{4.65}$$

where $|J(\nu(\bar{\Theta}))|$ is the determinant of the Jacobian matrix of parameterization ν. It is straightforward to verify that

$$|J(\nu(\bar{\Theta}))| = \exp(\bar{\theta}_0)\exp(\bar{\theta}_1)\exp(\bar{\theta}_5)\exp(\bar{\theta}_2)\exp(-\bar{\theta}_8)\exp(\bar{\theta}_9)\exp(\bar{\theta}_{10}). \tag{4.66}$$

In further sections we denote $P(\mathbf{U}|\nu(\bar{\Theta}))\pi(\nu(\bar{\Theta}))|J(\nu(\bar{\Theta}))|$ by $\bar{P}(\bar{\Theta}, \mathbf{U})$.

Updating $\bar{\Theta}$

Parameter vector $\bar{\Theta}$ is updated using a random-walk Metropolis algorithm. Let $\bar{\Theta}^{(i)}$ be the value of $\bar{\Theta}$ at iteration i of the MCMC algorithm. A new value $\bar{\Theta}^{(new)}$

is proposed from the symmetric proposal distribution

$$\bar{\Theta}^{(new)} \sim N(\bar{\Theta}^{(i)}, \Lambda).$$

The new value $\bar{\theta}_j^{(new)}$ is then accepted with probability

$$\min\left\{1, \frac{\bar{P}(\bar{\Theta}^{(new)}, \mathbf{U})}{\bar{P}(\bar{\Theta}^{(i)}, \mathbf{U})}\right\}.$$

If $\bar{\Theta}^{(new)}$ is not accepted then $\bar{\Theta}^{(i+1)} = \bar{\Theta}^{(i)}$. The covariance matrix of the proposal distribution Λ is tuned carefully in order to ensure an efficient exploration of the parameter space. If the proposed moves are too large, too small or do not reflect correlations between parameters then the convergence of the chain may be very slow. To obtain a "good" Λ we run prior, short simulation of a chain with an arbitrary Λ and estimate a new Λ based on the generated chain. The new Λ usually achieves good convergence in the main simulation.

Computation of the likelihood

Here we give a summary of the computation of the likelihood function 4.15. We assume that the initial condition $\varphi(t_0)$ and initial covariance matrix Ξ_{-1} have been found according to the procedure described in the section 4.8.3. Computation is performed as follows

```
1 Numerically find $\varphi(t)$ for $t \in [t_0, t_n]$ ;
2 For $i = 0, ..., n-1$ numerically find fundamental matrices $\Phi_{t_i}(t_{i+1}-t_i)$
3 Find numerically matrices $\Xi_{i-1}$ for $i = 0, ...n$;
4 Use matrices computed in steps 2 and 3 to construct covariance matrix $\hat{\Sigma}$ according to the procedure from section 4.8.3;
5 Extract covariance matrix $\Sigma$ from $\hat{\Sigma}$ (according to section 4.8.3);
6 For given data $\mathbf{u}$ evaluate multivariate normal density with mean vector
```

$\lambda(\varphi(t_0), ..., \varphi(t_{n-1}))$ `and covariance matrix` $\lambda^2\Sigma + \Sigma_\epsilon$, `where` λ `and` Σ_ϵ `are defined in section 4.4.`

Numerical approximation of fundamental matrices

Consider the linear ODE

$$\frac{d\Phi_s}{dt} = \mathbf{A}(s+t)\Phi_s, \tag{4.67}$$

where $\mathbf{A}(s+t)$ and Φ_s are an $N \times N$ matrices. Let $\Phi_s(t)$ be the solution of this with initial condition the identity matrix i.e. $\Phi_s(0) = I$. In order to compute the transition density covariances Ξ_{i-1} (eq. (4.53)), it is necessary to find these matrices. This can be done by solving the equation directly, which gives $\Phi_s(t)$ as t varies. For computation of matrices Ξ_{i-1}, however, it is more convenient to iterate backwards the equation for $\Phi_s(t-s)$, with an initial condition $\Phi_t(0) = I$, as a function of s.

To derive the equation for $\Phi_s(t-s)$ we use the fact that

$$\frac{d}{ds}\left(\Phi_0(s)^{-1}\Phi_0(s)\right) = \left(\frac{d}{ds}(\Phi_0(s)^{-1}\right)\Phi_0(s) + \Phi_0(s)^{-1}\frac{d}{ds}\Phi_0(s) = 0. \tag{4.68}$$

Therefore

$$\frac{d}{ds}\Phi_0(s)^{-1} = -\Phi_0(s)^{-1}\left(\frac{d}{ds}\Phi_0(s)\right)\Phi_0(s)^{-1} \tag{4.69}$$

and

$$\frac{d}{ds}\Phi_s(t-s) = \frac{d}{ds}\left(\Phi_0(t)\Phi_0(s)^{-1}\right) = -\Phi_0(t)\Phi_0(s)^{-1}\left(\frac{d}{ds}\Phi_0(s)\right)\Phi_0(s)^{-1}. \tag{4.70}$$

Finally

$$\frac{d}{ds}\Phi_s(t-s) = -\Phi_s(t-s)A(s). \tag{4.71}$$

Chapter 5

Inferring parametric distributions of reporter protein degradation rates with a Bayesian hierarchical model

5.1 Author Contributions and chapter's structure

This chapter is a result of a joint work between Michał Komorowski, Claire V. Harper and Bärbel Finkenstädt. MK did numerical simulations and wrote the chapter, CVH conduced cycloheximide experiment, BF suggested usage of hierarchical modelling and supervised the study.

Sections 5.2 - 5.5 are followed by supplementary section 5.6 that contains details of mathematical modeling and statistical methods.

5.2 Introduction

Understanding biological processes on the molecular level is fundamental to our investigation of cellular phenomena (e.g. [Blake et al., 2006, Losick and Desplan,

2008]). Careful use of mathematical models allows us to provide detailed description of the dynamics of complex biochemical interactions underlying functioning of living cells [Chabot et al., 2007, Hoffmann et al., 2002, Nelson et al., 2004]. The dynamics of gene expression is an example of a system where mathematical modelling proved to be particularly useful in bringing insight into understanding of cellular processes [Paulsson, 2005, 2006, Swain et al., 2002a, Thattai and van Oudenaarden, 2001].
In addition, recent development in fluorescent microscopy technology enabled us to measure levels of reporter proteins in vivo [Tsien, 1998, Wu and Pollard, 2005]. However, to understand how the observed fluorescence level relates to the dynamics of gene expression the knowledge of mRNA and reporter protein degradation rates is extremely useful [Chabot et al., 2007, Finkenstadt et al., 2008, Heron et al., 2007]. As reporter proteins are applied in a variety of different systems which may have different mRNA and protein degradation rates we need a robust method to determine these rates for the relevant experimental conditions.

