Probabilistic Reasoning and Inference for Systems Biology

by

Vladislav Vyshemirsky

A dissertation submitted to

The Department of Computing Science
of
The University of Glasgow

for the degree of

Doctor of Philosophy

July 2007

Abstract

One of the important challenges in Systems Biology is reasoning and performing hypotheses testing in uncertain conditions, when available knowledge may be incomplete and the experimental data may contain substantial noise.

In this thesis we develop methods of probabilistic reasoning and inference that operate consistently within an environment of uncertain knowledge and data. Mechanistic mathematical models are used to describe hypotheses about biological systems.

We consider both deductive model based reasoning and model inference from data. The main contributions are a novel modelling approach using continuous time Markov chains that enables deductive derivation of model behaviours and their properties, and the application of Bayesian inferential methods to solve the inverse problem of model inference and comparison, given uncertain knowledge and noisy data.

In the first part of the thesis, we consider both individual and population based techniques for modelling biochemical pathways using continuous time Markov chains, and demonstrate why the latter is the most appropriate. We illustrate a new approach, based on symbolic intervals of concentrations, with an example portion of the ERK signalling pathway. We demonstrate that the resulting model approximates the same dynamic system as traditionally defined using ordinary differential equations. The advantage of the new approach is quantitative logical analysis; we formulate a number of biologically significant queries in the temporal logic CSL and use probabilistic symbolic model checking to investigate their veracity.

In the second part of the thesis, we consider the inverse problem of model inference and testing of alternative hypotheses, when models are defined by non-linear ordinary differential equations and the experimental data is noisy and sparse. We compare and evaluate a number of statistical techniques, and implement an effective Bayesian inferential framework for systems biology based on Markov chain Monte Carlo methods and estimation of marginal likelihoods by annealing-melting integration. We illustrate the framework with two case studies, one of which involves an open problem concerning the mediation of ERK phosphorylation in the ERK pathway.

Computer Science is no more about computers than astronomy is about telescopes.

Edsger Dijkstra

Acknowledgements

I would like to thank everybody who provided me with any kind of support during my work. Especially:

My supervisor, Prof. Muffy Calder, who was always able to find some time for me despite her busy schedule; and without whom this work would have never been accomplished. Thank you, Muffy, for everything.

My second supervisor, Prof. Mark Girolami, whose scientific guidance made a vital contribution to this work.

Prof. David Gilbert, who leads the Bioinformatics Research Centre. Thanks for all the support and advice.

Prof. Alexander Letichevsky, who introduced me to the exciting quest of scientific research.

My teacher, Anatolij Abramovich, whose dedication and support I remember and appreciate. Rest in peace, Mr. Abramovich.

The Department of Trade and Industry for funding my work; and people of Scotland for being so friendly and breaking all my fears about being a foreigner.

My fellow researchers and my faithful friends for all the support.

My parents and family for their constant encouragement and support; and my sister, Victoria, for always thinking of me and how she can help me.

And most of all, Tanya, without whom I never would have made it.

Declaration

Some of the work reported in Chapter 3 has been published as a peer reviewed paper (Calder, Vyshemirsky, Gilbert and Orton 2006). The author's contribution to this paper was development of the PRISM models, theoretical derivation of the transition rates from kinetic constants, formulation and checking of the logical properties.

The biochemical experimental work referenced and described in Section 4.3.2 has been performed by Amélie Gormand under supervision of Prof. Walter Kolch and Prof. Miles D. Houslay.

All the work reported in this thesis has been performed by myself, unless specifically stated otherwise.

Vladislav Vyshemirsky

12 July 2007.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Overview

In this chapter we give a general background to the thesis and provide motivation for our work. We briefly describe the benefits of our work.

1.1 Challenges of Systems Biology

Systems Biology is a discipline of science which studies biological systems and their behaviour in an integrative way, using methods of mathematical modelling and analysis.

Burbeck and Jordan (2006) emphasise the central challenge of Systems Biology to assist in understanding how the various parts of a biological system fit and function together, and provide a list of major themes and focus areas which have emerged in Systems Biology:

- *Modelling.* Descriptive mathematical models are built to summarise and organise data.
- *Simulation.* Mathematical models are built to use researchers' knowledge about the parts of a biological system to better understand the implications of their interactions.
- *Automated analysis.* Automated analysis techniques to make inferences and predictions from accumulated knowledge and data.

- *Integration of computational biology and experimental biology.* Computational results are used to guide new experimentations followed by additional computational analysis and modelling of the new data.

All these themes play an important role in shaping Systems Biology, and a cyclic workflow from experimentations to computational analysis and back to experimentalists is a commonly accepted way to unravel the complexities of biological systems.

We consider Systems Biology as a methodological framework to operate on knowledge, experimental data, and hypotheses about biological systems, and assist in the development of a unified understanding of the involved biological processes.

In the scope of this thesis we concentrate our efforts on methods for computational analysis of biological data and knowledge in conditions of uncertainty. We focus on biochemical pathways and networks for practical applications for the methods considered in this thesis.

To establish the cyclic workflow between computational analysis and biochemical experimentation, we consider the methods of interest within the *scientific method* paradigm. The scientific method is a body of techniques for investigating phenomena and acquiring new knowledge, as well as for correcting and integrating with previous knowledge. It is based on gathering observable, empirical, measurable evidence, subject to principles of reasoning. Having the roots in ancient philosophy, it was preconditioned by the works by Sir Francis Bacon, René Descartes, and first formally introduced by Newton.

The scientific method suggests the following guideline for the research:

1. Formulate hypotheses;

2. Perform experiments and collect evidence;

3. Analyse the collected evidence and test the hypotheses formulated at the first step;

4. Interpret the result and draw conclusions that serve as a starting point for new hypotheses.

Sir Harold Jeffreys' book on Scientific Inference (Jeffreys 1937) argues for reasoning on probability inversion, the basis of what is known today as Bayesian

inference; this book emphasises the consistency of Bayesian inferential methodology with the scientific method, and demonstrates many examples from different areas of science on how this method is applied to provide sound and consistent evidential reasoning.

One of the important features in the life sciences is that there can be many competing hypotheses and corresponding models to explain some phenomenon. When evidence is collected all these competing hypotheses have to be tested. We propose to represent such competing hypotheses with mathematical models, and employ formal methods for the analysis and comparison of such models.

This thesis focuses on methods of formal reasoning and inference which allow knowledge and hypotheses to be operated in a natural and sound way according to established scientific method.

1.2 Models as Knowledge Representation

Mathematical models are used to represent knowledge and hypotheses about the structure and dynamics of a biological system. Generally speaking, there are two meta-approaches to biological systems modelling. The first one is data-driven modelling, and the second one is mechanistic modelling.

A data-driven modelling approach means that the main goal of the models is to mimic the observed behaviour of the model. Such models are often quantitative, and their main advantage is a precise simulation of the observed dynamic behaviour of the biological system. An example of a data-driven model is described in Example 1.1. There are a number of approaches to data-driven modelling ranging from Gaussian processes (see Rasmussen and Williams 2006) to S-systems (see Voit 2000). Data-driven models can sometimes be used to predict system behaviour in yet untested experimental conditions, however, their explanatory capabilities are limited due to a lack of structural information about the studied system.

Example 1.1 (Data-Driven Model of a Feedback Amplifier)

A feedback amplifier is a system which amplifies the input signal utilising a negative feedback to gain stability of amplification (see Figure 1.1). This structure can be found in many biological systems, this is possibly due to the evolutionary pressure to sustain a stable behaviour while amplifying some stimuli.

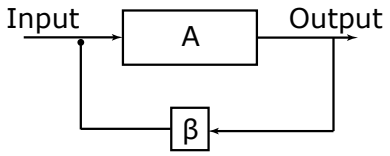


Figure 1.1: Feedback Amplifier

The basic idea of a feedback amplifier is to feed the output of the system back to the input of an amplification cascade with a negative feedback, thus inhibiting amplifier's input.

The data-driven model for the feedback amplifier does not consider the parts of the system separately, but rather defines the law by which the output of the system can be obtained from its input:

$$Output = \frac{A}{1 + \beta A} \cdot Input,$$

where A is the amplifier's gain, and β is the strength of the negative feedback.

Mechanistic models are used not only to mimic the observed behaviour but also to describe processes involved in producing such behaviour. A mechanistic model usually considers a system to be built from parts which interact with each other. In biological modelling, mechanistic models often employ the laws of molecular kinetics to describe the processes which contribute to the system behaviour. An example of a mechanistic model of a feedback amplifier is given in Example 1.2.

Example 1.2 (Mechanistic Model of a Feedback Amplifier)

Consider the same system as in Example 1.1. When modelling this system mechanistically, we consider a structural model of the feedback amplifier described using biochemical terms.

The structure depicted in Figure 1.2 gives more details about the processes involved in a feedback amplifier system. For example, one can see that the negative feedback is achieved through the competitive inhibition of the input I by the output O.

The system of differential equations which defines the dynamics of the mech-

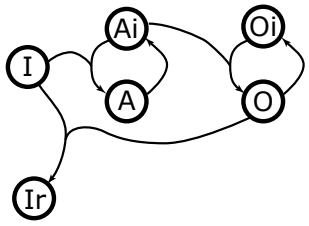


Figure 1.2: Structure of a Feedback Amplifier

anistic feedback amplifier model is the following:

$$\begin{cases} \dot{I} = -\dfrac{V_f \cdot I \cdot O}{K_f + I}, \\ \dot{I}_r = \dfrac{V_f \cdot I \cdot O}{K_f + I}, \\ \dot{A}_i = -\dfrac{V_a \cdot A_i \cdot I}{K_a + A_i} + \dfrac{V_d \cdot A}{K_d + A}, \\ \dot{A} = \dfrac{V_a \cdot A_i \cdot I}{K_a + A_i} - \dfrac{V_d \cdot A}{K_d + A}, \\ \dot{O}_i = -\dfrac{V_o \cdot O_i \cdot A}{K_o + O_i} + \dfrac{V_p \cdot O}{K_p + O}, \\ \dot{O} = \dfrac{V_o \cdot O_i \cdot A}{K_o + O_i} - \dfrac{V_p \cdot O}{K_p + O}. \end{cases}$$

In this thesis we will mainly use mechanistic biochemical models because these include more information about the structure of the modelled system, and consequently are more expressive when used to define the working hypotheses.

1.3 Reasoning and Inference

Deductive reasoning is a logical framework proposed by Aristotle in the 4th century B.C. which relies on the application of logical rules such as:

$$A \rightarrow B, A \vdash B, \text{also known as } \textit{Modus Ponens};$$
$$A \rightarrow B, \neg B \vdash \neg A, \text{also known as } \textit{Modus Tollens}.$$

For example, if we take $A \equiv$ *"It is raining"* and $B \equiv$ *"The sky is cloudy"* the above logical rules define the following deductions:

- The sky is cloudy when it is raining, and it is raining, therefore the sky is cloudy;
- The sky is cloudy when it is raining, and the sky is not cloudy, therefore it is not raining.

This kind of reasoning is the most desirable in practise, but unfortunately, we may not have all the information to apply deductive reasoning.

In the situations when the essential information is not available to perform deductive reasoning, *plausible reasoning* can be employed. Consider again the example above. Deductively, we cannot conclude that it is going to rain if the sky is cloudy, but we can say that it is more plausible that it will start to rain if the sky is cloudy. Plausible reasoning relies on a number of logical rules which allow one to make such conclusions, for instance:

$$A \to B, B \vdash A \text{ becomes more plausible,}$$
$$A \to B, \neg A \vdash B \text{ becomes less plausible.}$$

In the first case, observing the consequence B makes the reason A more plausible, when in the second case eliminating one of the reasons for B makes it less plausible.

We can assign a degree of plausibility to propositions or events using the Bayesian interpretation of probability. Bayesian theory defines the concept of probability as a degree to which a person believes in a proposition. This definition was first proposed by Ramsey (1931), and Bayesian theory was later developed on this foundation.

The name "Bayesian" comes from the use of Bayes' theorem which takes an important place in this theory. Bayes' theorem states how to update or revise beliefs in light of new evidence:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where

- $P(A)$ is the prior probability of A, which does not take into account any information about B;
- $P(A|B)$ is the posterior probability of A taking B into account;

- $P(B)$ is the prior probability of B;
- $P(B|A)$ is the conditional probability of B given A, or in other words the likelihood of B given A.

Automated inference and reasoning are useful not only for faster decision making, but also to investigate a large number of almost identical hypotheses. The goal of this thesis is to introduce automated approaches to assist reasoning about biological systems.

1.4 Thesis Statement

One of the important challenges in Systems Biology is to enable objective hypotheses testing and reasoning about models of biological systems. A major problem for such reasoning is uncertainty of knowledge and experimental observations.

We propose that model-based reasoning and inference based on probabilistic foundations are appropriate for tackling the problem of uncertainty representation. We demonstrate a novel modelling approach using continuous time Markov chains which enables quantitative model-based reasoning and develop an implementation of the Bayesian inferential framework for biological applications.

Both methodologies are applied to case studies in signal transduction pathways, thus demonstrating the feasibility of the proposed approaches.

1.5 Thesis Contribution

The problem of consistent reasoning and inference for Systems Biology in uncertain conditions is investigated.

Mechanistic mathematical models are used to describe the working hypotheses in biological research. A novel modelling approach using continuous time Markov chains (CTMCs) is proposed that enables deductive derivation of model behaviours and their properties; a Bayesian inferential methodology allows the inverse problem of model inference using uncertain knowledge and noisy data to be solved.

Alternative methods of model definition are considered in context of modelling using CTMCs, and the population-based approach is selected as the most

appropriate for modelling biochemical pathways. We demonstrate that the resulting models approximate the same dynamic system as traditionally defined using ordinary differential equations. Probabilistic symbolic model checking is then applied to derive model behaviours and investigate their properties.

A variety of algorithms which implement Bayesian inference methods are investigated and critically compared to solve the inverse problem of model inference and testing of alternative hypotheses. Markov chain Monte Carlo methods are selected as the "gold standard" as these are based on the least constrained foundations. Hypotheses testing using noisy experimental data is a challenging problem which requires the latest developments in applied statistics. We select path sampling methods to obtain stable results; this is demonstrated with several case studies in Systems Biology.

1.6 Outline of the Dissertation

The following diagram (see Figure 1.3) depicts the relationship between two major parts of this thesis.

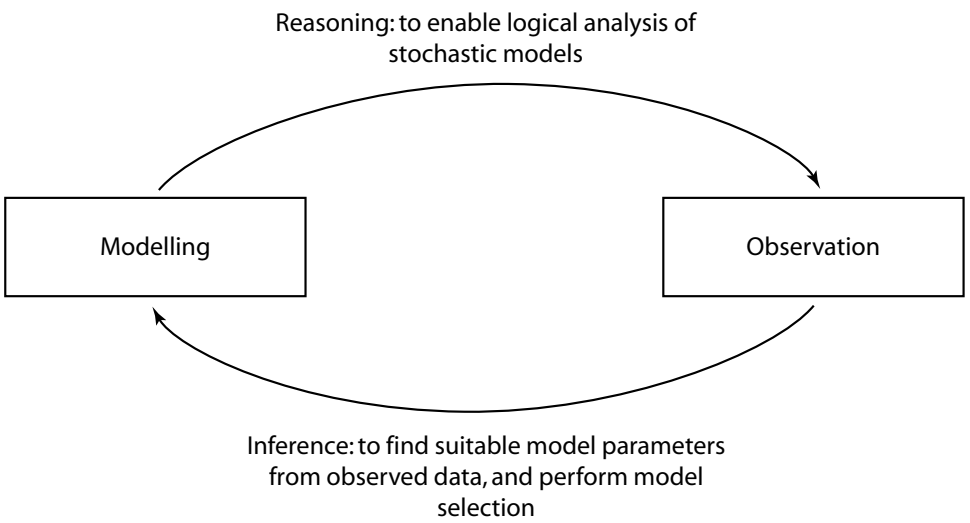


Figure 1.3: Schematic relationship between two parts of thesis.

From left to right, methods for probabilistic, logical reasoning about stochastic models allow us to analyse and map behaviours of such models to observed, experimental behaviour. In the opposite direction, inference based on the observed behaviour allows us to find suitable model parameters and perform model

selection.

In Chapter 2 we give an overview of the state of the art.

In Chapter 3 we describe the methods developed for quantitative reasoning about the dynamic behaviour of models of biological systems. We give examples of how this kind of reasoning can be applied to the analysis of signal transduction pathways.