The standard method for estimating the degradation rates of fluorescent reporter proteins is to treat a cell culture with a translational inhibitor to stop the formation of the protein [Gordon et al., 2007] and fit exponential curve to the obtained averaged measurements. This method is not without its problems. Firstly, inhibiting translation invariably results in the death of the cell so estimating the rates has to be done on a cell sample that is separate to the sample used in an experiment of interest. As such taking the average of this sample may not correspond to the samples used in the experiment, biasing the result. Furthermore, degradation rates naturally vary between cells so using a fixed degradation rate for all the samples in the experiment may introduce errors, especially if the variance of the degradation rates is large. Additionally, translation is never fully inhibited as some protein molecules will still be created biasing the degradation rate estimates.

In this chapter we present a method that overcomes these problems. It is achieved by the creation of suitable model that parameterizes protein degradation rates. Our model is extended to estimate the population distribution of the degradation rates which allows us to incorporate information on the cellular variation of the rates.

In the next section we derive the standard model for the determination of reporter protein degradation rates and propose an inference framework that is embedded in a Bayesian hierarchical approach. Then we apply the method to experimental, fluorescent reporter gene data.

5.3 Model

Under the influence of translational inhibitor translation level drops to a small basal level k_p and the initial protein level begins to decrease at rate γ_p [Gordon et al., 2007]. Therefore the natural model can be expressed by the single ordinary differential equation

$$\dot{\phi}_p \quad = \quad k_p - \gamma_p \phi_p, \tag{5.1}$$

where ϕ_p is the protein concentration. Henceforth we call ϕ_p the macroscopic protein concentration. As the single cell experimental data exhibit non negligible variability we use the stochastic model. To determine its analytical form we use the linear noise approximation [Elf and Ehrenberg, 2003, Komorowski et al., 2009a, Van Kampen, 2006] and obtain

$$dp \quad = \quad (k_p - \gamma_p p)dt + \sqrt{k_p + \gamma_p \phi_p} dW, \tag{5.2}$$

where p is the protein concentration and ϕ_p is the macroscopic protein concentration given by a solution of eq. (5.1) and W is a Wiener process. The unique solution of equation (5.2) requires the specification of the initial conditions $\phi_p(t_0), p(t_0)$. As we do not know the history of a cell before time t_0 we assume that $\phi_p(t_0) = p(t_0)$ and treat this quantity as a model parameter.

Considering the relation between experimental measurements cellular protein concentration we assume that single cell measurements u_t are taken at times $t_0, ..., t_n$ and that the data vector has the form

$$\mathbf{u} = (u_{t_0}, ..., u_{t_n}). \tag{5.3}$$

Fluorescent reporter data are usually assumed to be proportional to the number of fluorescent molecules [Wu and Pollard, 2005] and measurements are subject to measurement error. Therefore we assume that each measured u_{t_i} is related to the model variable p_{t_i} through the relation

$$u_{t_i} = \lambda p_{mt_i} + \epsilon_{t_i} \tag{5.4}$$

where λ is an unknown proportionality constant and ϵ_{t_i} is a measurement error. For mathematical convenience we assume that the joint distribution of the measurement error is normal with mean 0 and known covariance matrix Σ_ϵ, i.e. $(\epsilon_{t_0}, ..., \epsilon_{t_n}) \sim N(0, \Sigma_\epsilon)$. If measurement errors are independent with a constant variance σ_ϵ^2 then $\Sigma_\epsilon = \sigma_\epsilon^2 I$.

It can be shown that $\mathbf{u}$ has a multivariate Gaussian distribution (see Chapter 3 or [Komorowski et al., 2009a])

$$P(\mathbf{u}|\theta) = \psi(\mathbf{u}|\lambda\mu(\theta), \lambda^2\Sigma(\theta) + \Sigma_\epsilon), \tag{5.5}$$

where ψ denotes multivariate Gaussian density with mean vector $\lambda\mu(\theta)$ and covariance matrix $\lambda^2\Sigma(\theta) + \Sigma_\epsilon$. These are explicit functions of model parameters (see supplementary section 5.6).

Assume now that we observe l cells simultaneously. In this case data matrix has the form

$$\mathbf{U} = (\mathbf{u}^{(1)}, ..., \mathbf{u}^{(l)}). \tag{5.6}$$

Usually experimental data indicate that values of kinetic parameters may differ between cells. Bayesian hierarchical modelling [Gamerman and Lopes, 2006] provides a natural framework to account for this variability.

To build a hierarchical model we assume that each time series $\mathbf{u}_i$ is a realisation of the process (5.2) with initial condition $p^{(i)}(0)$ and parameters $\theta^{(i)} = (\gamma_p^{(i)}, k_p^{(i)}, \lambda^{(i)})$. We assume that for each cell these parameters are drown from the following distributions

$$\gamma_p^{(i)} \sim \Gamma(\mu_{\gamma_p}, \sigma^2_{\gamma_p}), \qquad k_p^{(i)} \sim \Gamma(\mu_{k_p}, \sigma^2_{k_p}), \qquad \lambda^{(i)} \sim \Gamma(\mu_\lambda, \sigma^2_\lambda),$$

where $\Gamma(\mu_\cdot, \sigma^2_\cdot)$ denotes a gamma density with mean $\mu_\cdot$ and variance $\sigma^2_\cdot$. Let $\Theta = (\mu_{\gamma_p}, \sigma^2_{\gamma_p}, \mu_{k_p}, \sigma^2_{k_p}, \mu_\lambda, \sigma^2_\lambda, \sigma^2_\epsilon)$. Assuming independence of $\gamma_p^{(i)}, k_p^{(i)}, \lambda^{(i)}$ the above give the distribution of a vector $\theta^{(i)}$ that we denote by $P(\theta^{(i)}|\Theta)$. Hierarchical modelling aims to estimate the parameter vector Θ via the posterior distribution $P(\Theta, \theta^{(1)}, ..., \theta^{(l)}, p^{(1)}(0), ..., p^{(l)}(0)|\mathbf{U})$ [Gamerman and Lopes, 2006]. Using Bayes' rule we can write

$$P(\Theta, \theta^{(1)}, ..., \theta^{(l)}, p^{(1)}(0), ..., p^{(l)}(0)|\mathbf{U}) \propto \prod_{i=1}^{l} P(\mathbf{u^{(i)}}|p^{(i)}(0), \theta^{(i)}) P(\theta^{(i)}|\Theta)\pi(\Theta), \tag{5.7}$$

where $\pi(\Theta)$ is a prior distribution of vector Θ. As $P(\mathbf{u^{(i)}}, p^{(i)}(0)|\theta^{(i)})$, $P(\theta^{(i)}|\Theta)$ are given by analytical formulae after the prior distribution $\pi(\Theta)$ is specified the standard Metropolis-Hasting algorithm [Gamerman and Lopes, 2006] may be used to generate samples from posterior distribution (5.7).

5.4 Results

Here we present results where we inferred the degradation rates distribution from data obtained in an experiment in which translation was blocked by the addition of cycloheximide (CHX) and fluorescence was imaged every 5.6 minutes for 10h in 42 cells (Details of the experiment can be found in supplementary section 5.6). The data are presented in Figure 5.1.

We used a standard Metropolis-Hastings algorithm to generate samples from the posterior distribution $P(\Theta, \theta^{(1)}, ..., \theta^{(l)}, p^{(1)}(0), ..., p^{(l)}(0)|\mathbf{U})$ and use posterior medians together with 95% credibility intervals as summary statistics. Results are presented in Table 5.1. The corresponding parametric distributions are plotted in Figure 5.2. Measurement errors were assumed to be independent across cells and times with mean zero and standard deviation σ_ϵ. An exponential distributions with mean 10 has been used as the uninformative prior for each of the parameters $\mu_{\gamma_p}, \sigma^2_{\gamma_p}, \mu_{k_p}, \sigma^2_{k_p}, \mu_\lambda, \sigma^2_\lambda, , \sigma^2_\epsilon$. For initial values $p^{(i)}(0)$ we used flat priors.

5.5 Discussion

In this chapter we have presented a novel method for the estimation of fluorescent protein degradation rates. The method is embedded in the Bayesian hierarchical modelling framework. Hierarchical approach allows us to quantify the variability of kinetic parameters between cells. Well established methods infer a single value of a degradation rate, whereas our approach provides not only values of degradation rates in individual cells but also a mean and variance of degradation rates in the population. Instead of a deterministic approach we used a stochastic model based on the linear noise approximation. It has the advantage that it accounts for stochasticity in degradation and translation events so that it may be applied to single-cell data without a risk of obtaining biased estimates.

To perform parameter inference we derived an analytic formula for the likelihood of data observed with error and used Metropolis-Hasting algorithm to generate samples from posterior distributions.

Although our hierarchical approach is applied here to the very simple stochastic model in future it may be extended to larger models of gene expression and gene regulation. This extension may be useful to quantify diversity in cellular populations, that is not only due to randomness of biochemical reactions but also result from other genetic and epigenetic factors.

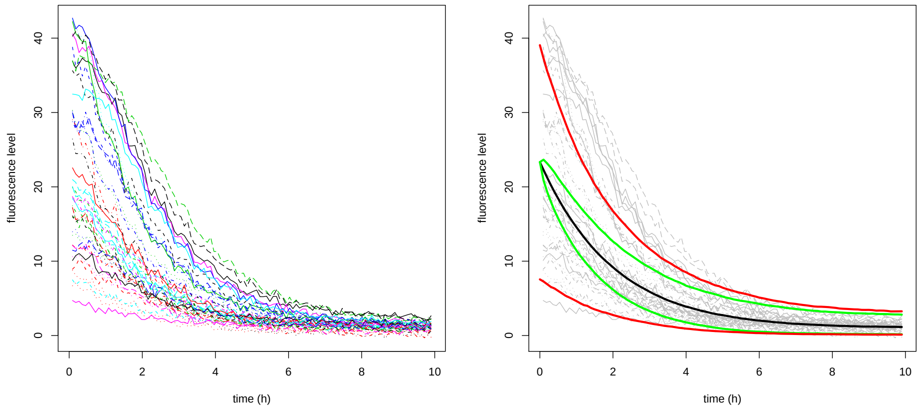

Figure 5.1: **Left:** Fluorescence level from cycloheximide experiment is plotted against time (in hours). Measurements were taken simultaneously in 42 cells every 5.6 minutes. **Right:** Fluorescence data (thin gray lines) and deterministic fit (black bold line) given by eq. 5.1 plotted using mean estimates presented in Table 5.1. The variation of SDE 5.2 is shown by the 10% and 90% quantiles (green lines) computed from 10000 simulations using mean estimates given in Table 5.1. Bold red lines present the 10% and 90% quantile of SDE 5.2 computed from 10000 simulation using rates drown from gamma distribution with parameters given in Table 5.1 and initial condition $p(0)$ drawn from normal distribution with mean and variance calculated directly from the data.

Parameter	Prior	Estimate
μ_{γ_p}	Exp(10)	0.49 (0.44-0.53)
$\sigma^2_{\gamma_p}$	Exp(10)	0.007 (0.002-0.016)
μ_{k_p}	Exp(10)	0.53 (0.37-0.73)
$\sigma^2_{k_p}$	Exp(10)	0.09 (0.02-0.29)
$\mu\lambda$	Exp(10)	6.74 (5.87-7.86)
σ^2_λ	Exp(10)	5.35 (2.57-11.26)
σ^2_ϵ	Exp(10)	1.04 (0.04-5.80)

Table 5.1: Posterior median and 95% credibility intervals of parameters μ_{γ_p}, $\sigma^2_{\gamma_p}$, $\mu_{k_p}\sigma^2_{k_p}$, μ_λ, σ^2_λ, σ^2_ϵ inferred from experimental data presented in Figure 5.1 . Rates are per hour.

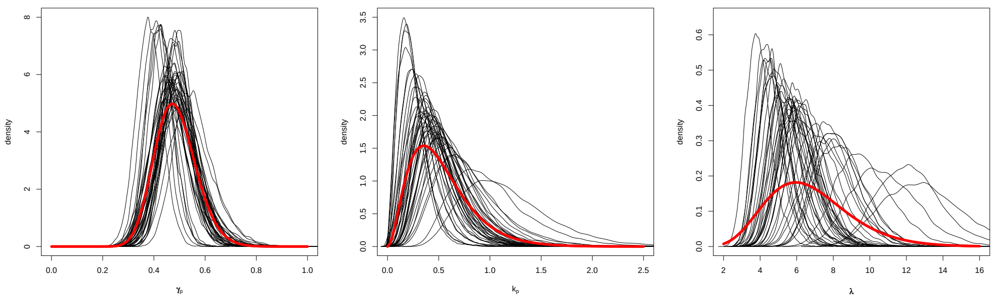

Figure 5.2: Distributions of parameters γ_p (left),k_p (middle), λ (right) inferred from data presented in Figure 5.1. Curves plotted in black correspond to posterior distributions of individual cells. Red curves present Gamma density with mean and variances presented in Table 5.1

5.6 Supplementary information

This section contains details about mathematical models and statistical methods used in the previous sections of this chapter.

5.6.1 Derivation of the mean vector and the covariance matrix

In this section we derive analytical formulae for the mean vector $\mu(\Theta)$ and the covariance matrix $\Sigma(\Theta)$.