In Chapter 4 we describe the implementation of Bayesian inference methodology in a biological context, considering the problems and solutions for model identification, hypotheses testing, and predictions. Section 4.3 contains two case studies which demonstrate how the proposed methods and algorithms can be employed to solve realistic research problems in Systems Biology.

We review our results and achievements and discuss ideas for future work in Chapter 5.

Summary

We have provided a brief outline of our work and its motivation.

Chapter 2

Related Work

Overview

In this chapter we review existing approaches in the areas of modelling, reasoning and inference for biological systems.

2.1 Ordinary Differential Equations

Modelling with differential equations is currently the most widely used approach in Systems Biology (see Voit 2000 de Jong 2003).

Definition 2.1: *An ordinary differential equation (ODE) is an equation which involves functions of only one independent variable, and one or more of its derivatives.*

For example,

$$\dot{y} = y$$

is an ordinary differential equation, where $\dot{y}$ denotes the first derivative of the function y by time. An alternative notation is $\frac{dy}{dt}$.

In the context of biological modelling, the independent variable is usually time, and dependent variables correspond to measurable quantities, e.g. protein concentrations.

The most common approach to building models of biological systems using ODEs relies on the use of kinetic laws, such as decay dynamics or binding dynamics.

Models defined with ODEs can be used to produce predictions of system behaviour by solving an *initial value problem*.

Definition 2.2 (Initial value problem)**:** *An initial value problem consists of a differential equation and the initial values which must belong to the solution.*

Example 2.1 (A Model of biochemical binding)

Consider a biochemical system of two proteins which can bind each other. We associate variables x, y, and z with the concentrations of the first protein, the second protein, and their complex respectively. Initially, at time $t = 0$, the concentrations of the proteins are $x|_{t=0}$, $y|_{t=0}$ and $z|_{t=0}$. At each instant of time, there is a chance that a molecule of the first protein will bind to a molecule of the second protein. The speed of protein concentration change depends on these concentrations at a given time. A system of ordinary differential equations which describes how the concentrations of the proteins change over time is

$$\begin{cases} \dot{x} = -k \cdot x \cdot y \\ \dot{y} = -k \cdot x \cdot y \\ \dot{z} = k \cdot x \cdot y \end{cases} \tag{2.1}$$

where $k \geq 0$ is a model parameter, usually called the binding rate. This type of chemical kinetics is called the mass action kinetic law (for more details see Stryer 1995).

Initial value problems are not always solvable analytically due to the structure of differential equations. Numerical solution methods can be employed in such situations. There is a wide range of differential equation solvers available at the moment (for an overview see Press et al. 2002). Different solvers are usually specialised for better performance on some classes of ordinary differential equations. For example, the Rosenbrock method is an implicit form of the Runge-Kutta solver that allows stiff[1] systems of ordinary differential equations

[1]Stiff systems of ordinary differential equations are those which cannot be solved effectively by basic adaptive step size solvers. This is mainly due to the fast changes in some dependent variables which require the step size of the solver to be reduced to very small values. Unfortunately, there is no formal definition of the stiff system of ODEs, and a system is usually declared stiff if the Runge-Kutta solver fails.

to be solved effectively in the cases when a Jacobean matrix of the system is available.

2.2 Petri Nets

Petri nets depict the structure of a distributed system as a directed bipartite graph with annotations. A Petri net has place nodes (depicted with circles in Figure 2.1), transition nodes (squares in Figure 2.1), and directed arcs connecting places with transitions.

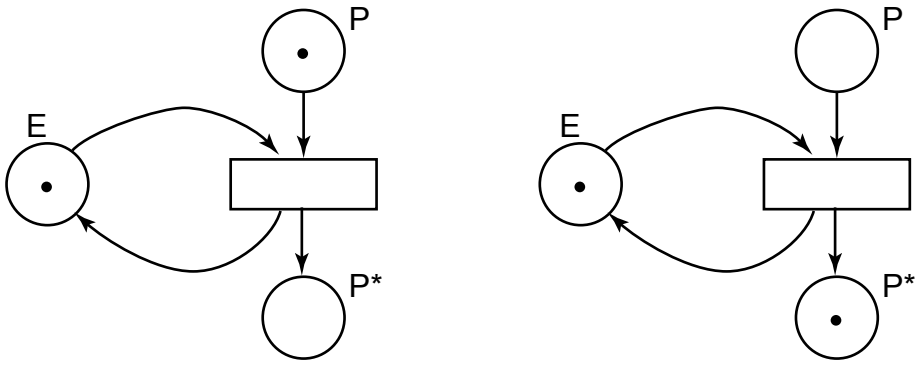
At any one time during a Petri net's execution, each place can hold zero or more tokens. Unlike more traditional data processing systems that can process only a single stream of incoming tokens, Petri net transitions can consume tokens from multiple input places, act on them, and output tokens to multiple output places. Before acting on input tokens, a transition waits until the required number of tokens appears in every one of its input places. Transitions act on input tokens by a process known as firing. When a transition fires, it consumes the tokens from its input places, performs some processing task, and places a specified number of tokens into each of its output places. It does this atomically, in one step. A set of tokens allocated at the places of Petri net is called *marking* of the Petri net.

Petri net representations can be used for qualitative modelling of biochemical networks (Goss and Peccoud 1998 Pinney et al. 2003 Sackmann et al. 2006 Heiner et al. 2004). In such applications species in the network are represented with places of the Petri net, and reactions are represented with transitions. The marking of this model with tokens represents the presence of some species in the system at different points of time.

Example 2.2 (Petri Net model of a biochemical reaction)

Protein P is activated (to become protein P) in presence of enzyme E. The reaction which converts P into P* is possible only when some E is available. The concentration of enzyme E will not be changed during the reaction.*

The initial state of the system is depicted in Figure 2.1(a). The initial marking (tokens in places P and E) corresponds to the presence of protein P and enzyme E in the system. As all input places for the transition have tokens, the transition can be fired and the Petri net will change its state to the one in Fig-



(a) before firing (b) after firing

Figure 2.1: Petri Net representation of an enzymatic activation

ure 2.1(b). In this state P is present in the system. At the same time, enzyme E is still present, as it was consumed and then reproduced during the transition.*

Above is an example of a catalytic reaction, which can usually be described with a pair of arcs in the Petri net, corresponding to consuming/producing the token at the catalyst place.

In general, inhibitory modifications are more difficult to formalise than catalytic ones. For example, a complementary place can be used to define a negated state of the system. Figure 2.2 illustrates Petri net for Kholodenko's model of the MAPK cascade (Kholodenko 2000), this model includes an inhibitory (negative) feedback loop. This system has been formalised using Petri nets (M. Heiner, private correspondence).

Petri nets enable a number of qualitative logical analysis techniques, such as automatic detection of loops in the system and checking of general topological properties. Additionally, some logical properties can be verified using temporal logic and model checking algorithms (see Clarke et al. 1999).

A number of extensions of this formalism have been created, such as Petri nets with inhibitory arcs, coloured Petri nets, stochastic Petri nets, timed Petri nets, and hybrid functional Petri nets. Using some of these extensions it is possible to simulate quantitative dynamics of the biochemical networks. We give a brief overview of a hybrid functional Petri nets approach, focusing on new capabilities provided and some limitations. Hybrid functional Petri nets extend

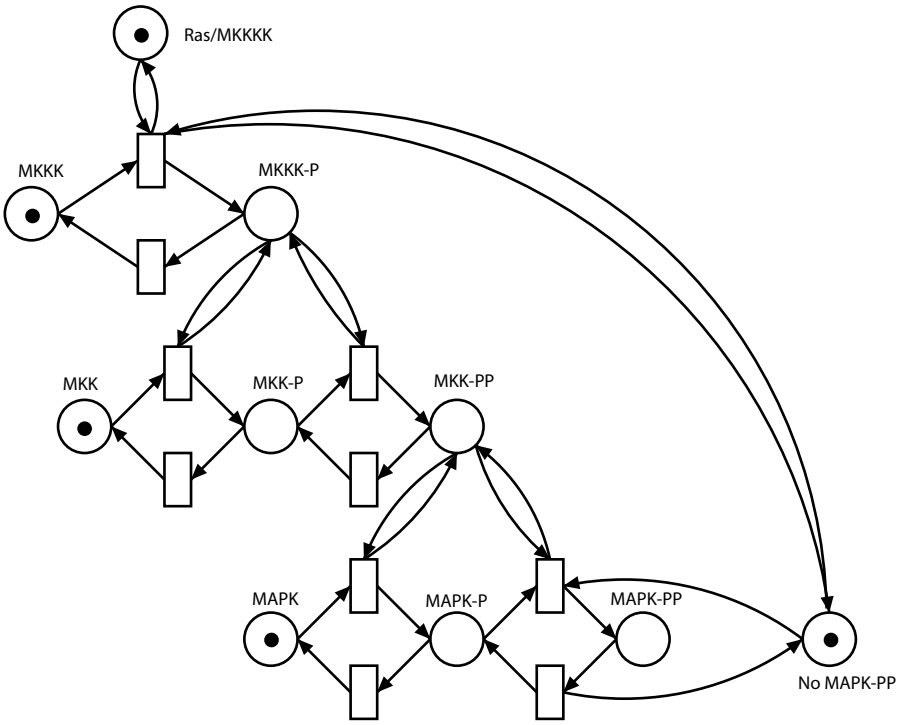


Figure 2.2: Petri net model of the MAPK cascade (features a complementary place for inhibition labelled "No MAPK-PP")

the basic definition of Petri nets allowing processing of continuous values on tokens; real continuous dynamics can be described using special kinds of places and transitions. This formalism also includes connectivity extensions such as inhibitory arcs. Another additional connectivity extension is a test arc (drawn as a dashed arrow) which defines a requirement for non-zero marking in a given place to allow a transition to proceed. This arc, however, defines that performing a transition does not impact the marking in a place connected to the transition with a test arc. Hybrid functional Petri nets are useful for illustrating system behaviour and mature simulation algorithms exist for these models. However these algorithms have several significant drawbacks, which we demonstrate with examples below. We refer to Cell Illustrator (Doi et al. 2004), which implements the hybrid functional Petri nets algorithms.

Example 2.3 (Hybrid functional Petri net model of a biochemical reaction)

The enzymatic activation of a protein with the following dynamics (2.2)*:*

$$\begin{cases} \dot{P} = -\dfrac{E \cdot P \cdot 2.7}{(E + 0.51) \cdot (P + 0.7)} \\ \dot{P}^* = \dfrac{E \cdot P \cdot 2.7}{(E + 0.51) \cdot (P + 0.7)} \\ \dot{E} = 0 \end{cases} \tag{2.2}$$

$$E|_{t=0} = 0.05, \;\; P|_{t=0} = 2.5, \;\; P^*|_{t=0} = 0;$$

can be described with the hybrid functional Petri net depicted in Figure 2.3.

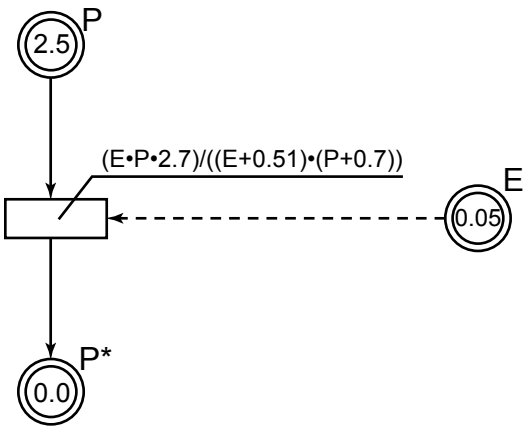
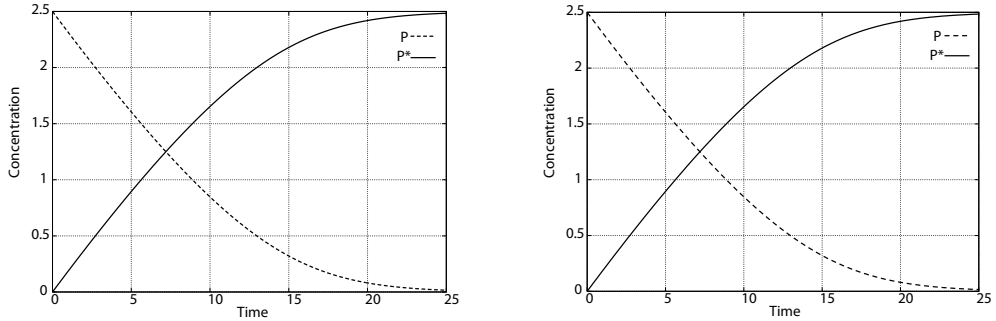


Figure 2.3: Hybrid functional Petri net model of enzymatic activation

Figure 2.4 shows the simulation results produced with MATLAB's (Moler 2004) ode15s differential equation solver (Moler 2004) (Figure 2.4(a)) and Cell Illustrator (Figure 2.4(b)). This comparison illustrates that the simulation results are the same.

However, due to the simulation strategy of hybrid functional Petri nets, the simulation results for some models can be incorrect. To illustrate this we constructed the following example. Consider a hypothetical biochemical network depicted in Figure 2.5(a). This model consists of six species x_1, x_2, x_3, m_1, m_2, and m_3. There are four reactions. All the reactions have mass action kinetics with coefficients k_1, k_2, k_3, and k_4. For this example the following coefficients values have been chosen: $k_1 = 0.013$, $k_2 = 1.0$, $k_3 = 2.5$, $k_4 = 0.087$, and the



(a) ODE solution

(b) Hybrid functional Petri net simulation

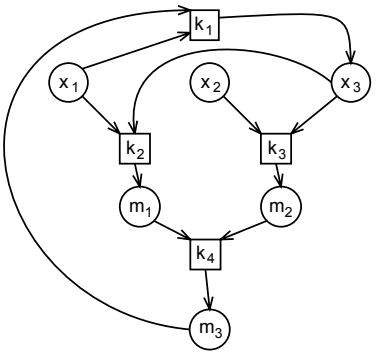
Figure 2.4: Enzymatic activation simulation results

initial concentrations for the species are the following: $x_1|_{t=0} = 10$, $x_2|_{t=0} = 10$, $x_3|_{t=0} = 1$, $m_1|_{t=0} = 0$, $m_2|_{t=0} = 0$, $m_3|_{t=0} = 0$. The network topology is depicted in Figure 2.5(a). The ODE model for this network is

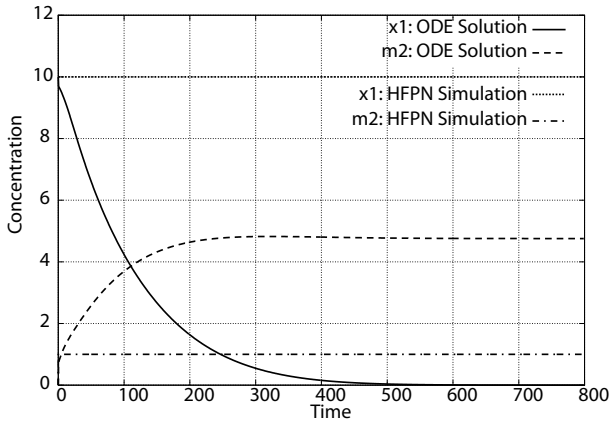
$$\begin{cases} \dot{x}_1 = -k_1 \cdot x_1 \cdot m_3 - k_2 \cdot x_1 \cdot x_3, \\ \dot{x}_2 = -k_3 \cdot x_2 \cdot x_3, \\ \dot{x}_3 = k_1 \cdot x_1 \cdot m_3 - k_2 \cdot x_1 \cdot x_3 - k_3 \cdot x_2 \cdot x_3, \\ \dot{m}_1 = k_2 \cdot x_1 \cdot x_3 - k_4 \cdot m_1 \cdot m_2, \\ \dot{m}_2 = k_3 \cdot x_2 \cdot x_3 - k_4 \cdot m_1 \cdot m_2, \\ \dot{m}_3 = k_4 \cdot m_1 \cdot m_2 - k_1 \cdot x_1 \cdot m_3, \end{cases} \tag{2.3}$$

$$k_1 = 0.013,\ k_2 = 1.0,\ k_3 = 2.5,\ k_4 = 0.087,$$
$$x_1|_{t=0} = x_2|_{t=0} = 10.0,\ x_3|_{t=0} = 1.0,\ m_1|_{t=0} = m_2|_{t=0} = m_3|_{t=0} = 0.0.$$

We solved this problem with MATLAB's *ode15s* solver and compare it with the hybrid functional Petri net (HFPN) simulation produced with Cell Illustrator. The results are plotted in Figure 2.5(b). The comparison shows that the hybrid functional Petri nets approach does not produce the same behaviour as the ODEs. This is due to the fact that the simulation algorithm for hybrid functional Petri nets can evaluate a flux through only one reaction at each given time. The reaction to be performed is chosen randomly. In this case, the reaction $x_2 + x_3 \rightarrow m_2$ has been chosen first, and this reaction consumed all available x_3. This took the model into a deadlock. But, continuous dynamics consume x_3 in both $x_2 + x_3 \rightarrow m_2$ and $x_1 + x_3 \rightarrow m_1$ reactions simultaneously, which produces the correct trace. We conclude that hybrid functional Petri nets do



(a) Schematic representation of model structure. Circles correspond to chemical species and rectangles correspond to reactions.