It is straightforward to verify that solution of eq. (5.1) is given by

$$\begin{aligned} \phi_p(t) &= \phi_p(0)\exp(-\gamma_p t) + k_p \int_0^t \exp(-\gamma_p(t-s))ds \\ &= \phi_p(0)\exp(-\gamma_p t) + \frac{k_p}{\gamma_p}(1 - exp(-\gamma_p t)). \end{aligned} \tag{5.8}$$

Similarly solution of eq. (5.2) has the form [Arnold, 1974]

$$\begin{aligned} p(t) &= p(0)\exp(-\gamma_p t) + k_p \int_0^t \exp(-\gamma_p(t-s))ds \\ &+ \int_0^t \exp(-\gamma_p(t-s))\sqrt{k_p + \gamma_p \phi(s)}dW. \end{aligned} \tag{5.9}$$

Basic properties of Itô integral imply

$$\begin{aligned} \langle p(t) \rangle &= \langle p(0) \rangle \exp(-\gamma_p t) + k_p \int_0^t \exp(-\gamma_p(t-s))ds \\ &= \langle p(0) \rangle \exp(-\gamma_p t) + \frac{k_p}{\gamma_p}(1 - exp(-\gamma_p t)), \end{aligned} \tag{5.10}$$

where $\langle X\rangle$ denotes expected value of a random variable X.

In order to find the covariance matrix $\Sigma(\Theta) = \{\sigma^2_{i,j}\}_{i,j=0,\dots,n}$ we calculate the autocorrelation function $\langle(p(t_1) - \langle p(t_1)\rangle)(p(t_2) - \langle p(t_2)\rangle)\rangle$. Formulae (5.9), (5.8) and basic properties of Itô integral give for $t_1 \leq t_2$

$$\begin{aligned}&\langle(p(t_1) - \langle p(t_1)\rangle)(p(t_2) - \langle p(t_2)\rangle)\rangle = \\ &\langle(p(0) - \langle p(0)\rangle)^2\rangle \exp(-\gamma_p(t_1 + t_2)) \\ &+ \exp(-\gamma_p(t_1 + t_2)) \int_0^{t_1} \exp(2\gamma_p s)(k_p + \gamma_p\phi(s))ds.\end{aligned} \tag{5.11}$$

Therefore

$$\sigma^2_{i,j} = \sigma^2_{p(0)} \exp(-\gamma_p(t_i + t_j)) + \exp(-\gamma_p(t_j + t_i)) \int_0^{t_i} \exp(2\gamma_p s)(k_p + \gamma_p\phi(s))ds, \tag{5.12}$$

where $\sigma^2_{p(0)} = \langle(p(0) - \langle p(0)\rangle)^2\rangle$. To simplify formula (5.12) we use eq. (5.8) and perform further integration

$$\begin{aligned}&\int_0^{t_i} \exp(2\gamma_p s)(k_p + \gamma_p\phi_p(s))ds = \\ &\frac{k_p}{\gamma_p}(\exp(2\gamma_p t_i) - 1) + \left(\phi_p(0) - \frac{k_p}{\gamma_p}\right)(\exp(\gamma_p t_i) - 1).\end{aligned} \tag{5.13}$$

Finally we obtain

$$\begin{aligned}&\sigma^2_{i,j} = \\ &\exp(-\gamma_p(t_i + t_j))\left(\sigma^2_{p(0)} + \frac{k_p}{\gamma_p}(\exp(2\gamma_p t_i) - 1) + \left(\phi_p(0) - \frac{k_p}{\gamma_p}\right)(\exp(\gamma_p t_i) - 1)\right).\end{aligned} \tag{5.14}$$

As the history of a cell before time 0 is unknown we treat $p(0)$ as a parameter of the model and assume that $p(0) = \phi_p(0) = \langle p(0)\rangle$ and $\sigma^2_{p(0)} = 0$.

5.6.2 Cycloheximide experiment

Cycloheximide is an inhibitor of protein biosynthesis in eukaryotic organisms. It is widely used to determine degradation rates of proteins. In the experiment GH3 rat pituitary cells stably transfected with 5kb human prolactin promoter estabilised EGFP reporter construct (hPRL-d2EGFP) were seeded onto 35 mm glass coverslip-based dishes (IWAKI, Japan) and cultured in 10% FCS for 24h prior to imaging. Cells were transferred to the stage of a Zeiss Axiovert 200 equipped with an XL incubator (maintained at 37C, 5% CO2, in humid conditions) and images were obtained using a Fluar x20, 0.75 numerical aperture (Zeiss), air objective. Excitation of d2EGFP was performed using an Argon ion laser at 488nm. Emitted light was captured through a 505-550 nm bandpass filter from a 545 nm dichroic mirror. Images were captured every 5.4 min. 5 μM forskolin and 0.5 μM BayK 8644 was added directly to the dish for 6h followed by the addition of 10μg/ml cyclohexamide to inhibit translation. Data was captured and analysed using LSM510 software with consecutive autofocus. Analysis was performed using Kinetic Imaging software AQM6. Regions of interest were drawn around each single cell and mean intensity data was collected over 10h.

Chapter 6

Discussion

6.1 Summary

In this thesis we have presented a set of statistical methods for the inference of biochemical kinetic parameters. We have demonstrated the use of well established methods using experimental data as well as developed new methods testing them using data generated in computer simulations as well experimental data.

In Chapter 2 we used methods based on ordinary differential equations and stochastic differential equations to infer the kinetic rates of three different genes. A deterministic model was applied to the averaged measurements of Arabidopsis thaliana CAB2 protein and mRNA. A stochastic model was used for single rat pituitary cell prolactin promoter fluorescence measurements. For Arabidopsis thaliana CCA1 clock gene data both deterministic and stochastic approach were used. We discuss the advantages and limitations of fitting either stochastic or ordinary differential equations and address the problem of parameter identifiability when model variables are unobserved. Our results demonstrate that MCMC methods for ODEs and SDEs provide practical algorithms for the reconstruction transcription profiles whilst estimating some of the kinetic rates involved. As the single-cell dynamics is naturally stochastic SDEs provide the superior theoretical model. However the mean ODE approach can be useful as a vehicle for estimation when the data are not fully compatible with the SDE assumptions.

The SDEs used for inference in Chapter 2 were derived from the chemical master equation using the diffusion approximation. Transition densities for this type of SDE are usually unknown, therefore we used data augmentation techniques to obtain an approximation of transition densities. Our experience from work with diffusion approximation based methods is that their implementation is challenging especially for data that are sparsely sampled in time because the need for imputation of unobserved time points leads to a very high dimensionality of the posterior distribution. This usually results in highly autocorrelated traces affecting the speed of convergence of the Markov chain.

The aim of the Chapter 3 is to introduce the linear noise approximation as a useful and novel approach to the inference of biochemical kinetics parameters. Its major advantage is that an explicit formula of the likelihood can be derived even for systems with unobserved variables and data observed with measurement error. In contrast to the diffusion approximation based methods the computationally costly methods of data augmentation to approximate the transition densities and to integrate out unobserved model variables are not necessary. Our method considerably reduces the dimension of the posterior distribution to the number of unknown parameters of a model only and is independent of the number of unobserved components. Applicability of the LNA is demonstrated using a model of single gene expression together with experimental and simulated data.

In Chapter 4 we extend the standard model of gene expression used in Chapters 2 and 3 in order to propose a reliable framework for the interpretation of fluorescent reporter gene, single-cell, steady state and out-of-steady-state data. The new model, firstly, incorporates extrinsic noise, modelled as stochastic fluctuations in the transcription rate. Secondly, a maturation process was introduced to take account of variability generated by the stochastic maturation of a fluorescent reporter protein. Therefore, apart from stochasticity resulting from randomness of transcription, translation and degradation events, our approach accounts for extrinsic noise as well as variability arising from the kinetics of fluorescent protein maturation. Our method allows to infer properties of extrinsic noise such as variance and half-life from single reporter gene time-lapse data, whereas other established methods require experiments employing two reporter genes.

In Chapter 5 we propose a Bayesian hierarchical model for estimation of fluorescent reporter protein degradation rates. In contrast to standard methods the hierarchical approach allows us to quantify the variability of kinetic parameters between cells. Instead of a single average value of the degradation rate the distribution of the degradation rates in a population of cells can be inferred. Our model also accounts for stochasticity in degradation and translation events and measurement error.

6.2 Relevance

The functioning of living cells depends on the execution of a genetic program determined by a complex networks of genes. Reliability of this program requires steadfast signal transduction from one gene to another [Tkačik et al., 2008]. This process is perturbed by spontaneous fluctuations arising from gene expression [Raj and van Oudenaarden, 2008]. Reverse engineering of gene regulatory networks as well as understanding their dynamical properties is a fundamental problem in current biology. In vivo measurement technologies provide more and more data that refer to the functioning of gene regulatory networks [Megason and Fraser, 2007]. Stochastic mathematical modelling is necessary in providing their systematic description [Wilkinson, 2009]. In order to explain experimentally observed behaviour in a mathematical model statistical methods are needed for the the estimation of its parameters [Jaqaman and Danuser, 2006]. This is important because the picture obtained using inference techniques integrates the theoretical knowledge resulting from mathematical modelling with experimental data and therefore offers insight that is otherwise unavailable.

The methodological concepts presented in this thesis seem to provide a reliable and convenient framework that addresses the problem of integrating single-cell reporter gene data with stochastic mathematical models. The main advantage of the proposed approach based on the linear noise approximation is that it allows for unobserved variables which is a common situation in current experimental settings. In addition, a Bayesian framework allows for the portability of results across different studies, providing a natural solution to parameter identifiability problems. These advantages allow us to construct a model that constitutes a general framework for interpretation of fluorescent time-lapse data and can extract richer, more insightful description of gene expression process from experimental data than currently available methods. As fluorescent reporters are primary tools to observe the dynamics of gene expression reliable interpretation of this type of data is of special significance.

6.3 Future extensions

The explicit formula for the likelihood of observed steady-state and out-of-steady-state data allows us to ask questions that have only been vaguely addressed before. For instance, the following problems can be approached:

1. How do the dynamical properties of noise differ between genes?

The method proposed in Chapter 4 can be used to estimate kinetic parameters of gene expression as well as variance/strenght and half-life of extrinsic noise from single fluorescent reporter gene time-lapsed data. These parameters provide a detailed description of noise. Their comparison would provide a new insight into differences in stochastic regulatory properties between different genes.

2. Can we assign variability in protein concentrations to individual reactions using single reporter gene data?

The total observed variability of the fluorescent proteins arises from ten major sources: extrinsic noise, transcription, translation, protein folding, protein oxidation, mRNA and protein degradation and measurement error. Schematic description of a model of expression of a fluorescent protein with depicted noise sources is presented in Figure 6.1. Existing methods employ a double reporter gene construct to distinguish between intrinsic and extrinsic noise. It can be shown (proof not presented) that the variance of the matured proteins $\sigma^2_{p_m}(t) = \langle (p_m(t) - \langle p_m(t) \rangle)^2 \rangle$ can be decomposed as follows $\sigma^2_{p_m}(t) = \sigma^2_1(t) + ... + \sigma^2_9(t)$, where $\sigma^2_i(t)$ is the contribution of ith source (reaction). More interestingly, the contributions $\sigma^2_i(t)$ for $i = 1, ..., 9$ can be inferred from single reporter gene time-lapsed data using the method proposed in Chapter 4. Inference of the individual contribution will provide a very precise description of origins of the observed fluctuations.

3. Can we detect origins of extrinsic variability?

Although the definition of extrinsic noise is precise, the source of extrinsic fluctuations is mostly unknown. The determination of its origins will help us to understand how stochastic fluctuations influence gene expression in complex information processing networks. The availability of an explicit formula for the likelihood of observed data enables us to employ a statistical model selection framework for testing hypotheses about the origins of extrinsic fluctuations. For instance, it would be possible to test the hypothesis whether extrinsic noise occurs as fluctuations in transcription or in translation rate.

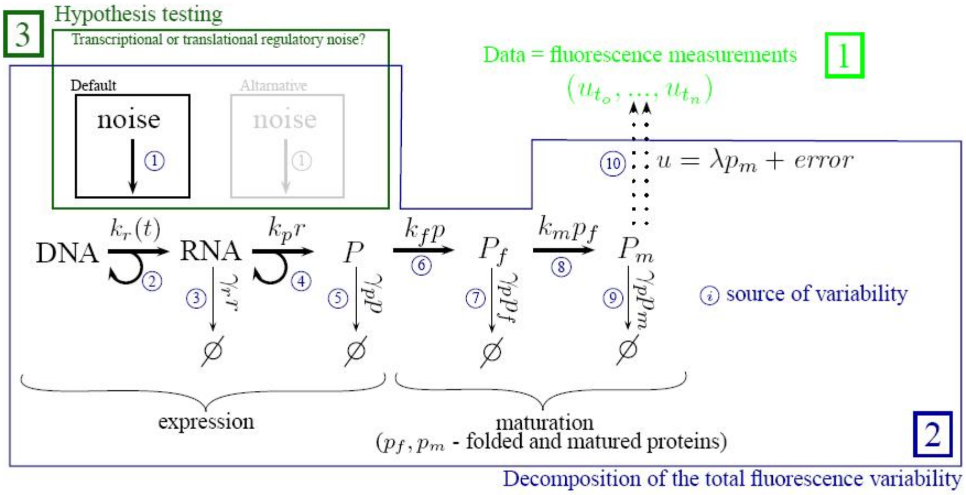
3
Hypothesis testing
Transcriptional or translational regulatory noise?
Default
noise
Altarnative
noise
1
Data = fluorescence measurements
$(u_{t_o}, ..., u_{t_n})$
$u = \lambda p_m + error$
DNA
$k_r(t)$
RNA
$k_p r$
P
$k_f p$
P_f
$k_m p_f$
P_m
$\gamma_r r$
$\gamma_p P$
$\gamma_p p_f$
$\gamma_p p_m$
$\emptyset$
source of variability
expression
maturation
(p_f, p_m - folded and matured proteins)
2
Decomposition of the total fluorescence variability

Figure 6.1: Further work: systems analysis of gene expression data at the single cell level. The explicit formula for the likelihood function allows for detailed analysis of fluorescent reporter gene data. The parameters of the model of single expression can be estimated and compared between different genes (*1*); the total observed fluorescence variability can be decomposed into the sum of contributions resulting from different sources (*2*); the statistical model selection methods can be employed to detect the origins of extrinsic variability (*3*)

Bibliography

S. O. Aase and P. Ruoff. Semi-algebraic optimization of temperature compensation in a general switch-type negative feedback model of circadian clocks. *Journal of Mathematical Biology*, 56(3):279–292, 2008.