(b) A comparison of ODE solution result to a hybrid functional Petri net simulation.

Figure 2.5: An example of a biochemical model for which hybrid functional Petri nets predict an incorrect behaviour.

not produce correct predictions for this model.

We found that HFPN simulation algorithm also fails when analysing stiff models. In such a case, the simulation algorithm uses a very small value for the simulation step size, and the simulation cannot be completed in reasonable time. On the other hand, fixing the step size on larger values does not give suitable precision of the simulation results. Thus, hybrid functional Petri nets cannot be used for modelling stiff systems.

Stochastic Petri Nets can be used for modelling biochemical systems in a way similar to the approach proposed in Chapter 3. However, such Petri net models become quite complex to understand when the models become larger.

We conclude that the Petri nets can be quite illustrative for small examples, but they do not allow any quantitative reasoning and simulation of complex pathways are often imprecise or incorrect. This motivates the development of alternative modelling techniques which support structural view of the system, and also allow quantitative modelling.

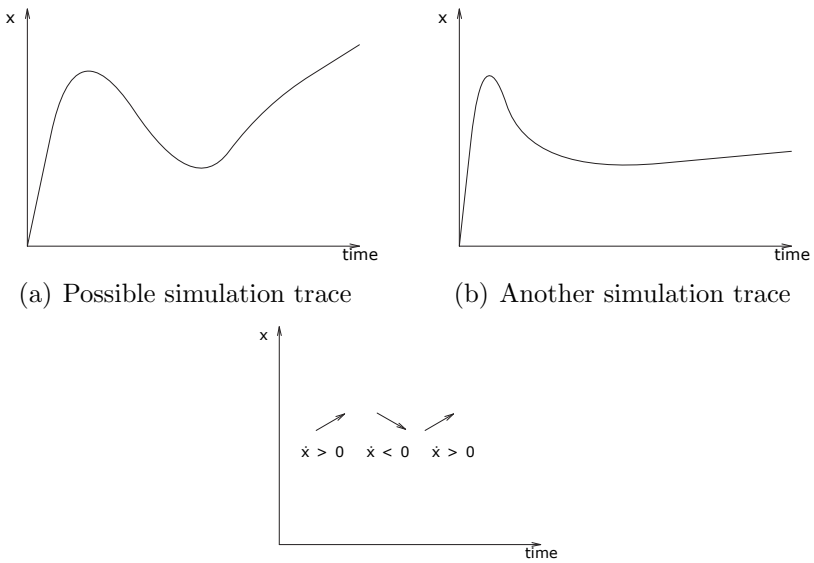
2.3 Hybrid Systems

Hybrid systems describe both discrete signals (or variables) and continuous signals or variables. There have been several attempts to use hybrid systems

for modelling and analysis of biochemical networks (Alur et al. 2002 Belta et al. 2004 Lincoln and Tiwari 2004).

The state space of possible variable values can be partitioned into rectangular *domains*, in which the flow is qualitatively identical. From this we derive a qualitative transition system, consisting of the set of all domains, the set of all transitions between the domains, and a labelling function that associates the sign of the derivatives of the concentration variables to every domain. A sequence of states in the qualitative transition system is called a *path*. A path describes a possible behaviour of the system. The qualitative transition system is designed such that it provides an *conservative approximation* of the dynamics of the original system, in the sense that to every solution of the model corresponds a path in the state transition graph. Note, that the converse is not true: some paths may not correspond to any solution, and therefore represent spurious behaviours.

The qualitative transition system can be used for model validation with model checking techniques. As the transition system is labelled with the signs of the derivatives, model checking queries can only describe trends of the concentration plots. Usually *Computational Tree Logic* (CTL) (see Clarke et al. 1999) is used to describe such model properties.



(a) Possible simulation trace (b) Another simulation trace

(c) Derivative sign pattern

Figure 2.6: Qualitative properties for hybrid systems analysis

For example, the following is a property:

$$\mathbf{EF}(\text{dsign\_x} = 1 \wedge \mathbf{EF}\ (\text{dsign\_x} = \text{-}1 \wedge \mathbf{EF}(\text{dsign\_x} = 1))),$$

where dsign_x is a built-in variable for the sign of the derivative of variable x, and the $\mathbf{EF}\phi$ quantifier means that there must exist at least one path on which ϕ eventually holds. The expression above defines a property for the value of x as depicted in Figure 2.6(c). Note that this property describes both possible behaviours plotted in Figure 2.6(a) and Figure 2.6(b).

The analysis of realistic models leads to large state transition graphs, which make verification of dynamic properties practically infeasible.

It is possible to implement another approximation of model behaviour, that conserves the quantitative characteristics of the system. In this case, the state space of the system is decomposed into a number of hyperrectangles defining approximation domains. The behaviour of the modelled system is approximated with a linear behaviour in each of these domains. It is possible to evaluate more quantitative properties using specialised reasoning algorithms. This approach can be successfully applied to modelling of small networks, but does not scale up to larger models. For example, a model of the MAPK cascade proposed by Schoeberl et al. (2002) consists of 94 species, therefore the state space for this model will have 94 dimensions. The problem to generate a decomposition of this space into a number of domains is infeasible by itself. Even an approximation with three linear segments per specie is described with $3^{94} > 10^{44}$ rectangles.

We conclude that due to the complexity of state space decomposition into multiple domains, this approach cannot be applied effectively to the simulation and analysis of large biological systems.

2.4 Chemical Master Equation and Stochastic Simulation

Stochastic simulation involves modelling individual molecules. Most simulations abstract away from location and motion of individual molecules, which is justified if one assumes that the system is *well stirred*, which means that the molecules of all kinds are uniformly distributed through the spatial volume. The following is usually assumed when considering stochastic simulations: the system is in thermodynamic equilibrium, and the volume is fixed. The state of such system is described by a vector $\mathbf{X}(t) = (X_1(t), X_2(t), \ldots, X_N(t))$, where $X_i(t)$ is a non-

negative integer which expresses a number of molecules of the ith kind at time t. $\mathbf{X}(0)$ describes the initial state of the system.

At each time when one of the reactions takes place, vector $\mathbf{X}(t)$ can change its value. This leads to the *Chemical Master Equation* (CME), which is a system of ODEs, one ODE per each possible state of the system. At time t, the kth equation defines the probability that the system is in the kth state. Unlike the models considered in Section 2.1, the dimensionality of such a system depends not on the number of chemical species N, but on the number of possible states of the system $\mathbf{X}$, which in turn depends on the total number of molecules.

Usually, the dimension of the CME is so huge that it is not possible to work with it either analytically or numerically.

Gillespie (1977) proposed a stochastic simulation algorithm which produces model behaviours using the CME indirectly. Instead of solving the system of ODEs to get the probability distribution over the possible states of the system at each time t, the algorithm produces samples from such distributions.

In spite of the simplicity of the Gillespie's algorithm, it can be quite inefficient when some reactions take place frequently. The basic algorithm can be improved by "lumping together" several reactions, and changing the state vector only when several reactions took place. This method is called tau-leaping approximation (see Wilkinson 2006). The error of this approximation is small in the cases where the system state changes are small.

Several software platforms implement the algorithm (see Kierzek 2002 Adalsteinsson et al. 2004 Gillespie et al. 2006), but they do not offer additional reasoning or analysis capabilities (beside simulation).

2.5 Process Algebras and Process Calculi

2.5.1 π-Calculus and Stochastic π-Calculus

The original π-calculus (sometimes referenced as pi-calculus) was developed by Milner (1999) as a formal language for concurrent computational processes. The π-calculus provides a framework for representation, simulation, analysis and verification of mobile communicating systems. In fact, the π-calculus, just as the μ-calculus (see Kozen 1983), is so minimal that it does not contain primitives such as arithmetic (no numbers, no operations), boolean values, flow control

statements usual for programming languages, data structures, variables or functions.

When modelling biochemical systems with π-calculus, individual molecules and their domains are treated as computational processes, while their complementary structural and chemical determinants correspond to communication channels. Processes can be composed in parallel, which means that they are performed at the same time, and communication between such processes is achieved through rendezvous. Communication channels between computational processes express chemical bonds between molecules or molecular domains. Biochemical interaction and subsequent modification is usually expressed with communication involving channel transmission. Biochemical reactions are modelled by passing the communication channels from one process to another, thus altering the communication topology, and therefore describing new bonds between molecules and molecular domains. For an example of how a model of a signal transduction pathway can be defined with π-calculus see (Regev et al. 2001).

There exists a number of analysis techniques which can be applied to π-calculus models of biochemical systems, such as simulation or reachability analysis.

The basic formulation of π-calculus does not allow quantitative dynamics. Therefore, only qualitative analysis is possible on such models. However, there exists a number of extensions which allow association of stochastic time delays with interactions to be made. One of such extensions is the stochastic π-calculus (see Priami 1995), and there is a specific implementation of stochastic π-calculus tools for modelling and simulation of biochemical networks called BioSPi (see Priami et al. 2001).

Considering analysis techniques for the stochastic π-calculus, these are limited to quantitative simulations of model behaviour which is usually achieved with the Gillespie algorithm (see Section 2.4). Since π-calculus and its stochastic extension consider individual molecules as basic components of a model, there is a problem of a state-space explosion. These approaches, however, are suitable for modelling systems with only a few molecules, such as protein-DNA interactions, transcription and translation modelling.

We conclude that, as no tools for quantitative reasoning or inference are available at the moment, these modelling formalisms are mostly suitable for simulation.

2.5.2 Performance Evaluation Process Algebra (PEPA)

PEPA (see Hillston 1996) is a stochastic process algebra which has also been applied to modelling biochemical systems (see Calder, Gilmore and Hillston 2006). Unlike π-calculus, the PEPA models do not consider protein structure directly, but operate on the level of abstract interactions associated with biochemical reactions. Instead of describing molecular domains and binding them together using communication channels, PEPA models assign symbolic names to different species involved in a biochemical network.

Semantically, PEPA is different to stochastic π-calculus, as the former is a proper process algebra. This allows a natural comparison of models to be made (e.g. by bisimulation).

Considering an algorithmic support of modelling using PEPA, there exist a number of software frameworks which implement steady state and transient analysis of models, for example with the PEPA workbench (see Gilmore and Hillston 1994). The PRISM model checker (see Kwiatkowska et al. 2002) also supports stochastic model checking and Monte-Carlo simulation of such models.

PEPA allows multi-way synchronisation (synchronous actions for more than two processes at the same time) therefore allowing more abstract modelling of biochemical reactions which involve multiple reactants. A stochastic rate can be associated with each event in this process algebra, therefore enabling quantitative modelling of dynamic systems. In some cases a component can be passive with respect to some activity. This means that the rate of the activity will be left unspecified (denoted $\top$) and is determined upon cooperation, by the rate of the activity on the other component.

Calder, Gilmore and Hillston (2006) model the RKIP inhibited ERK pathway using this process algebra. Two different models of the system have been developed: a reagent-centric model and a pathway-centric one. In the former model, each protein in the system is associated with a computational process, while the reactions are the actions performed by the processes. The performance rate is associated with each reaction, which allows the reaction speed to be controlled; all the proteins can have only two states: low concentration and high concentration. In the low state the protein cannot participate in reactions as a reagent, but can be a product of a different reaction. If the protein is produced, it changes its state from low to high.

In the pathway-centric model, parts of the pathway are considered as pro-

cesses. Such pathways can be synchronised together to reproduce a behaviour of the complete network.

Calder, Gilmore and Hillston (2006) provided extensive analysis of these models. It was shown that both models are bisimilar, and therefore define the same observable network behaviour. This could be useful at the stage of building the model. For example, users can abstract from defining fine-grained players of the network, concentrating on pathway composition. Particular players can be defined at the next stage. This approach allows different networks to be compared, and decide whether they simulate each other.

2.6 Pathway Logic

Pathway Logic (Eker et al. 2002) is a qualitative analysis technique based on term rewriting.

Pathway Logic is currently used for modelling and analysis of signal transduction and metabolic networks. Pathway Logic models are represented using the Maude term rewriting system (see Clavel et al. 2003). Models can be queried and computational experiments can be carried out using the execution, search and model-checking tools of the Maude system. Some current capabilities of Pathway Logic include:

- Models with different levels of detail. This means that a model can be described either on the scale of species and reaction, or on the scale of molecular domains. For example, it is also possible to define cellular compartments and therefore model spacial characteristics of the system, and molecule transport.

- Analysis of models using search and model-checking. This allows one to verify logical properties of the models in addition to performing simulations. Though, this approach does not allow quantitative modelling, as the steps of rewriting do not carry the timing information.

- Transformation to Petri nets for analysis and visualisation.

Using Pathway Logic, biological molecules, their states, and their roles in network elements can be modelled at very different levels of abstraction. For example, a complex signalling protein can be modelled either according to an

overall state or as a collection of functional domains (protein functional domains, PFDs) and their internal or external interactions (Talcott et al. 2004). The same hierarchy of organisation levels can be modelled with π-calculus (see Section 2.5.1), but not with, for example, Petri nets.

The main disadvantage of Pathway Logic is that it does not support quantitative modelling, as neither time nor variables can take quantitative values.

2.7 BIOCHAM

BIOCHAM (Chabrier and Fages 2003) is a logical approach which allows classical model checking algorithms to be applied for biochemical network analysis. In particular, BIOCHAM translates biological models into model definitions for the SMV model checker (see Clarke et al. 1994), which then can be used to perform property validation.

The basic approach uses qualitative models. Chabrier-Rivier et al. (2004) propose a number of logical properties, which have biological meaning, using Computational Tree Logic, CTL (Clarke et al. 1999). Example queries are

- Does the system have a stable state?
- Given an initial state of the system, is there a series of reactions that will produce some compound P?
- Is state s_2 a necessary checkpoint for reaching state s?

The basic approach is extended with quantitative kinetic parameters. The system behaviour can then be simulated using these parameters. Calzone et al. (2005) introduce additional features to allow quantitative reasoning. These extensions enable checking LTL (Linear Temporal Logic) properties on quantitative simulation traces. Some machine learning techniques are proposed in (Calzone et al. 2006) which allow model reconstruction and parameter search to fit model behaviour to the LTL properties.

In the case of quantitative model checking, each time point of the system trace is represented by a different state of a transition system. A simulation trace is therefore a sequence of states with unlabelled transitions between them, as depicted in Figure 2.7. The properties can be described using LTL extended by constraints over variable values and their derivatives.

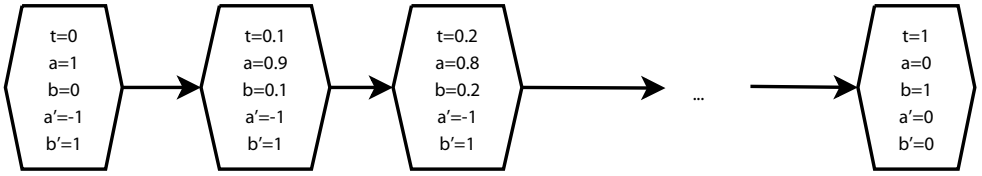


Figure 2.7: Transition system generated from the model simulation trace. t denotes time, a and b are model variables, a' and b' are their derivatives, respectively.

The parameter search is implemented by iterative scanning through the intervals of possible parameter values.

BIOCHAM implements algorithms for qualitative reasoning about structural models, and quantitative model checking of properties of model simulation traces. In Chapter 3 we extend the idea by enabling quantitative model checking of structural models.

This concludes our review of the approaches for modelling biological systems. In the following sections we review alternative methods for model identification using experimental data as the main source of information.