A. Arkin, J. Ross, and H.H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics*, 149(4): 1633–1648, 1998.

L. Arnold. *Stochastic differential equations: theory and applications.* Wiley-Interscience, 1974.

L. Ashall, C.A. Horton, D.E. Nelson, P. Paszek, C.V. Harper, K. Sillitoe, S. Ryan, D.G. Spiller, J.F. Unitt, D.S. Broomhead, et al. Pulsatile Stimulation Determines Timing and Specificity of NF-{kappa} B-Dependent Transcription. *Science*, 324 (5924):242, 2009.

DW Austin, MS Allen, JM McCollum, RD Dar, JR Wilgus, GS Sayler, NF Samatova, CD Cox, and ML Simpson. Gene network shaping of inherent noise spectra. *Nature*, 439:608–611, 2006.

M. Barenco, D. Tomescu, D. Brewer, R. Callard, J. Stark, and M. Hubank. Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biology*, 7(3):R25, 2006.

A. Becskei, B. Séraphin, and L. Serrano. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *The EMBO journal*, 20(10):2528, 2001.

A. Beskos, O. Papaspiliopoulos, G. O. Roberts, and P. Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):333–382, 2006. doi: 10.1111/j.1467-9868.2006.00552.x.

W. Bialek and S. Setayeshgar. Physical limits to biochemical signaling. *Proceedings of the National Academy of Sciences*, 102(29):10040–10045, 2005.

W.J. Blake, M. KAern, C.R. Cantor, and JJ Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003.

W.J. Blake, G. Balázsi, M.A. Kohanski, F.J. Isaacs, K.F. Murphy, Y. Kuang, C.R. Cantor, D.R. Walt, and J.J. Collins. Phenotypic consequences of promoter-mediated transcriptional noise. *Molecular cell*, 24(6):853–865, 2006.

R.J. Boys, D.J. Wilkinson, and T.B.L. Kirkwood. Bayesian inference for a discretely observed stochastic kinetic model. *Statistics and Computing*, 18(2):125–135, 2008.

P.J. Brockwell and R.A. Davis. *Introduction to time series and forecasting*. Springer New York, 2002.

K.S. Brown and J.P. Sethna. Statistical mechanical approaches to models with many poorly known parameters. *PHYSICAL REVIEW E Phys Rev E*, 68:021904, 2003.

Y. Cao, D. Gillespie, and L. Petzold. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *Journal of Computational Physics*, 206(2):395–411, 2005.

J. R. Chabot, J. M. Pedraza, P. Luitel, and A. van Oudenaarden. Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock. *Nature*, 450:1249–1252, 2007. doi: doi:10.1038/nature06395.

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

S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49:327–336, 1995.

J. Chubb. Faculty of 1000 Biology: evaluations for Dong GQ and McMillen DR Phys Rev E Stat Nonlin Soft Matter Phys 2008 Feb 77 (2 Pt 1) :021908 . URL `http://www.f1000biology.com/article/id/1119120/evaluation`.

A. Cornish-Bowden. *Fundamentals of enzyme kinetics*. Portland Press London, 1995.

C.G. Dong, L. Jakobowski, and D.R. McMillen. Systematic Reduction of a Stochastic Signalling Cascade Model. *Journal of Biological Physics*, 32(2):173–176, 2006.

G.Q. Dong and D.R. McMillen. Effects of protein maturation on the noise in gene expression. *Physical Review E*, 77(2):21908, 2008.

M.R. Doyle, S.J. Davis, R.M. Bastow, H.G. McWatters, L. Kozma-Bognar, F. Nagy, A.J. Millar, and R.M. Amasino. The elf4 gene controls circadian rhythms and flowering time in arabidopsis thaliana. *Nature*, 419:74–77, 2002.

Y. Dublanche, K. Michalodimitrakis, N. Kümmerer, M. Foglierini, and L. Serrano. Noise in transcription negative feedback loops: simulation and experimental analysis. *Molecular Systems Biology*, 2(1), 2006.

Gallant A.R. Durham G. B. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business and Economic Statistics*, 20:297–338(42), 2002.

M. Ehrenberg, J. Elf, E. Aurell, R. Sandberg, and J. Tegner. Systems Biology Is Taking Off. *Genome Res.*, 13(11):2377–2380, 2003. doi: 10.1101/gr.1763203.

O. Elerian, S. Chib, and N. Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–993, 2001. doi: 10.1111/1468-0262.00226.

Johan Elf and Mans Ehrenberg. Fast Evaluation of Fluctuations in Biochemical Networks With the Linear Noise Approximation. *Genome Res.*, 13(11):2475–2484, 2003. doi: 10.1101/gr.1196503.

M. B. Elowitz, A. J. Levine, E. D. Siggia, and P. S. Swain. Stochastic Gene Expression in a Single Cell. *Science*, 297(5584):1183–1186, 2002a. doi: 10.1126/science.1070919.

M.B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, 2000.

M.B. Elowitz, A.J. Levine, E.D. Siggia, and P.S. Swain. Stochastic gene expression in a single cell, 2002b.

B. Eraker. MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19:177–191, 2001.

W.R. Esposito and C.A. Floudas. Global optimization for the parameter estimation of differential-algebraic systems. *Industrial and Engineering Chemistry Research*, 39(5):1291–1310, 2000. ISSN 0888-5885.

L.C. Evans. *Partial differential equations*. American Mathematical Society, 1998.

L. Ferm, Lötstedt P., and A. Hellander. A Hierarchy of Approximations of the Master Equation Scaled by a Size Parameter. *Journal of Scientific Computing*, 34(2):127–151, 2007. doi: 10.1007/s10915-007-9179-z.

B. Finkenstadt, E.A. Heron, M. Komorowski, K. Edwards, S. Tang, C.V. Harper, J.R.E. Davis, M.R.H. White, A.J. Millar, and D.A. Rand. Reconstruction of transcriptional dynamics from gene reporter data using differential equations. *Bioinformatics*, 24(24):2901, 2008.

N. Friedman, L. Cai, and X.S. Xie. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. *Physical Review Letters*, 97(16):168302, 2006.

D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference, 2nd ed.* Chapman & Hall/CRC, 2006.

C. Gardiner. *Handbook of stochastic methods*. Springer, 1985.

D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977. doi: 10.1021/j100540a008.

D. T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A*, 188(1-3):404–425, 1992a. doi: 10.1016/0378-4371(92)90283-V.

D.T. Gillespie. *Markov processes: An introduction for physical scientists*. Academic Press, 1992b.

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

D.T. Gillespie and L.R. Petzold. Improved leap-size selection for accelerated stochastic simulation. *The Journal of Chemical Physics*, 119:8229, 2003.

A. Goldbeter. Computational approaches to cellular rhythms. *Nature*, 420(6912): 238–245, 2002.

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61(3):781–788, 2005. doi: 10.1111/j.1541-0420.2005.00345.x.

A. Gordon, A. Colman-Lerner, T.E. Chin, K.R. Benjamin, R.C. Yu, and R. Brent. Single-cell quantification of molecules and rates using open-source microscope-based cytometry. *Nature methods*, 4(2):175–182, 2007.

P.D. Gould, J.C. Locke, C. Larue, M.M. Southern, S.J. Davis, S. Hanano, R. Moyle, R. Milich, J. Putterill, A.J. Millar, and A. Hall. The molecular basis of temperature compensation in the arabidopsis circadian clock. *Plant Cell*, 18:1177–1187, 2006.

J. Gunawardena. Models in systems biology: the parameter problem and the meanings of robustness. *Elements of Computational Systems Biology. John Wiley and Sons, New York*, 2009.

P. Guptasarma. Does replication-induced transcription regulate synthesis of the myriad low copy number proteins of Escherichia coli? *Bioessays*, 17(11):987–97, 1995.

AC Guyton, TG Coleman, and HJ Granger. Circulation: overall regulation. *Annual Review of Physiology*, 34(1):13–44, 1972.

DA Henderson, RJ Boys, CJ Proctor, and DJ Wilkinson. Linking systems biology models to data: a stochastic kinetic model of p53 oscillations, 2009.

S. Hengl, C. Kreutz, J. Timmer, and T. Maiwald. Data-based identifiability analysis of non-linear dynamical models. *Bioinformatics*, 23(19):2612–2618, 2007.

E. A. Heron, B. Finkenstadt, and D. A. Rand. Bayesian inference for dynamic transcriptional regulation; the Hes1 system as a case study. *Bioinformatics*, 23 (19):2596–2603, 2007. doi: 10.1093/bioinformatics/btm367.

D. J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM*, 43(3):525–546, 2001.

A.L. Hodgkin and A.F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Bulletin of Mathematical Biology*, 52(1):25–71, 1990.

A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The Ikappa B-NF-kappa B Signaling Module: Temporal Control and Selective Gene Activation. *Science*, 298(5596):1241–1245, 2002. doi: 10.1126/science.1071914.

S. Hooshangi, S. Thiberge, and R. Weiss. Ultrasensitivity and noise propagation in a synthetic transcriptional cascade. *Proceedings of the National Academy of Sciences*, 102(10):3581–3586, 2005.

S. Huang. Back to the biology in systems biology: What can we learn from biomolecular networks? *Briefings in functional genomics and proteomics*, 2(4): 279–297, 2004.

M.A.J. Iafolla and D.R. McMillen. Extracting Biochemical Parameters for Cellular Modeling: A Mean-Field Approach. *JOURNAL OF PHYSICAL CHEMISTRY B*, 110(43):22019, 2006.

K. Jaqaman and G. Danuser. Linking data to models: data regression. *Nature Reviews Molecular Cell Biology*, 7(11):813–819, 2006.

M. H. Jensen, K. Sneppen, and G. Tiana. Sustained oscillations and time delays in gene expression of protein $Hes1$. *Febs Letters*, 541(1-3):176–177, 2003.

H. Kacser, JA Burns, and DA Fell. The control of flux. *Biochemical Society Transactions*, 23(2):341–366, 1995.

S. Kalir, J. McClure, K. Pabbaraju, C. Southward, M. Ronen, S. Leibler, M. G. Surette, and U. Alon. Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria. *Science*, 292(5524):2080–2083, 2001. doi: 10.1126/science.1058758.

Joel Keizer. *Statistical Thermodynamics of Nonequilibrium Processes*. Springer, 1987.

T.B. Kepler and T.C. Elston. Stochasticity in Transcriptional Regulation: Origins, Consequences, and Mathematical Representations. *Biophysical Journal*, 81(6): 3116–3136, 2001.

T.R. Kiehl, R.M. Mattheyses, and M.K. Simmons. Hybrid simulation of cellular behavior, 2004.

S. Kim, N. Shephard, and S. Chib. Stochastic volatility: Likelihood inference and comparison with ARCH models. *Review of Economic Studies*, 65:361–393, 1998.

P. E. Kloeden and Platen E. *Numerical Solution of Stochastic differential equations*. Springer, 1999.

MS 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, 1990.

M. Komorowski, B. Finkenstadt, C.V. Harper, and Rand D. Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *Submitted.*, 2009a. URL `http://arxiv.org/PS_cache/arxiv/pdf/0907/0907.0759v1.pdf`.

M. Komorowski, J. Miekisz, and A. Kierzek. Translational Repression Contributes Greater Noise to Gene Expression than Transcriptional Repression. *Biophysical Journal*, 96(2), 2009b.

Thomas G. Kurtz. The Relationship between Stochastic and Deterministic Models for Chemical Reactions. *The Journal of Chemical Physics*, 57(7):2976–2978, 1972.

T. Lipniacki, P. Paszek, A.R. Brasier, B. Luxon, and M. Kimmel. Mathematical model of NF-κB regulatory module. *Journal of theoretical biology*, 228(2):195–215, 2004.

T. Lipniacki, P. Paszek, A.R. Brasier, B.A. Luxon, and M. Kimmel. Stochastic regulation in early immune response. *Biophysical journal*, 90(3):725–742, 2006.

J. C. W. Locke, A. J. Millar, and M. S. Turner. Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *J Theor Biol*, 234:383–393, 2005a.

J. C. W. Locke, M. M. Southern, L. Kozma-Bognar, V. Hibberd, P. E. Brown, M. S. Turner, and A. J. Millar. Extension of a genetic network model by iterative experimentation and mathematical analysis. *Mol Syst Biol*, 1:E1–E9, 2005b.

R. Losick and C. Desplan. Stochasticity and cell fate. *Science*, 320(5872):65, 2008.

V. Maple. Waterloo Maple Inc. *Waterloo, Canada*.

A. Martinez Arias and P. Hayward. Filtering transcriptional noise during development: concepts and mechanisms. *Nature reviews. Genetics*, 7(1):34–44, 2006.

H.H. McAdams and A. Arkin. Stochastic mechanisms in gene expression, 1997.

D.A. McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, pages 413–478, 1967.