2.8 Maximum Likelihood Estimation for Inference

Maximum likelihood estimation is a method used to make inferences about parameters of an underlying model from experimental data. The goal is to find parameters that fit the experimental data as closely as possible.

This approach can work with any parametric quantitative model provided the likelihood function, which defines the probability that the model reproduces the experimental data, is defined. We assume that models are formulated using ODEs, however, the arguments below are relevant to any kind of quantitative models used within the maximum likelihood estimation framework.

The problem of finding values of model parameters which maximise the likelihood is not trivial, as many of models of biological systems are nonlinear. There are numerous algorithms for maximum likelihood estimators, e.g. simulated annealing (Kirkpatrick et al. 1983),, genetic algorithms (Crosby 1973), stochastic gradient descent (Spall 2003), tabu search (Glover and Laguna 1997).

Maximum likelihood estimators propose a single value representing the “best

estimate" of the model parameters. This is so-called *point estimation* problem. Note that, in such formulation, the final answer is an element of the parameter space, with no explicit recognition of the uncertainty involved. Pragmatically, a point estimate may be motivated as being the simplest possible summary of the inferences to be drawn from the experimental data about the value of the model parameters.

There is a number of methods designed for estimation of corresponding uncertainty for the point estimates. This is achieved either through confidence limits estimation (see e.g. Cox and Hinkley 1974), through local approximations to the posterior density (these will be discussed in more detail in Section 4.1.2), or using Markov chain Monte Carlo to "inflate" the modes around the maximum likelihood point estimate to evaluate the local variance of the posterior (this method is referred as "ensemble" method and discussed in great detail by Brown et al. (2004)). The main problem with these methods is that none of them considers the complete parameter posterior and thus will fail to account for alternative inferences following from possible multimodal posteriors or posteriors of peculiar shape. Non-trivial posterior distributions are not rare when considering models of biological systems due to nonlinearities involved.

The hypotheses testing problem and the problem of model comparison are usually solved with either simple likelihood ratio test (which would prefer more complex hypotheses over simple ones as long as they provide better fit to data) or, in an attempt to compensate for that effect, using the Akaike Information Criterion (see Akaike 1973). We discuss different Information Criteria in detail in Section 4.1.4. The Akaike Information Criterion is well justified for cases when predictions of the prior[2] are compatible to those of the likelihood, and not in the more usual situation when prior information is small in comparison to the information provided by the data. Another major problem with the Akaike Information Criterion is that it is based on an asymptotic approximation to the parameter posterior, and therefore it is only valid when the posterior distribution is approximately multivariate normal.

The latter problem with the Akaike Information Criterion (as well as several other methods based on the same principle, which are discussed in Section 4.1.4) is the major drawback of this approach, because nonlinear parameter posteriors

[2]Knowledge about possible model parameter values available before the experimental data is observed.

are very common in biological modelling[3] context.

2.9 Bayesian Networks

Bayesian network inference algorithms (Sachs et al. 2002), (Li and Chan 2004), (Woolf et al. 2005), (Werhli et al. 2006), (Sachs et al. 2005) construct graphs in which nodes represent the measured species, and arcs represent statistical relations and dependencies between species. There are several attractive properties of Bayesian networks for the inference of signalling pathways from biological data sets. As emphasised by Sachs et al. (2005), Bayesian networks can represent complex stochastic nonlinear relationships among multiple interacting species, and their probabilistic nature can accommodate noise that is inherent in biologically derived data. Such networks can describe direct molecular interactions as well as indirect influences that proceed through additional unobserved components. Therefore, very complex relationships can be modelled and discovered.

The key problem for applying Bayesian networks to signal transduction pathways studies (which are considered as case studies in this thesis) is that this approach requires quantitative data for every player involved in a biological system (e.g. every specie of a biochemical network) to be available. Bayesian networks are capable of inferring the models which include only the players which are observed. Collection of all this data is very difficult in most cases due to the limitations of current experimental methods. The vast majority of the species would, therefore, be omitted from the network identified with the Bayesian networks approach, and replaced with arcs which specify statistical dependency of the observed species. The most discouraging evidence against the application of Bayesian networks is that, in practice, the amount of available data is very small, especially when the study of a network is at an early stage.

2.10 Discussion

Some of the modelling approaches considered in this section aim to reproduce qualitative properties of modelled systems, these are basic Petri nets, original

[3]As mentioned earlier, nonlinear models are common in biological applications, and very often such models have peculiar parameter posteriors, and in some cases even multimodal ones.

(non-stochastic) π-calculus, Pathway logic or the qualitative part of the methodology implemented in BIOCHAM. Other approaches are essentially quantitative: ODEs, hybrid systems and hybrid functional Petri nets, stochastic calculi and process algebras, and the quantitative methodology in BIOCHAM.

The overview provided in this chapter demonstrates that the majority of the existing quantitative approaches are aimed primarily for simulation of model behaviours. In the following chapters, we develop a modelling methodology and an inference framework which allow reasoning and inference about models of biological systems.

BIOCHAM supports logical matching of simulation traces to behaviour templates. However, this kind of logical analysis can hardly be used for deductive or inferential reasoning about biological models. PEPA supports stochastic descriptions of biological models, and supports some quantitative analysis techniques (including model simulation/bisimulation checking), we suggest a closely related modelling approach presented in Chapter 3 which allows logical reasoning about temporal properties using probabilistic symbolic model checking. In Chapter 3 we build upon the idea of using abstract levels of concentration for quantitative modelling, and propose to use multiple symbolic intervals of the concentration for more precise quantitative modelling.

We discussed some drawbacks of the inferential methodology based on maximum likelihood estimation methods. In Chapter 4 we propose to adopt Bayesian inferential methodology which overcomes these problems.

Bayesian nets seem to be an interesting approach which allows model inference from empirical data, but this approach requires large amounts of experimental data. Whilst good quality data is available in some areas of biological research (such as metabolic networks research), it is very expensive and mostly unavailable in other areas. The methodology we propose in Chapter 4 operates on predefined mechanistic models of biochemical pathways, so instead of reconstructing the models from data, we propose evidence-based hypotheses testing, treating these models as a description of working hypotheses.

Summary

We have provided an overview of the existing modelling and reasoning approaches in Systems Biology. We considered several qualitative modelling methodologies: Petri Nets, Pathway logic, BIOCHAM,

and π-calculus. We also considered quantitative methodologies, such as stochastic simulation and the Chemical Master Equation, systems of ordinary differential equations, stochastic π-calculus, and PEPA.

We reviewed existing methods for analysis and inference of quantitative models using the methods of maximum likelihood estimation. We identified some weaknesses of the existing approaches and suggest a methodology to tackle those weaknesses in the following chapters.

Chapter 3

Model-Based Reasoning

Overview

In this chapter we investigate how the methods of probabilistic model checking on continuous time Markov chains can be employed to reason about behaviours of biological systems.

3.1 Background

Traditionally, systems of differential equations are employed for modelling biological systems; models are usually defined using ordinary differential equations or, in some cases, partial differential equations. In this chapter we consider stochastic modelling approaches, namely those based on continuous time Markov chains. There exist a number of alternative stochastic approaches, such as *Chemical Master Equation* which can be simulated using the Gillespie algorithm (see Gillespie 1977), or *stochastic π-calculus* (see Priami 1995). The main problem with these approaches is that they are mainly designed for simulation of model behaviour and little, if any, logical analysis methods are developed for them. The approach proposed in this chapter is designed to allow logical analysis of the models. Additionally, it overcomes some of the drawbacks with ODE modelling, namely

- The structure of the ODE system is flat, thus it is hard to recognise particular interactions in system behaviour;
- ODEs are mainly suitable for simulation of system behaviours and not for reasoning;

- ODEs are deterministic and therefore they may over-specify the knowledge about biological systems.

The approach described in this chapter resolves these problems; as follows:

1. The structure of a biological system is preserved in a natural way, making interactions between components clear.

2. There are logical analysis and reasoning methods for stochastic models described with continuous time Markov chains. We employ probabilistic model checking of Continuous Stochastic Logic (CSL) for reasoning. CSL was shown to be decidable (Aziz et al. 1996).

3. It is possible to reason about system behaviours in a semi-quantitative way, for example, using intervals instead of fixed values.

We start this chapter with a concise overview of the underlying theory of continuous time Markov chains and Continuous Stochastic Models, and then demonstrate how these concepts can be applied to modelling and reasoning about biological systems.

3.1.1 Continuous Time Markov Chains

This section gives a brief introduction to the concept of CTMCs. The definitions are taken from (Kwiatkowska 2003).

Definition 3.1: *A continuous time Markov chain (CTMC) is a tuple* $(S, \bar{s}, \mathbf{R}, L)$ *where*

1. S *is a finite set of states*

2. $\bar{s}$ *is an initial state*

3. $\mathbf{R} : S \times S \rightarrow \mathbb{R}_{\geq 0}$ *is the rate matrix, and*

4. $L : S \rightarrow 2^{AP}$ *is a labelling with atomic propositions from set* AP.

If $\mathbb{R}(s_i, s_j) > 0$ for a pair of states s_i, s_j we say that there is a *transition* from state s_i to state s_j. The elements of matrix $\mathbf{R}$ are *rates* for the transitions, meaning that the probability that a transition from state s_i to state s_j will

be taken within time t is $1 - e^{-\mathbf{R}(s_i,s_j)t}$. If for a pair of states s_i, s_j the rate $\mathbf{R}(s_i, s_j) = 0$, then we say that there is no direct transition from state s_i to state s_j.

Note, that these chains have the Markov property, as the exponential distribution used to define the transition rates is "memoryless" and thus the stochastic process is conditioned only by its current state.

In the cases when $\mathbf{R}(s_i, s_j) > 0$ for more than one state s_j, a *race* (often referenced as a *racing condition*) between the outgoing transitions from s exists. That is, the probability $\mathbf{P}(s, s')$ of taking a transition from s to s' in a single step equals the probability that the delay of going from s to s' is smaller than the delays for any other outgoing transition from s.

Definition 3.2: *A path in a CTMC is a non-empty sequence $s_0 t_0 s_1 t_1 s_2 \ldots$ where $\mathbf{R}(s_i, s_{i+1}) > 0$ and $t_i \in \mathbb{R}_{\geq 0}$ for all $i \geq 0$. The value t_i represents the amount of time spent in the state s_i.*

Analysis of CTMCs is based on transient (the state of the CTMC at a particular time instant) and steady-state (the state of the CTMC in the long run) behaviour.

Definition 3.3: *The transient probability $\pi_{s,t}(s')$ is defined as the probability, having started in state s, of being in state s' at time instant t.*

Definition 3.4: *The steady-state probability $\pi_s(s')$ is defined as $\lim_{t\to\infty} \pi_{s,t}(s')$.*

An example of a CTMC is given in Example 3.1.

Example 3.1 (A continuous time Markov chain)

Consider a set of states $S = \{s_0, s_1, s_2, s_3, s_4, s_5\}$, state s_0 is the initial state;

transition matrix:

$$\mathbf{R} = \begin{array}{c} \\ s_0 \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{array}{c} \begin{array}{cccccc} s_0 & s_1 & s_2 & s_3 & s_4 & s_5 \end{array} \\ \begin{bmatrix} 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

the set of atomic propositions:

$$AP = \{A = 0,\ A = 1,\ A = 2,\ B = 0,\ B = 1,\ B = 2,\ C = 0,\ C = 1,\ C = 2\}$$

and labelling:

$$L = \begin{cases} s_0 \rightarrow \{A = 2, B = 0, C = 0\} \\ s_1 \rightarrow \{A = 1, B = 1, C = 0\} \\ s_2 \rightarrow \{A = 1, B = 0, C = 1\} \\ s_3 \rightarrow \{A = 0, B = 2, C = 0\} \\ s_4 \rightarrow \{A = 0, B = 0, C = 2\} \\ s_5 \rightarrow \{A = 0, B = 1, C = 1\} \end{cases}$$

The CTMC $(S, s_0, \mathbf{R}, L)$ *is schematically depicted in Figure 3.1.*

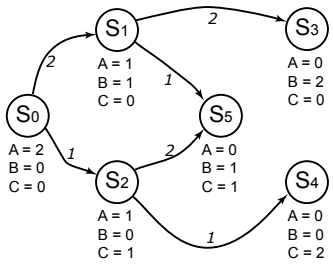


Figure 3.1: Structure of the CTMC.

Transient probabilities for states $s_1, s_2,$ *and* s_3 *are plotted in Figure 3.2. Steady-state probabilities for all of the states are:*

$$\pi_{s_0}(s_0) = 0 \qquad \pi_{s_0}(s_3) = 4/9$$
$$\pi_{s_0}(s_1) = 0 \qquad \pi_{s_0}(s_4) = 1/9$$
$$\pi_{s_0}(s_2) = 0 \qquad \pi_{s_0}(s_5) = 4/9$$

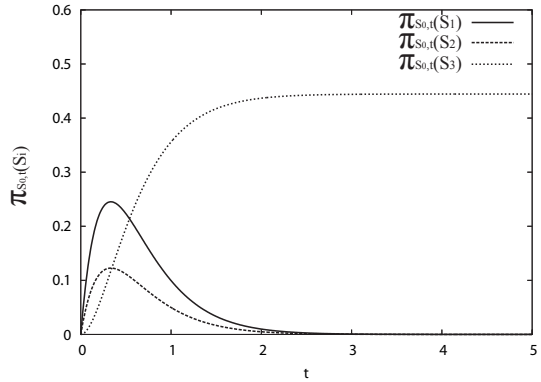


Figure 3.2: Transient probabilities for some states in the CTMC

3.1.2 Continuous Stochastic Logic

Properties of CTMCs can be described using *continuous stochastic logic* (CSL, see Aziz et al. 1996) which is based on computational tree logic (CTL, see Emerson 1990) and contains probabilistic operators evaluated with respect to path-based measures on CTMCs. Some definitions of CSL (e.g. Kwiatkowska et al. 2005) also include a steady-state operator.

Consider a CTMC $M = (S, \bar{s}, \mathbf{R}, L)$ with labelling L over atomic propositions AP. Two types of formulae are used to define CSL: *state formulae* (which are true or false in a specific state), and *path formulae* (which are true or false along a specific path).

Definition 3.5: *A state formula is defined inductively:*

1. $a \in AP$ *is a state formula,*
2. *if* ϕ_1 *and* ϕ_2 *are state formulae, then so are* $\neg\phi_1$, $\phi_1 \wedge \phi_2$, $\phi_1 \vee \phi_2$.
3. *if* ψ *is a path formula, then* $P_{\lessgtr c}(\psi)$ *is a state formula, where* $c \in [0, 1]$, *and* $\lessgtr \in \{<, \leq, \geq, >\}$,
4. *there are no other state formulae.*

Definition 3.6: *Path formulae are formulae of the form*

$$\phi_1 \mathbf{U}_{[a_1,b_1]} \phi_2 \mathbf{U}_{[a_2,b_2]} \ldots \phi_n,$$

where $\phi_1, \phi_2, \ldots \phi_n$ *are state formulae, and* $\forall i \in [1 \ldots n]\,(a_i \in \mathbb{R}_{\geq 0}, b_i \in \mathbb{R}_{\geq 0})$.

Definition 3.7: *Continuous stochastic logic is the set of state formulae generated by the above rules.*

Note, that CSL contains only state formulae; path formulae are used only in definitions of some state formulae.

The formal semantics for continuous stochastic logic can be found in (Aziz et al. 1996). The only unusual operator requiring explanation is a *time limited until*:

$$P_{\lessgtr c}(\phi_1 \mathbf{U}_{[a,b]} \phi_2). \tag{3.1}$$

Formula (3.1) holds at a state s_i if and only if the probability to reach a state which satisfies ϕ_2, from the state s_i, passing only through states which satisfy ψ_1 within time interval $[a, b]$, is $\pi \lessgtr c$.

3.1.3 Verification of CTMCs

Properties of CTMCs expressed in CSL can be verified using model checking techniques. Kwiatkowska (2003) argues that the complexity of CSL model checking is linear in the size of the formula, and polynomial in the state space of CTMC.

Some properties of interest for Example 3.1 are given in Example 3.2.

Example 3.2 (CSL properties for the CTMC defined in Example 3.1)

- $A = 2$ – *all the states of the CTMC have labelling* $A = 2$. *This property* ***does not hold***.
- $P_{\geq 1}((true)\mathbf{U}_{[0,+\infty]}(B = 2))$ – *a state with labelling* $B = 2$ *is eventually reachable in infinite time with probability 1. This property* ***does not hold***, *as once a state with* $C > 0$ *is reached it is not possible to get to a state where* $B = 2$ *any more.*
- $P_{\geq 0.4}((true)\mathbf{U}_{[0,+\infty]}(B = 2))$ – *a state with labelling* $B = 2$ *is reachable in infinite time with probability greater or equal to 0.4. This property* ***holds*** *as the probability to reach a state where* $B = 2$ *is equal to* $\frac{4}{9}$, *which is greater than* 0.4.

3.1.4 The PRISM Model Checker

The PRISM model checker is a tool for defining and analysing stochastic models. It supports a description language to define models with continuous time Markov chains. It also supports discrete time Markov chains and Markov decision processes, however we confine our interest to CTMCs. PRISM allows users to define properties of CTMCs in CSL, and check them, employing numerical solutions of linear equation systems and linear optimisation problems. For a detailed description see Kwiatkowska et al. (2005).

The PRISM model description language is based on the Reactive Modules formalism of Alur and Henzinger (1999).

In this section we give a short overview of the PRISM language; the complete syntax and semantics of this language can be found in the PRISM user's guide (see Kwiatkowska et al. 2006).

The fundamental components of the PRISM description language are *modules* and *variables*. A model consists of modules which can interact with each other. Variables can be either local (in the scope of a module) or global (accessible to all modules in a model). Each variable has a type: boolean, integer or real. A module contains local variables. The local state of a module is a vector of values for local variables of this module. The *global state* of the model is determined by the global variables together with the local state of all modules.

The behaviour of each module is defined by a set of *commands*. A command is defined as:

$$[a]\ g \rightarrow \lambda : v_i' = f_i(v_1, \ldots, v_n) \& \ldots \& v_j' = f_j(v_1, \ldots, v_n);$$

The *guard* g is a predicate over all the variables in the model (including variables which belong to other modules). a is a symbolic label associated with this command, it is usually called *an action*, and is used for synchronisation. A transition is specified by giving the new values of the variables in the module, possibly as an expression formed from other variables or constants using assignment expressions:

$$v_i' = f_i(v_1, \ldots, v_n),$$

where v_i' denotes the new value for variable v_i, which determines a new local state of the module which is the destination of the transition defined with such a command. The expression λ assigns a rate to the transition.

The CTMC defined in Example 3.1 can be described using one module with the following PRISM language definition (see Example 3.3):

Example 3.3 (CTMC from Example 3.1 defined in PRISM language.)

```
ctmc

module module1
    pA: [0..2] init 2;
    pB: [0..2] init 0;
    pC: [0..2] init 0;

    [] (pA > 0)&(pB < 2) -> 2: (pA' = pA - 1)&(pB' = pB + 1);
    [] (pA > 0)&(pC < 2) -> 1: (pA' = pA - 1)&(pC' = pC + 1);
endmodule
```

The keyword `ctmc` *specifies that the model is a continuous time Markov chain. Note, the overlapping guards indicate a racing condition between transitions.*

The CTMC generated from the PRISM language description is the parallel composition of all the modules. PRISM supports multi-way *synchronisation* in the style of process algebras. To enable such synchronisation, commands are labelled with *actions* given within square brackets. A multi-modular model with synchronisation is illustrated in Example 3.4:

Example 3.4 (A PRISM model with multiple modules)

The following defines the CTMC from Example 3.1:

```
ctmc

const double lambda = 2;
const double gamma = 1;

module module1
    pA: [0..2] init 2;

    [action1] (pA > 0) -> lambda: (pA' = pA - 1);
```

```
    [action2] (pA > 0) -> gamma: (pA' = pA - 1);
endmodule

module module2
    pB: [0..2] init 0;

    [action1] (pB < 2) -> 1: (pB' = pB + 1);
endmodule

module module3
    pC: [0..2] init 0;

    [action2] (pC < 2) -> 1: (pC' = pC + 1);
endmodule
```

The ***action1*** *action in the following command:*

```
    [action1] (pA > 0) -> lambda: (pA' = pA - 1);
```

is used to force two or more modules to make transitions simultaneously (i.e. to synchronise). For example, in state $(pA = 2,\ pB = 0,\ pC = 0)$*, the model can move to state* $(pA = 1,\ pB = 1,\ pC = 0)$*, synchronising over* ***action1****. The rate of the synchronised transition is the product of the individual rates for the synchronising action (in this case,* $lambda \cdot 1 = lambda$*).*

Note that by default all modules are synchronised over all their common actions.

PRISM also supports definition and verification of properties based on *rewards*, therefore enabling reasoning about expected values. The basic idea is to associate numerical values with states or transitions of the model. Rewards can be associated with models using the `rewards...endrewards` construct. State rewards can be specified using multiple reward items, each of the form:

$$guard : reward;$$

where *guard* is a predicate (over any variables of the model) and *reward* is an expression (containing any variables or constants from the model). For example,

```
rewards
    x = 0: 100;
    x > 0 & x < 10 :  2*x;
    x = 10: 100;
endrewards
```

assigns a reward of 100 to states satisfying $x = 0$ or $x = 10$ and a reward of $2 * x$ to states satisfying $x > 0 \,\&\, x < 10$. Any states which do not satisfy the guard of any reward will have no reward assigned to them. For the states which satisfy the guards of several rewards, the reward assigned is a sum of the rewards for all the corresponding reward items. Rewards can also be assigned to transitions of a model (see Kwiatkowska et al. 2006). We do not use transition rewards for our analysis.

Properties in PRISM are expressed in a language based on CSL and a number of additional customisations and extensions are supported. One of the customisations is evaluation of a property to a numeric value. This is achieved by replacing the probability bounds in CSL properties with "=?" and is illustrated in the following example:

Example 3.5 (Evaluation of property probabilities using PRISM)

Some probabilities evaluated using PRISM on the CTMC defined in Example 3.1:

- $P_{=?}((true)\mathbf{U}_{[0,+\infty]}(B = 2)) = 0.444$
- *probabilities* $P_{=?}((true)\mathbf{U}_{[t,t]}(B = 1))$ *and* $P_{=?}((C = 0)\mathbf{U}_{[t,t]}(B = 1))$ *for different values of* t *are plotted in Figure 3.3.*

Note that the probabilities are evaluated on the paths which start in the initial state of a CTMC.

The satisfaction of a property (i.e. whether it is true or false) is defined for a single state of a model. When analysing a property, PRISM considers it to be true if it is satisfied in *all* the states of the model, and false otherwise. The satisfaction of a property in a particular state can be verified by considering subsets of model states. This is expressed in CSL by formulae of the form:

$$\phi \rightarrow \psi.$$

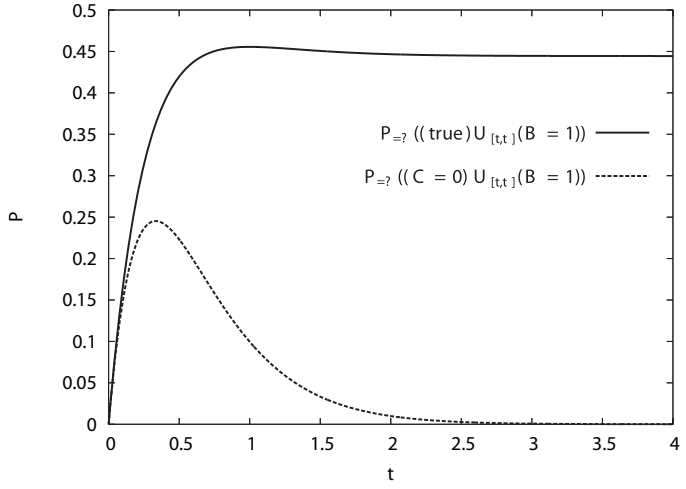


Figure 3.3: Property probabilities computed with the PRISM model checker.

Such a formula holds when ψ is satisfied at all the states satisfying ϕ. This is illustrated by the following example:

Example 3.6 (Verifying properties in a subset of model states)

Consider the model defined in Example 3.1 and the CSL property

$$P_{>0.4}((true)\mathbf{U}_{[0,+\infty]}(B = 2)).$$

PRISM evaluates this property to false because when $C > 0$ no state with $B = 2$ can be reached, but this property holds at the initial state of the model (as probability of reaching $B = 2$ from the initial state is 0.444 as shown in Example 3.5). The expression above can be extended with an implication to evaluate this property only on the initial state thus:

$$(A = 2) \rightarrow P_{>0.4}((true)\mathbf{U}_{[0,+\infty]}(B = 2)).$$

This property is evaluated to be true. Note that

$$(A = 2) \rightarrow P_{>0.5}((true)\mathbf{U}_{[0,+\infty]}(B = 2)).$$

is evaluated to false, since 0.444 is less than 0.5.

The operator $S_{\lessgtr c}(\phi)$ is used for steady-state properties. In the cases when $\lessgtr c$ is substituted with "=?", the actual steady-state probability will be evaluated.

This is illustrated by the following example:

Example 3.7 (Steady-state properties in PRISM)

Consider the CTMC defined in Example 3.1. Some steady-state properties and their evaluations are as following:

- $S_{=?}(B > 0) = 0.888$
- $S_{=?}(C > 0) = 0.555$
- $(A = 2) \rightarrow S_{>0.75}(B > 0) = true$
- $(A = 2) \rightarrow S_{>0.75}(C > 0) = false$
- $(A = 2) \rightarrow S_{<0.75}(C > 0) = true$

Reward-based properties are specified using the operator R in a similar fashion to the P and S operators. The following are some typical examples:

- $R_{=?}[I = 100]$ – "what is the expected reward value at exactly 100 seconds?";
- $R_{=?}[C \leq 24]$ – "what is the expected sum of rewards reachable within 24 seconds?";
- $R_{<10}[S]$ – "is the expected steady-state reward less than 10?".

Here I stands for instantaneous, C for cumulative, and S for steady-state reward value. The time units are, of course, model specific, and are usually interpreted in the context of each particular model.

This concludes our overview of PRISM. In Section 3.2 we model one biochemical reaction and evaluate some of the model's properties using PRISM. In Section 3.3 we build a stochastic model of a complex signalling pathway and demonstrate how the analysis can be performed using PRISM.

3.2 Modelling Single Reaction with CTMCs

In this section we demonstrate two approaches to modelling biochemical reactions using CTMCs. The first approach is based on modelling *individual*

molecules while the second one is based on modelling *populations*. In this section we demonstrate both modelling approaches with reference to the following example of a single biochemical reaction:

Example 3.8 (Single biochemical reaction: protein binding)

Consider a system consisting of one (reversible) biochemical reaction binding proteins A and B together to form a complex AB:

$$A + B \underset{k_2=0.5}{\overset{k_1=1}{\longleftrightarrow}} AB.$$

Assume that molecules of proteins A and B are uniformly distributed in one spatial compartment, the probability of binding proteins to form a complex AB is proportional to the probability that molecules will meet in the space, and the proportion factor is some affinity coefficient which corresponds to the likelihood of binding when molecules meet in space.

The probability that molecules will meet in space is proportional to the number of molecules. The proportionality coefficient is not usually considered separately from the affinity coefficient: they are both combined in a common kinetic rate.

There is also a probability that complex AB will decompose producing proteins A and B. The probability of this event is proportional to the likelihood of such an unbinding event.

The mean dynamic behaviour of a (reversible) binding reaction is traditionally modelled using the following system of differential equations:

$$\begin{cases} \dot{A} = -k_1 \cdot A \cdot B + k_2 \cdot AB, \\ \dot{B} = -k_1 \cdot A \cdot B + k_2 \cdot AB, \\ \dot{AB} = k_1 \cdot A \cdot B - k_2 \cdot AB. \end{cases} \tag{3.2}$$

Assume that initially we have an equal concentration of proteins A and B, both 10 millimolars[1] *(10 mmol/L), the compartment size is 1 millilitre. No complexes AB are present initially.*

Solving the initial value problem, we obtain the simulation trace depicted in Figure 3.4.

---

[1]The Molar (M) is a unit of concentration, or molarity, of solution equal to 1 mol/L.

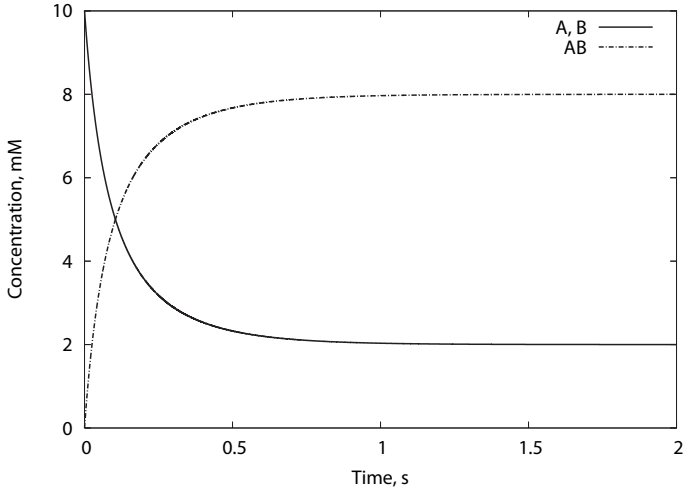


Figure 3.4: Simulation of a reversible binding reaction. $A|_{t=0} = 10$, $B|_{t=0} = 10$, $AB|_{t=0} = 0$.

3.2.1 Modelling Individual Molecules

Heath et al. (2006) propose the following approach for modelling interactions of individual molecules with CTMCs. In the case of modelling individual molecules, stochastic effects of individual molecule-molecule interactions form the basis of the model.

Consider Example 3.8 above. To model this system the following atomic propositions are defined:

- $A = 1$ – a molecule of protein A present,
- $A = 0$ – no molecules of A present,
- $B = 1$ – a molecule of protein B present,
- $B = 0$ – no molecules of B present,
- $AB = 1$ – a molecule of a complex AB present,
- $AB = 0$ – no complexes.

The system described in Example 3.8 can be modelled as individual molecules, with a CTMC consisting of two states: s_1 and s_2. The labelling for these states is the following:

- $L(s_1) = \{A = 1, B = 1, AB = 0\}$,

- $L(s_2) = \{A = 0, B = 0, AB = 1\}$.

Below is the PRISM model for this system. The rate of protein binding is λ_1; the rate of unbinding is λ_2. The variables `A`, `B` and `AB` denote the presence/absence of a molecule of A, B or the complex AB respectively. Initially, there is one molecule each of A and B, and none of AB.

```
ctmc

   rate k1 = λ1;
   rate k2 = λ2;

module binding
    A: [0..1] init 1;
    B: [0..1] init 1;
    AB:[0..1] init 0;

    [bind] (A = 1)&(B = 1) -> k1: (A' = 0)&(B' = 0)&(AB' = 1);
    [unbind] (AB = 1) -> k2: (A' = 1)&(B' = 1)&(AB' = 0);
endmodule
```

Assume that constants λ_1 and λ_2 are the kinetic rates defined in the original example (i.e. $\lambda_1 = 1$, and $\lambda_2 = 0.5$). This assumption is sound if we consider the system within a fixed volume of uniform size. Adjustments for system volume are required otherwise.

To calculate the probability that there will be a molecule of AB at time point t, we check the following property:

$$P{=}?(true\,\mathbf{U}_{[t,t]}(AB = 1))$$

The result of this formula over the range of t $[0, 10]$ is depicted in Figure 3.5.

We also employ *rewards* to evaluate the probability that there is molecule AB at time t. To do so we define the following reward:

```
rewards
    (AB > 0) : 1;
endrewards
```

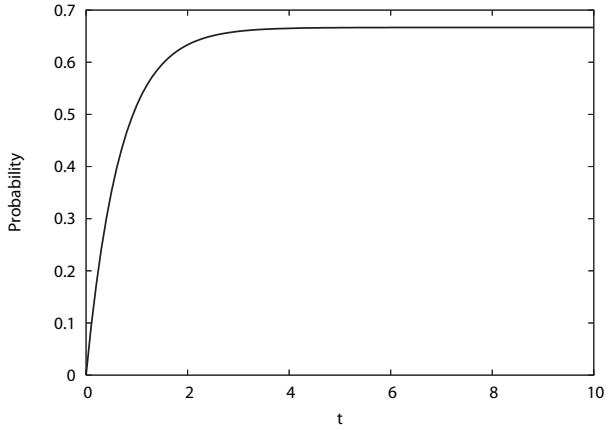


Figure 3.5: Probability that there is complex AB at time t

thus assigning a reward of 1 point to each state where $AB > 0$ (only s_2 in this case). The expected value for this reward is evaluated using the following property:

$$R_{=?}[I = t].$$

The result obtained with this experiment is precisely equivalent to the one depicted in Figure 3.5. Thus we can employ any of the methods to compute such probabilities.