S.G. Megason and S.E. Fraser. Imaging in systems biology. *Cell*, 130(5):784–795, 2007.

P Mendes and D Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14 (10):869–883, 1998. doi: 10.1093/bioinformatics/14.10.869.

A. J. Millar and S. A. Kay. Integration of circadian and phototransduction pathways in the network controlling cab gene transcription in arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America*, 93:15491–15496, 1996.

A. J. Millar, I. Carre, C. Strayer, N. Chua, and S. Kay. Circadian clock mutants in arabidopsis identified by luciferase imaging. *Science*, 267:1161–1163, 1995.

C. G. Moles, P. Mendes, and J. R. Banga. Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods. *Genome Res.*, 13 (11):2467–2474, 2003. doi: 10.1101/gr.1262503.

T. Nagai, K. Ibata, E.S. Park, M. Kubota, K. Mikoshiba, and A. Miyawaki. A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. *Nature biotechnology*, 20(1):87–90, 2002.

D. E. Nelson, A. E. C. Ihekwaba, M. Elliott, J. R. Johnson, C. A. Gibney, and et al. Oscillations in NF-kappaB Signaling Control the Dynamics of Gene Expression. *Science*, 306(5696):704–708, 2004. doi: 10.1126/science.1099962.

A. Novick and M. Weiner. Enzyme induction as an all-or-none phenomenon. *Proceedings of the National Academy of Sciences*, 43(7):553–566, 1957.

B. Oksendal. *Stochastic differential equations (3rd ed.): an introduction with applications*. Springer, 1992.

E. M. Ozbudak, M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden. Regulation of noise in the expression of a single gene. *Nature Genetics*, 31:69 – 73, 2002. doi: 10.1038/ng869.

J. Paulsson. Models of stochastic gene expression. *Physics of life reviews*, 2(2): 157–175, 2005.

J. Paulsson. Summing up the noise in gene networks. *Nature*, 427:415–418, 2006. doi: 10.1038/nature02257.

A. Pedersen. A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, pages 55–71, 1995.

J.M. Pedraza and A. van Oudenaarden. Noise Propagation in Gene Networks, 2005.

M. Ptashne. On the use of the word'epigenetic'. *Current Biology*, 17(7):233, 2007.

J. Puchałka and A.M. Kierzek. Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophysical journal*, 86(3):1357–1372, 2004.

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

J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007. doi: 10.1111/j.1467-9868.2007.00610.x.

TA Rapoport, R. Heinrich, G. Jacobasch, and S. Rapoport. A linear steady-state treatment of enzymatic chains. A mathematical model of glycolysis of human erythrocytes. *European journal of biochemistry/FEBS*, 42(1):107, 1974.

J. M. Raser and E. K. O'Shea. Noise in Gene Expression: Origins, Consequences, and Control. *Science*, 309(5743):2010–2013, 2005. doi: 10.1126/science.1105891.

J. Rausenberger and M. Kollmann. Quantifying Origins of Cell-to-Cell Variations in Gene Expression. *Biophysical Journal*, 95(10):4523–4528, 2008.

S. Reinker, R.M. Altman, and J. Timmer. Parameter estimation in stochastic biochemical reactions. *Systems Biology, IEE Proceedings*, 153(4):168–178, 2006. ISSN 1741-2471. doi: 10.1049/ip-syb:20050105.

M. Ronen, R. Rosenberg, B. I. Shraiman, and U. Alon. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10555–10560, 2002.

N. Rosenfeld, J.W. Young, U. Alon, P.S. Swain, and M.B. Elowitz. Gene regulation at the single-cell level. *Science*, 307(5717):1962–1965, 2005.

M.A. Savageau. *Biochemical systems analysis: a study of function and design in molecular biology*. Addison Wesley Publishing Company, 1976.

M. Schena, D. Shalon, R.W. Davis, and P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270 (5235):467, 1995.

V. Shahrezaei, J.F. Ollivier, and P.S. Swain. Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4(196), 2008.

O. Shimomura, F.H. Johnson, Y. Saiga, et al. Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan, Aequorea. *Journal of Cellular and Comparative Physiology*, 59(3):223–239, 1962.

A. Sigal, R. Milo, A. Cohen, N. Geva-Zatorsky, Y. Klein, Y. Liron, N. Rosenfeld, T. Danon, N. Perzov, and U. Alon. Variability and memory of protein levels in human cells. *NATURE-LONDON-*, 444(7119):643, 2006.

S.D. Silvey. *Statistical inference*. Chapman & Hall, 1975.

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

P.S. Swain, M.B. Elowitz, and E.D. Siggia. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences*, 99(20):12795–12800, 2002b.

M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540, 1987.

M. Thattai and A. van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, page 151588598, 2001. doi: 10.1073/pnas.151588598.

T. Tian, S. Xu, J. Gao, and K. Burrage. Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics*, 23(1):84, 2007.

G. Tkačik, C.G. Callan, and W. Bialek. Information flow and optimization in transcriptional regulation. *Proceedings of the National Academy of Sciences*, 105(34):12265, 2008.

R. Tomioka, H. Kimurab, T. J. Kobayashib, and K. Aihara. Multivariate analysis of noise in genetic regulatory networks. *Journal of Theoretical Biology*, 229(4): 501–521, 2004.

R.Y. Tsien. THE GREEN FLUORESCENT PROTEIN. *Annual Reviews in Biochemistry*, 67(1):509–544, 1998.

N.G. Van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 2006.

V. Vyshemirsky and M.A. Girolami. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833, 2008.

X. Wang, B. Errede, and T.C. Elston. Mathematical Analysis and Quantification of Fluorescent Proteins as Transcriptional Reporters. *Biophysical Journal*, 94 (6):2017, 2008.

Z. Wang and E. M. Tobin. Constitutive expression of the circadian clock associated 1 (cca1) gene disrupts circadian rhythms and suppresses its own expression. *Cell*, 93(7):1207–1217, 1998.

D.J. Wilkinson. *Stochastic modelling for systems biology*. Chapman & Hall/CRC, 2006.

D.J. Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.

J.Q. Wu and T. D. Pollard. Counting Cytokinesis Proteins Globally and Locally in Fission Yeast. *Science*, 310(5746):310–314, 2005. doi: 10.1126/science.1113230.

S. X. Xie, P. J. Choi, G. W. Li, N. K. Lee, and G. Lia. Single-molecule approach to molecular biology in living bacterial cells. *Annual Review of Biophysics*, 37 (1):417–444, 2008.

E. Yakir, D. Hilman, M. Hassidim, and R. Green. Circadian clock associated1 transcript stability and the entrainment of the circadian clock in arabidopsis. *Plant Physiology*, 145:925–932, 2007.

E. Ziv, I. Nemenman, and C.H. Wiggins. Optimal signal processing in small stochastic biochemical networks. *PLoS One*, 2(10), 2007.

D. Zwillinger. *Handbook of Differential Equations*. San Diego, 1989.