Figure 3.5 suggests that the probability that there is a molecule of the complex initially rises, and then stays at the same level. This suggests that the system reaches an equilibrium state, which is indeed one of the features of this system.

The concentrations of proteins and complexes can be determined from the model of individual molecules using the following rewards (by scaling the number of molecules to concentrations with a factor of 10, because the maximal concentration possible is the model is 10 M.):

- `true:(A*10);` for protein A,
- `true:(B*10);` for protein B,
- `true:(AB*10);` for complex AB.

The concentration of complex AB is compared to the one found by simulation of the ODEs in Figure 3.6. Notice, however, that the concentration produced with these rewards does not match the behaviour simulated with the ODE model

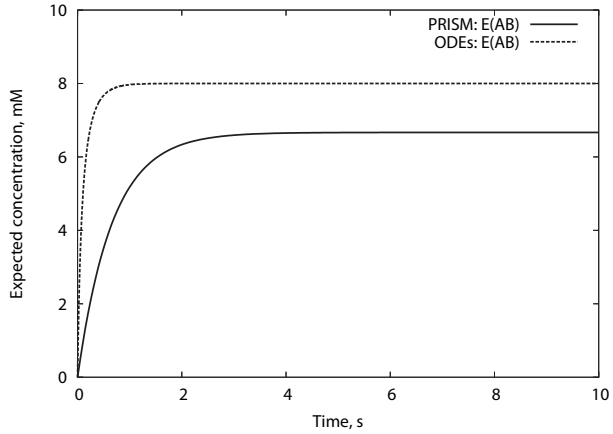


Figure 3.6: Predicted concentration of complex AB.

(introduced in Example 3.8 and originally depicted in Figure 3.4). We propose that this is due to the lack of the *population* dynamics in the molecule-based model. Namely, the rates of actions do not take the numbers of molecules into account.

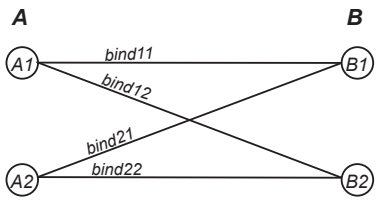


Figure 3.7: Possible binding events for two molecules of protein A and B.

Below, we demonstrate how to represent population dynamics in the case of a small number of molecules. Consider the case of two molecules. Each molecule of protein A can bind to any of the molecules of protein B, thus we consider four possible binding events (see Figure 3.7). Initially, all four binding events are enabled, thus a racing condition exists in the corresponding CTMC; reflecting the impact of population size onto the dynamics of this reaction. Example 3.9 defines a model with two molecules of A, B and AB each:

Example 3.9 (PRISM model of a small population of individual molecules)

```
ctmc

rate k1 = λ1;
rate k2 = λ2;
```

```
module A1
A1:[0..1] init 1;

[bind11] (A1=1) -> k1:(A1'=0);
[bind12] (A1=1) -> k1:(A1'=0);
[unbind11] (A1=0) -> 1:(A1'=1);
[unbind12] (A1=0) -> 1:(A1'=1);
endmodule

module A2
A2:[0..1] init 1;

[bind21] (A2=1) -> k1:(A2'=0);
[bind22] (A2=1) -> k1:(A2'=0);
[unbind21] (A2=0) -> 1:(A2'=1);
[unbind22] (A2=0) -> 1:(A2'=1);
endmodule

module B1
B1:[0..1] init 1;

[bind11] (B1=1) -> 1:(B1'=0);
[bind21] (B1=1) -> 1:(B1'=0);
[unbind11] (B1=0) -> 1:(B1'=1);
[unbind21] (B1=0) -> 1:(B1'=1);
endmodule

module B2
B2:[0..1] init 1;

[bind12] (B2=1) -> 1:(B2'=0);
[bind22] (B2=1) -> 1:(B2'=0);
[unbind12] (B2=0) -> 1:(B2'=1);
[unbind22] (B2=0) -> 1:(B2'=1);
endmodule
```

```
module A1B1
A1B1:[0..1] init 0;

[bind11](A1B1=0) -> 1:(A1B1'=1);
[unbind11](A1B1=1) -> k2:(A1B1'=0);
endmodule

module A1B2
A1B2:[0..1] init 0;

[bind12](A1B2=0) -> 1:(A1B2'=1);
[unbind12](A1B2=1) -> k2:(A1B2'=0);
endmodule

module A2B1
A2B1:[0..1] init 0;

[bind21](A2B1=0) -> 1:(A2B1'=1);
[unbind21](A2B1=1) -> k2:(A2B1'=0);
endmodule

module A2B2
A2B2:[0..1] init 0;

[bind22](A2B2=0) -> 1:(A2B2'=1);
[unbind22](A2B2=1) -> k2:(A2B2'=0);
endmodule
```

We can analyse the transient probabilities concerning quantities of molecules A, B or AB at any given time using the following CSL properties:

- $P_{=?}((true)\mathbf{U}_{[t,t]}(A_1 = 1)|(A_2 = 1))$ - at least one molecule of A is present at time t;
- $P_{=?}((true)\mathbf{U}_{[t,t]}(B_1 = 1)|(B_2 = 1))$ - at least one molecule of B is present at time t;

- $P_{=?}((true)\mathbf{U}_{[t,t]}(A_1B_1 = 1)|(A_1B_2 = 1)|(A_2B_1 = 1)|(A_2B_2 = 1))$ - at least one molecule of AB is present at time t;

The evaluation results for these properties are depicted in Figure 3.8.

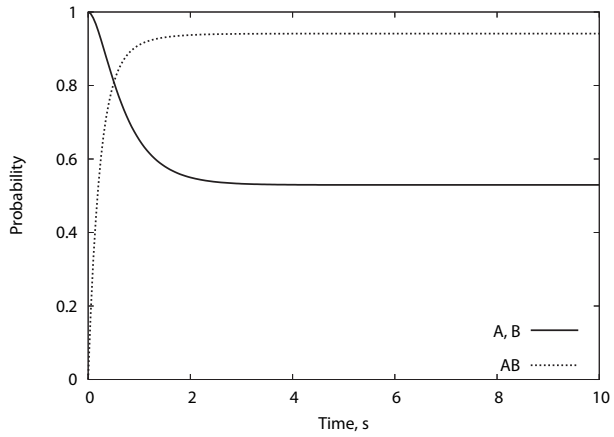


Figure 3.8: Transient probabilities for a CTMC model of individual molecules. The solid line corresponds to the probability that at least one molecule of A and B (out of two available) is present at a certain time. Initially, this probability equals 1, but then the probability decreases, until stabilising between 0.5 and 0.6. The dashed line corresponds to the analogous property for complex AB. Initially, probability equals 0, and then it increases and stabilises between 0.9 and 1.

We can also compute the concentrations of proteins using rewards. For example, for AB such a reward is

```
rewards
    (true):(A1B1 + A1B2 + A2B1 + A2B2)*5;
endrewards
```

Now, we use the scaling factor of 5 in order to map the maximal amount of complex molecules (2 in this case) on to the allowed range of concentrations (0 to 10 Molars). The computed (PRISM) concentration plot is depicted in Figure 3.9. This time the result is closer to the simulation trace produced with ODEs (introduced in Example 3.8 and originally depicted in Figure 3.4), but there is still a significant difference as the population dynamics is simulated by the racing condition between parallel binding events. The quality of this approximation will improve if a model with a larger population of molecules is used. We investigated how the results change when using larger populations of

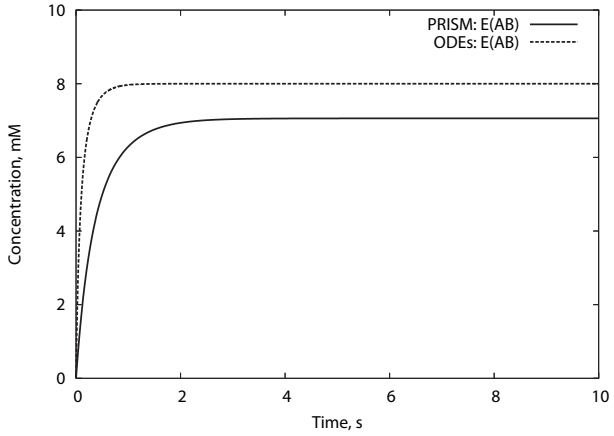


Figure 3.9: Comparison of protein concentrations obtained with a model of individuals using PRISM rewards with ODE results.

molecules by building a programme which generates individual-based models for a given size of molecules population.

Figure 3.10 demonstrates how the result for the concentration of AB approaches the simulation trace generated with the ODE model as the number of molecules increases.

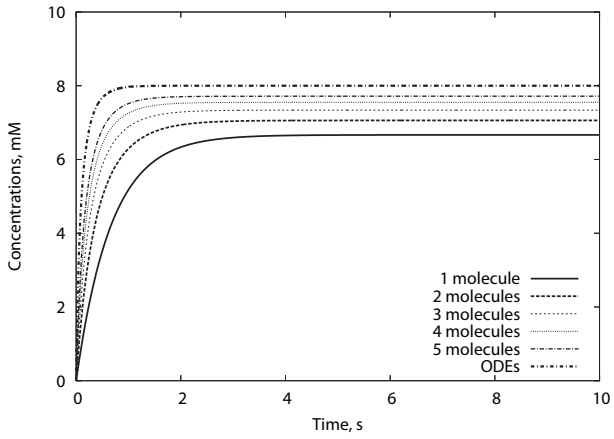


Figure 3.10: Estimates generated from CTMC models with different number of molecules.

The size of the CTMCs, however, increases dramatically when modelling larger populations. Table 3.1 lists sizes of CTMCs for different models of Example 3.8. Notice that models of 6 or 7 molecules per species fitted into computer's memory, however, we could not calculate the estimated concentrations as reward computations became intractable (e.g. took more than 3 days on a state of art desktop computer).

Number of Molecules per species	***States***	***Transitions***
1	2	2
2	7	16
3	34	126
4	209	1088
5	1546	10450
6	13327	111312
7	130922	1306046
8	Does not fit to computer's memory	

Table 3.1: Sizes of CTMCs for models of individual molecules.

Models of larger populations are intractable for both the description (because a separate variable, or a module is required for each molecule, and a separate command is required for each possible binding event) and computational experiments.

We conclude that due to complexity limitations (state space and time) it is problematic to employ the individual-based modelling approach of Heath et al. (2006) for systems with large populations of molecules when population dynamics plays an important role. However, this approach may be suitable when modelling systems consisting of few molecules, where population dynamics does not serve an important role in shaping the system behaviour.

We propose that the scalability problem can be resolved by adopting a population-based approach as defined in the following section (see Section 3.2.2).

3.2.2 Population-Based Modelling

To resolve the scalability problem demonstrated in the previous section we propose an approach which takes the population dynamics directly into account. This means considering concentrations rather than individual molecules. We extend an approach published in (Calder, Gilmore and Hillston 2006) which was implemented in PEPA. Instead of considering two symbolic levels of concentration ("high" and "low" in (Calder, Gilmore and Hillston 2006)) we propose to use multiple intervals of the concentration. Additionally, we propose a new form of analysis of such models using probabilistic symbolic model checking. We reported some of this work in (Calder, Vyshemirsky, Gilbert and Orton 2006).

The key concept is the introduction of discrete concentrations for each bio-

chemical species.

Each species has a concentration which changes with time, i.e. $m = f(t)$, where m is a concentration of the species and t is time. We make discrete abstractions as follows. When the maximum concentration is M, then for a given N, the abstract values $0 \ldots N$ represent the concentration intervals $[0, 1 * M/N), [1 * M/N, 2 * M/N), \ldots [(N-1) * M/N, N * M/N]$. We refer to $0 \ldots N$ as *levels* of concentration. We note that we could define a different N for each species but in this chapter, without loss of generality, we assume the same N, for all species.

To model a biochemical system we associate a concurrent, computational process with each of the proteins in the network and define these processes by PRISM modules. We note that this description could be produced automatically from a topological description of the pathway. Consider a model of a binding reaction (Example 3.10) described in Example 3.8.

Example 3.10 (Population model for Example 3.8.)

```
ctmc

const double M=10.0;
const int N=3;
const double L=M/N;

rate k1=1.0;
rate k2=0.5;

module A
    A: [0..N] init N;
    [bind] (A>0) -> A*L:  (A' = A - 1);
    [unbind] (A<N) -> 1: (A' = A + 1);
endmodule

module B
    B: [0..N] init N;
    [bind] (B>0) -> B*L: (B' = B - 1);
    [unbind] (B<N) -> 1: (B' = B + 1);
```

```
endmodule

module AB
    AB: [0..N] init 0;
    [bind] (AB < N) -> 1: (AB' = AB + 1);
    [unbind] (AB > 0) -> AB*L: (AB' = AB - 1);
endmodule

module Constants
    x: bool init true;
    [bind] (x) -> k1/L: (x'=true);
    [unbind] (x) -> k2/L: (x'=true);
endmodule

rewards
   (true) : AB*L;
endrewards
```

The model begins with the keyword `ctmc`, consists of some preliminary constants:

- `M` – the maximal (continuous) concentration of any species,
- `N` – the number of discrete abstract concentration levels,
- `L` – the length of each interval, an abbreviation for M/N;

and four modules: `A`, `B`, `AB`, and `Constants`. Consider the first three modules, representing species A, B, and AB. Each module has the form: a state variable which denotes the species concentration (we use the same name for process and variable, the type can be deduced from the context) followed by commands with actions `[bind]` and `[unbind]`.

In order to define the rates of the transitions, we distinguish species which are spent in a reaction (reactants) from species which are produced in a reaction (products). Since the transitions are synchronised on common actions, the rate of each transition will be a product of rates defined in individual modules. We assign these rates to the concentration of reactants (in corresponding modules)

multiplied by the scaling coefficient `L`, which corresponds to the length of the discrete interval. The rates in modules which correspond to products are set to 1.

There is a fourth module, `Constants`. This module simply defines the coefficients for reaction kinetics. This module uses a "dummy" state variable called `x`, and (two) always enabled transitions with actions `bind` and `unbind`.

The assignments for the transitions of reactants decrease the value of corresponding variables by 1, thus decreasing corresponding species' concentration. The assignments for products increase corresponding variables by 1.

All the transitions with the same action are synchronised. For example, the resulting transition from the initial state for action `bind` has rate:

$$\lambda = \frac{N \cdot L \cdot N \cdot L \cdot k_1}{L} = 30. \tag{3.3}$$

The resulting CTMC consists of 4 states and 6 transitions. The schematic representation of this CTMC is depicted in Figure 3.11.

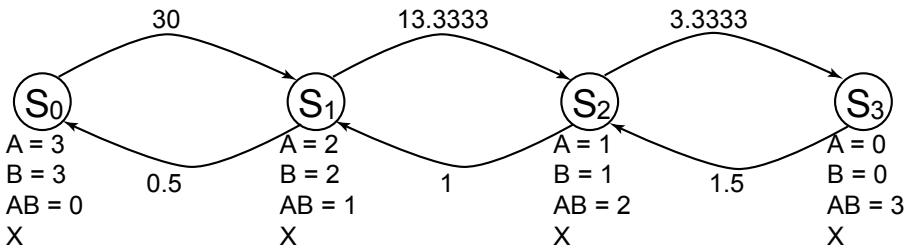


Figure 3.11: CTMC for model from Example 3.10.

Consider the soundness of this modelling approach. The crucial question is *how do the transition rates compare with, or relate to, the binding kinetics explained in Example 3.8?*

First, consider how the variables relate to each other: the system of differential equations (3.2) in Example 3.8 refers to continuous concentrations, whereas the PRISM model operates on *discrete* natural concentration levels. Let m_i be a continuous variable (e.g. $m_1 = A$, $m_2 = B$ and $m_3 = AB$ in (3.2)). and let m_i^d be the corresponding PRISM variable (e.g. `A`, `B`, `AB`). Then

$$m_i = m_i^d \cdot L = m_i^d \cdot \frac{M}{N} \tag{3.4}$$

Second, derive a rate expressed in terms of the PRISM variables. From the

continuous rate:

$$\frac{dm_3}{dt} = k_1 \cdot m_1 \cdot m_2 \tag{3.5}$$

the simplest way to derive a new concentration m_3' from m_3 is by Euler's method thus:

$$m_3' = m_3 + (k \cdot m_1 \cdot m_2 \cdot \Delta t) \tag{3.6}$$

But the discrete (PRISM) concentrations can only increase in units of 1 level of concentration, or $\frac{M}{N}$ molars, so the time to perform such concentration change is:

$$\Delta t = \frac{M}{k_1 \cdot m_1 \cdot m_2 \cdot N} \tag{3.7}$$

PRISM implements rates as "memoryless" negative exponential, that is for given rate λ, $P(t) = 1 - e^{-\lambda t}$ is the probability that the transition will be taken before time t. Taking λ as $\frac{1}{\Delta t}$, in this example we have

$$\lambda = \frac{k_1 \cdot m_1 \cdot m_2 \cdot N}{M} \tag{3.8}$$

Replacing the continuous variables by their discrete forms, we have

$$\lambda = \frac{k_1 \cdot (m_1^d \cdot \frac{M}{N}) \cdot (m_2^d \cdot \frac{M}{N}) \cdot N}{M} \tag{3.9}$$

or

$$\lambda = \frac{m_1^d \cdot L \cdot m_2^d \cdot L \cdot k_1}{L} \tag{3.10}$$

which in an initial state of Example 3.10 is

$$\lambda = \frac{N \cdot L \cdot N \cdot L \cdot k_1}{L} = 30, \tag{3.11}$$

which is exactly the rate given in (3.3).

Now consider simulation of behaviour predictions using these population-based models. In Example 3.10, we included the factor (`/L`) in the `Constants` module, and multiply the concentrations by the scaling factor `L` in the protein processes. The following reward

```
(true) : AB*L;
```

can be used to determine the expected concentration of complex AB. The scaling factor `L` is used again to convert the discrete concentration to the scale

of continuous concentrations. The estimated concentration of AB for different values of N are compared with the simulation trace produced with ODEs in Figure 3.12.

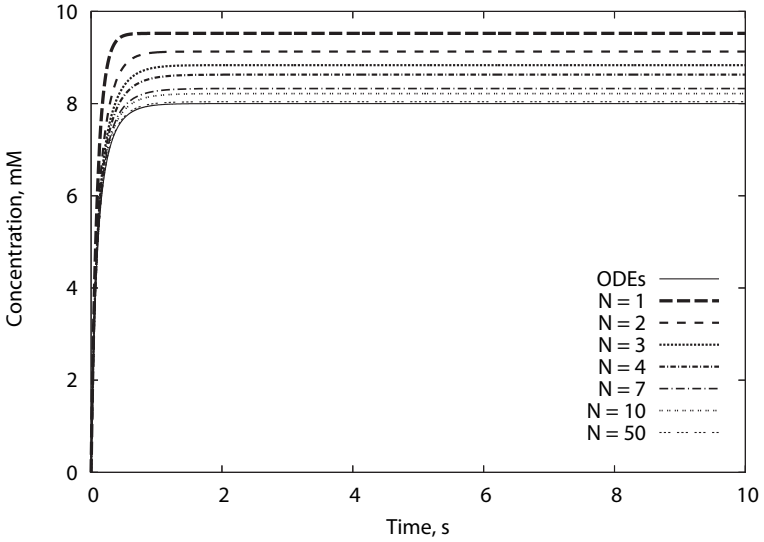


Figure 3.12: Model behaviour predictions for different values of N. Note, that in comparison to Figure 3.10, the predicted behaviours here provide better approximations to the behaviour simulated with ODEs.

The size of CTMCs when using this approach depends linearly on the number of intervals employed. Table 3.2 lists sizes of CTMCs for different population-based models of Example 3.8.

N	***States***	***Transitions***
1	2	2
2	3	4
3	4	6
4	5	8
7	8	14
10	11	20
50	51	100
200	201	400

Table 3.2: Sizes of CTMCs for population-based models.

This illustrates how, using a population-based approach, we have overcome the problem of model complexity. At the same time we have managed to achieve a more precise simulation of population dynamics. In the next section (see Section 3.3) we apply this methodology to a more complex biochemical pathway.

3.3 Modelling a Pathway with CTMCs

In this section we consider modelling and analysing the RKIP inhibited ERK pathway. Here we give only a brief overview of the pathway structure, further details are presented in (Cho et al. 2003).

This pathway is a ubiquitous pathway that conveys mitogenic and differentiation signals from the cell membrane to the nucleus.

The kinase inhibitor protein RKIP inhibits activation of Raf and we conjecture that it can reduce the strength of the signal passing through the pathway.

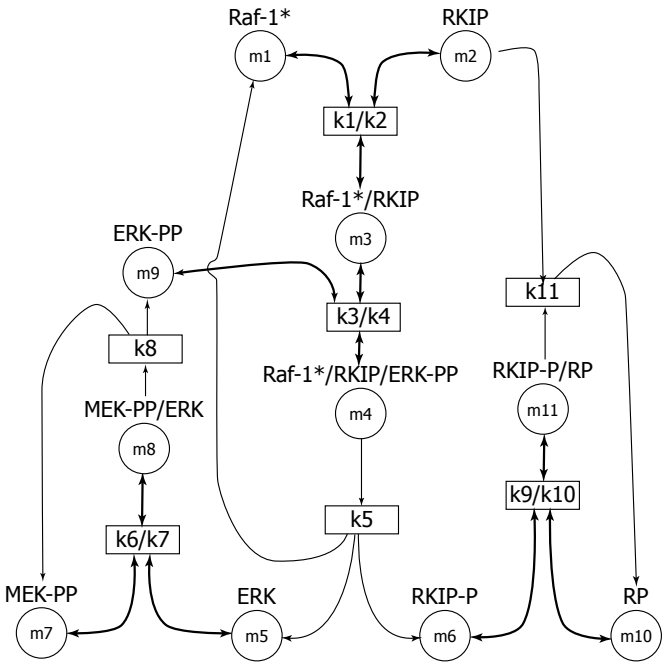


Figure 3.13: The RKIP inhibited ERK pathway

We consider the pathway as given in the graphical representation of Figure 3.13. This figure is taken from (Cho et al. 2003), where a number of nonlinear ordinary differential equations (ODEs) representing the kinetics are given. We take Figure 3.13 as our starting point, and explain informally, its meaning. Each node is labelled by a protein (or species). For example, Raf-1*, RKIP and Raf-1*/RKIP are proteins, the last being a complex built up from the first two. A suffix -P or -PP denotes a (single or double, resp.) phosphorylated protein, for example RKIP-P and ERK-PP. Each protein has an associated concentration, given by m1, m2 etc. Reactions define how proteins are built up and broken down. Propagation of a signal corresponds to the "wave" of binding/unbinding

events. In Figure 3.13, bidirectional arrows correspond to both forward and backward reactions; unidirectional arrows to forward reactions. Each reaction has a rate given by the rate constants $k1$, $k2$, etc. These are given in the rectangles, with $kn/kn+1$ denoting that kn is the forward rate and $kn+1$ the backward rate. Initially, all concentrations are equal to zero, except for $m1$, $m2$, $m7$, $m9$, and $m10$ (Cho et al. 2003). The dynamic behaviour of the pathway is quite complex, because proteins are involved in more than one reaction and there are several feedbacks. We note that the example system is part of a larger pathway which can be found elsewhere (e.g Kholodenko et al. 1999 Schoeberl et al. 2002).

ERK-PP is a protein which is capable of entering the cell nucleus, and for this reason is usually considered the "output" of the signalling pathway.

We model this biochemical pathway employing the population-based approach described in Section 3.2.2.

Example 3.11 (PRISM model of the RKIP inhibited ERK pathway.)

```
ctmc

const double M=3.0;
const int N=12;
const int I1=floor((2.5*N)/3.0);
const int I2=N;
const double L=M/N;

rate k1=0.53;
rate k2=0.0072;
rate k3=0.625;
rate k4=0.00245;
rate k5=0.0315;
rate k6=0.8;
rate k7=0.0075;
rate k8=0.071;
rate k9=0.92;
rate k10=0.00122;
rate k11=0.87;
```

```
module RAF1
RAF1: [0..I1] init I1;
[r1] (RAF1 > 0) -> RAF1*L: (RAF1' = RAF1 - 1);
[r2] (RAF1 < I1) -> 1: (RAF1' = RAF1 + 1);
[r5] (RAF1 < I1) -> 1: (RAF1' = RAF1 + 1);
endmodule

module RKIP
RKIP: [0..I1] init I1;
[r1] (RKIP > 0) -> RKIP*L: (RKIP' = RKIP - 1);
[r2] (RKIP < I1) -> 1: (RKIP' = RKIP + 1);
[r11] (RKIP < I1) -> 1: (RKIP' = RKIP + 1);
endmodule

module RAF1RKIP
RAF1RKIP: [0..I1] init 0;
[r1] (RAF1RKIP < I1) -> 1: (RAF1RKIP' = RAF1RKIP + 1);
[r2] (RAF1RKIP > 0) -> RAF1RKIP*L:(RAF1RKIP' = RAF1RKIP - 1);
[r3] (RAF1RKIP > 0) -> RAF1RKIP*L:(RAF1RKIP' = RAF1RKIP - 1);
[r4] (RAF1RKIP < I1) -> 1: (RAF1RKIP' = RAF1RKIP + 1);
endmodule

module ERKPP
ERKPP: [0..I1] init I1;
[r3] (ERKPP > 0) -> ERKPP*L: (ERKPP' = ERKPP - 1);
[r4] (ERKPP < I1) -> 1: (ERKPP' = ERKPP + 1);
[r8] (ERKPP < I1) -> 1: (ERKPP' = ERKPP + 1);
endmodule

module RAF1RKIPERKPP
RAF1RKIPERKPP: [0..I1] init 0;
[r3] (RAF1RKIPERKPP < I1) -> 1:(RAF1RKIPERKPP' = RAF1RKIPERKPP + 1);
[r4] (RAF1RKIPERKPP > 0) ->RAF1RKIPERKPP*L:
                            (RAF1RKIPERKPP' = RAF1RKIPERKPP - 1);
[r5] (RAF1RKIPERKPP > 0) ->RAF1RKIPERKPP*L:
                            (RAF1RKIPERKPP' = RAF1RKIPERKPP - 1);
```

```
endmodule

module ERK
ERK: [0..I1] init 0;
[r5] (ERK < I1) -> 1: (ERK' = ERK + 1);
[r6] (ERK > 0) -> ERK*L: (ERK' = ERK - 1);
[r7] (ERK < I1) -> 1: (ERK' = ERK + 1);
endmodule

module RKIPP
RKIPP: [0..I1] init 0;
[r5] (RKIPP < I1) -> 1: (RKIPP' =RKIPP + 1);
[r9] (RKIPP > 0) -> RKIPP*L: (RKIPP' =RKIPP - 1);
[r10] (RKIPP < I1) -> 1: (RKIPP' =RKIPP + 1);
endmodule

module RP
RP: [0..I2] init I2;
[r9] (RP > 0) -> RP*L: (RP' = RP - 1);
[r10] (RP < I2) -> 1: (RP' = RP + 1);
[r11] (RP < I2) -> 1: (RP' = RP + 1);
endmodule

module MEKPP
MEKPP: [0..I1] init I1;
[r6] (MEKPP > 0) -> MEKPP*L: (MEKPP' = MEKPP - 1);
[r7] (MEKPP < I1) -> 1: (MEKPP' = MEKPP + 1);
[r8] (MEKPP < I1) -> 1: (MEKPP' = MEKPP + 1);
endmodule

module MEKPPERK
MEKPPERK: [0..I1] init 0;
[r6] (MEKPPERK < I1) -> 1: (MEKPPERK' = MEKPPERK + 1);
[r7] (MEKPPERK > 0) -> MEKPPERK*L:(MEKPPERK' = MEKPPERK - 1);
[r8] (MEKPPERK > 0) -> MEKPPERK*L:(MEKPPERK' = MEKPPERK - 1);
endmodule
```

```
module RKIPPRP
RKIPPRP: [0..I1] init 0;
[r9] (RKIPPRP < I1) -> 1: (RKIPPRP' = RKIPPRP + 1);
[r10] (RKIPPRP > 0) -> RKIPPRP*L: (RKIPPRP' = RKIPPRP - 1);
[r11] (RKIPPRP > 0) -> RKIPPRP*L: (RKIPPRP' = RKIPPRP - 1);
endmodule

module Constants
x: bool init true;
[r1] (x) -> k1/L: (x' = true);
[r2] (x) -> k2/L: (x' = true);
[r3] (x) -> k3/L: (x' = true);
[r4] (x) -> k4/L: (x' = true);
[r5] (x) -> k5/L: (x' = true);
[r6] (x) -> k6/L: (x' = true);
[r7] (x) -> k7/L: (x' = true);
[r8] (x) -> k8/L: (x' = true);
[r9] (x) -> k9/L: (x' = true);
[r10] (x) -> k10/L: (x' = true);
[r11] (x) -> k11/L: (x' = true);
endmodule
```

The model defined in Example 3.11 uses special constants `I1` and `I2` to tackle different maximal concentration for different species. For example, the original model (see Cho et al. 2003) of this pathway states that the initial concentration for *RP* is 3.0 Molars, while the initial concentrations for *Raf-1**, *RKIP*, *ERK-PP* and *MEK-PP* are 2.5 Molars. Thus, we define `I1` as an integer part of $\frac{2.5 \cdot N}{3.0}$ which is the discrete value for 2.5 Molars, and `I2` equal to N which is the discrete value for 3.0 Molars.

We use PRISM rewards to evaluate the estimates for protein concentrations. The following reward:

```
rewards
   true: ERKPP*L;
endrewards
```

produces the estimate for the MEK-PP concentration. To compare simulation results between this stochastic model and the deterministic model defined by the

system of differential equations (3.2), consider the concentration of *MEK-PP*, over the time interval $[0 \ldots 100]$. Concentration is the vertical axis. Figure 3.14 demonstrates the results, using the ODE model and two instances of our stochastic model, with $N = 6$ and $N = 12$. The "upper" curve is the ODE simulation, the "lower" curve is the stochastic simulation, when $N = 3$; the curve in between the two is the stochastic behaviour when $N = 7$. As N increases, the closer the plots; with $N = 12$ the difference is barely discernible.

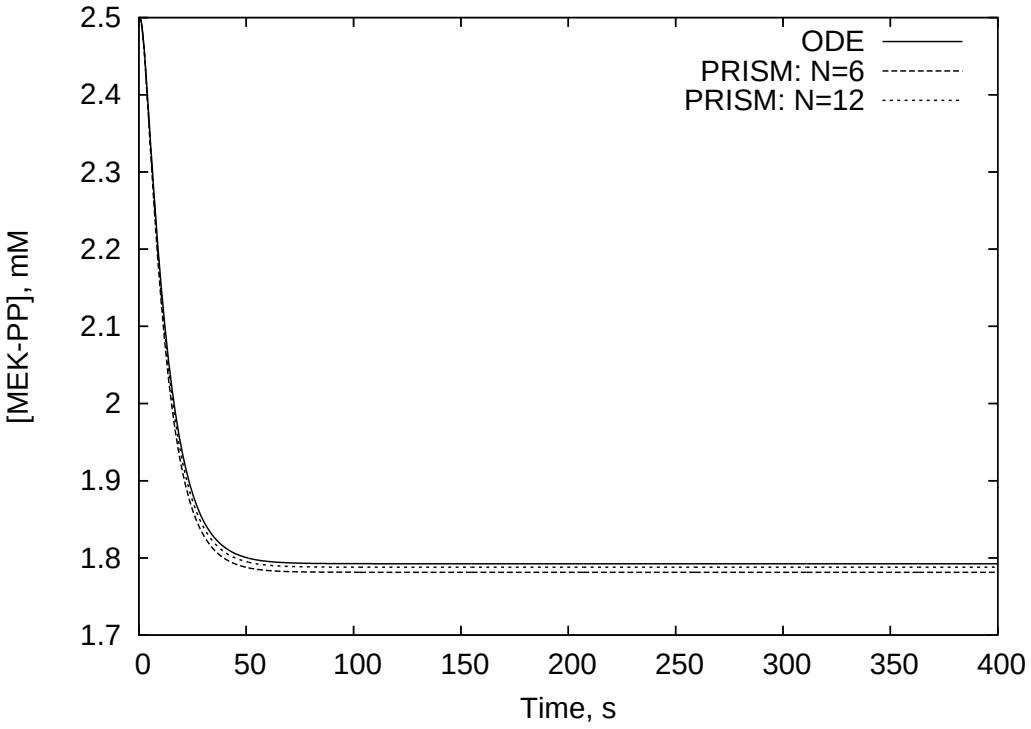


Figure 3.14: Concentration estimates for *ERK-PP*

In this model the size of the CTMCs depends entirely on the number of abstract concentration levels. This is due to concurrency and interleaving in this model, e.g. binding of *Raf-1** to *RKIP* can either precede or succeed binding of *RKIP-P* to *RP*. The sizes of CTMCs for different numbers of concentration levels are given in Table 3.3.

While it is very interesting to see that a stochastic model produces the behaviour very similar to the one produced with the deterministic ODE model, even when employing quite a small number of concentration levels, the primary motivation for this methodology is logical analysis, with respect to temporal logic properties.

N	***States***	***Transitions***
2	13	30
3	73	276
6	1974	12236
8	4326	28896
10	16071	118932
12	47047	372372
16	175644	1485848

Table 3.3: Sizes of CTMCs for models of the RKIP inhibited ERK pathway.

Temporal logics are powerful tools for expressing temporal queries which may be generic (e.g. state reachability, deadlock) or application specific (e.g. referring to variables representing application characteristics). Whereas simulation is the exploration of a *single* behaviour over a given time interval, model checking allows us to investigate the truth (or otherwise) of temporal queries over (possibly infinite) sets of behaviours over (possibly) unbounded time intervals.

For this example, we consider three different kinds of temporal property:

1. steady state analysis of stability of a protein, i.e. a protein reaches a level and then remains there, within certain bounds,

2. steady state analysis of protein stability when varying reaction rates, i.e. a protein is more likely to be stable for certain reaction rates,

3. transient analysis of protein activation sequence, i.e. concentration peak ordering.

3.3.1 Stability of Protein in Steady State

This type of property is particularly applicable to the analysis of networks where temporary and sustained signal responses can produce markedly different cellular outcomes. For example, a transient signal could lead to cell proliferation, whereas a sustained signal would result in differentiation.

Consider the concentration of *Raf-1**. Stability for this protein (at level D) is expressed by the CSL formula:

$$S_{=?}[(RAF1 = D)] \tag{3.12}$$

The results are given Figure 3.15, with D ranging over $0 \ldots 12$ ($N = 12$). They illustrate that *Raf-1** is most likely to be stable at level 1, with a relatively high probability of stability at levels 0, 2 or 3. It is unlikely to be stable at levels 4 or more.

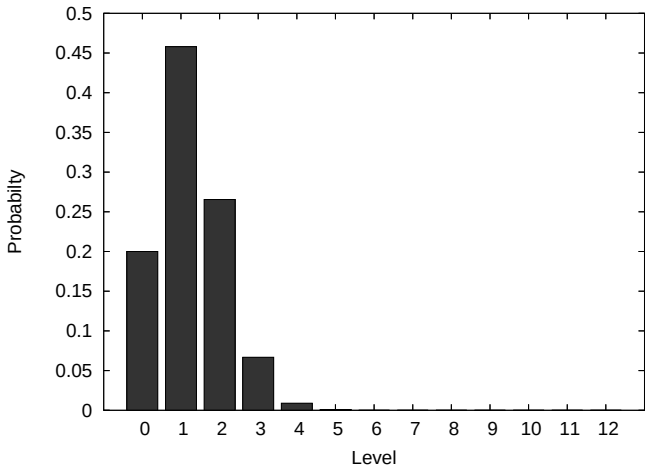


Figure 3.15: Stability of *Raf-1** at level D in steady state.

3.3.2 Protein Stability in Steady State while Varying Rates

This type of property is particularly useful during model fitting, i.e. fitting the model to experimental data. As an example, consider evaluating the probability that *Raf-1** is stable in concentrations interval $[0.25 \ldots 0.5]$ mM (discrete level 2) in the steady state, whilst varying the rate of the reaction $r1$ (the reaction which binds *Raf-1** and *RKIP*). We vary the parameter k_1 (which determines the rate of $r1$) over the interval $[0 \ldots 1]$. The stability property is expressed by:

$$S_{=?}[RAF1 = 2] \tag{3.13}$$

Consider also the probability that *Raf-1** is stable at level 1 ($[0 \ldots 0.25]$ mM); the formula for this is:

$$S_{=?}[RAF1 = 1] \tag{3.14}$$

Figure 3.16 depicts results for both these properties, when $N = 12$. The probability density of property (3.13) (solid line) peaks at $k_1 = 0.2$ and then decreases; the probability density of property (3.14) (dashed line) increases dramatically, reaching a maximum when $k_1 > 0.6$.

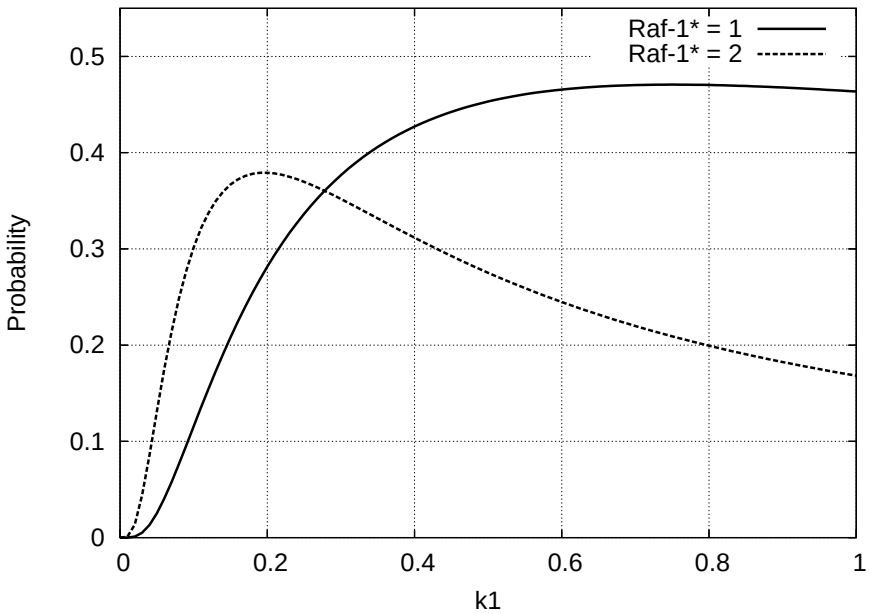


Figure 3.16: Stability of *Raf-1** for different values of rate $k1$

3.3.3 Activation Sequence Analysis

The last example illustrates queries over several proteins: sequences of protein activations. Consider two complexes: *Raf-1*/RKIP* and *Raf-1*/RKIP/ERK-PP*. Is it possible that the (concentration of the) former "peaks" before the latter?

Let M be the peak level of *Raf-1*/RKIP/ERK-PP*, and D be the level of *Raf-1*/RKIP*. The formula for this property is:

$$P_{=?}[(RAF1RKIPERKPP < M)\mathbf{U}(RAF1RKIP = D)] \tag{3.15}$$

This property expresses *"What is the probability that the concentration of Raf-1*/RKIP/ERK-PP does not exceed level M, until Raf-1*/RKIP reaches concentration level D?"* The results of this query, for $N = 6$, D ranging over $\{1, 2\}$ and M ranging over $\{1, 2, 3, 4, 5\}$ are given in Table 3.4. For example, the probability *Raf-1*/RKIP* reaches concentration level 2 before *Raf-1*/RKIP/ERK-PP* reaches concentration level 5 is 98.7%, the probability *Raf-1*/RKIP* reaches concentration level 2 before *Raf-1*/RKIP/ERK-PP* reaches concentration level 2 is 90.9%.

To confirm these results, we conducted the inverse experiment – *is it possible*

	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
$D = 1$	100%	100%	100%	100%	100%
$D = 2$	73.1%	90.9%	95.9%	97.5%	98.7%

Table 3.4: Protein activation sequence (Property (3.15)).

for Raf-1/RKIP/ERK-PP to reach concentration level 5* before *Raf-1*/RKIP reaches concentration level 2 with* $N = 6$*?* The property is:

$$P_{=?}[(RAF1RKIP < D)\mathbf{U}(RAF1RKIPERKPP = M)] \tag{3.16}$$

This property expresses *"What is the probability that the concentration of Raf-1*/RKIP is less than level D until Raf-1*/RKIP/ERK-PP reaches concentration level M?"* The results are given in Table 3.5 which is complementary to Table 3.4: for example, the probability *Raf-1*/RKIP/ERK-PP* reaches concentration level 5 before *Raf-1*/RKIP* reaches concentration level 2 is 1.39%. This confirms the results obtained with property (3.15).

	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$
$D = 1$	0%	0%	0%	0%	0%
$D = 2$	26.9%	9.1%	4.1%	2.5%	1.3%

Table 3.5: Inverse protein activation sequence (Property (3.16)).

3.4 Discussion: Population-Based Modelling and Deductive Reasoning

We have described a new population-based modelling and quantitative deductive analysis approach for signal transduction networks. We model the dynamics of networks by continuous time Markov chains, making discrete approximations to concentrations. We describe the models in the high level PRISM modelling language: proteins are synchronous processes and concentrations are state variables. We have illustrated our approach with an example, the RKIP inhibited ERK pathway, a pathway previously modelled by ODEs (Cho et al. 2003).

The PRISM model checker has been a useful tool for model checking, experimentation, and even simulation. All computations have been tractable on a single standard processor (the times are trivial and have been omitted).

The main advantage of our approach is that using a continuous time stochastic logic and the PRISM model checker, we can perform quantitative analysis such as *what is the probability that a protein reaches a stable concentration level?* and *how does varying a reaction rate affect that probability?* The approach offers considerably more expressive power than simulation or qualitative analysis. We can also perform standard simulations and we have compared our results with traditional ordinary differential equation-based (simulation) methods. An interesting and useful result is that in the example pathway, only a small number of discrete data values is required to render the simulations practically indistinguishable.

In Section 3.2.2 we showed how our population based PRISM model relates to mass action kinetics as defined by ODEs. While simulation is not the primary goal of our approach, in Section 3.3 we demonstrated that with small N, our model provides (more than) sufficient simulation accuracy, for the example system. This is because the example pathway has reactions which are all on a similar scale. If we were to apply our approach to a pathway where the changes of concentrations are on different scales, i.e. the corresponding ODE model is a set of stiff equations, then we could still reason about the stochastic model using temporal logic queries. However, simulations would not be as accurate, for small N. If more accuracy of simulation was required, then we would have to increase N.

We demonstrated how deductive reasoning using temporal logic properties can be performed to show that the pathway model considered in Section 3.3 reaches the steady state, and to verify that one of the proteins (*Raf-1*/RKIP*) is likely to "peak" earlier than another one (*Raf-1*/RKIP/ERK-PP*).

3.5 Model Analysis: from Reasoning to Inference

In this chapter we have developed a *deductive* reasoning approach to analyse model behaviour. We have assumed that the structure of a model and particular quantitative kinetic parameters are known: the aim of the reasoning is to derive model behaviours and deduce properties of such behaviours. We formulated models using Continuous Time Markov Chains (defined with a high level language), and expressed properties using CSL.

In the next chapter we consider an *inferential* approach which solves the inverse problem: given observed behaviour, we search for suitable model parameters and compare alternative model topologies. Models will be formulated using ODEs, since these are traditional models for time course data.

Summary

In this chapter we considered two modelling approaches for quantitative reasoning about biological systems. Both approaches use continuous time Markov chains. Initially, we considered an approach proposed by Heath et al. (2006) that is based on modelling individual molecules. We have demonstrated that while being suitable for some cases when only a few molecules are involved in a model (such as Protein-DNA interaction models) this approach is inadequate for modelling systems where any impact of population dynamics is important. We tried to fix this flaw by explicitly defining small populations of molecules. We have shown how several molecules can be modelled using this approach, however this solution does not appear to be practical, as it leads to a well known state space explosion problem.

To resolve these problems, we proposed a novel population-based modelling approach which employs abstract discrete concentration levels. These provide approximations for continuous concentration values. We compared the simulated model behaviours to the solutions of the traditional ODE models, and found that even a small number of abstract concentration levels can be sufficient for a satisfactory simulation of biochemical system dynamics.

The main advantage of the proposed modelling approach is the ability to apply model checking techniques to enable logical reasoning. We have demonstrated how this can be done on a case study of the RKIP inhibited ERK pathway by proposing and verifying a number of logical properties described with Continuous Stochastic Logic.

Chapter 4

Model-Based Inference

Overview

In this chapter we consider a problem of model inference and testing of alternative hypotheses, when models are defined by non-linear ordinary differential equations and the experimental data is noisy and sparse. We compare and evaluate a number of statistical techniques, and implement an efficient Bayesian inferential framework for Systems Biology based on Markov chain Monte Carlo methods and estimation of marginal likelihoods by annealing-melting integration. We illustrate the application of this framework with two case studies, one of which involves an open problem concerning the mediation of ERK phosphorylation in the ERK signal transduction pathway.

4.1 Background

In this chapter we develop an inferential framework for Systems Biology. The inferential framework consists of methods to perform inference and hypotheses testing tasks. It requires alternative hypotheses to be defined with parametric statistical models. The initial knowledge and beliefs have to be defined as *prior* distributions of model parameters. The likelihood function which compares behaviour produced by a model to the experimental data allows us to infer plausible *posterior* distributions of the model parameters that explain the observed data. Alternative models can be ranked by the degree of their support by experimental data. The main contribution of this chapter is the selection of

methods that perform well with nontrivial models used in modelling biochemical systems.

As we base the inferential framework for Systems Biology on the foundations of Bayesian theory, we begin this chapter with an introduction of the main concepts of Bayesian analysis, and practical methods for Bayesian inference. This overview should be considered as a general one, though sufficient to understand the work described in this chapter. A complete overview of the Bayesian inference philosophy, background and methods can be found in, for example, (Lindley 1965 Box et al. 1983 Bernardo and Smith 1994 Jaynes 2003).

4.1.1 Bayesian Inference

(Bernardo and Smith 1994) demonstrate how Bayesian theory is built on the foundation of axiomatic utility theory, and therefore is conceptually sound. In this section we introduce the main concepts for Bayesian inference, and provide an overview of the methods which can be used to perform such inference.

Formally, Bayesian inference is statistical inference in which evidence or observations are used to update or to infer the probability that a hypothesis may be true. To perform such inference we need to define a way to express our initial beliefs and describe the process by which some evidence or observations can be used to update these beliefs.

Applying Bayesian inference methods requires formal representation of the available knowledge. This should include the statistical model for the problem, and *a priori* information about the model parameters, as we assume that the statistical model is parametric.

In the cases when we have several competing hypotheses about some phenomenon, and therefore several competing models of it, we also associate an *a priori* probability $p(M_i)$ to each model, which describes the degree of initial belief that a particular model is the most appropriate one to describe the observed phenomenon.

Our initial beliefs (initial state of information) about the values of parameters of each available statistical model of the system are, most often, uncertain and therefore distributed according to some probability density function $p(\theta_i|M_i)$. This probability distribution function is called *"a prior distribution of model parameters"*.

When some new information D about the modelled phenomenon is acquired,

we update our beliefs according to Bayes' theorem. The updated distribution of our beliefs is called *"a posterior distribution of model parameters"*. D can correspond to the data from a newly performed experiment, or new information published in a recent paper. Bayes' theorem defines how the posterior can be obtained from the prior, generally:

$$p(\theta_i|M_i, D) = \frac{p(D|M_i, \theta_i) \cdot p(\theta_i|M_i)}{\int p(D|M_i, \theta_i) \cdot p(\theta_i|M_i) d\theta_i}.$$

Here the probability $p(D|M_i, \theta_i)$ to produce data D with model M_i given parameters θ_i is called *"the likelihood"* (see, for example, Cox and Hinkley 1974 Gelman et al. 1995).

Consider Example 4.1 as an illustration of a consistent beliefs update using this methodology:

Example 4.1 (Inference about a genetic probability)

This example is proposed by Gelman et al. (1995) to illustrate how inference can be performed using Bayes' theorem.

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Haemophilia is a disease that inhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene which causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her two X-chromosomes is not affected. The disease is generally fatal for women who inherit two such genes, and this is very rare, since the frequency of occurrence of the gene is low in human populations.

Consider a woman who has an affected brother, which implies that her mother must be a carrier of the haemophilia gene with one "good" and one "bad" X-chromosome. We are also told that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene. The unknown quantity of interest, the state of the woman, has only two possible values: the woman is a carrier of the gene ($\theta = 1$) or not ($\theta = 0$). Based on the information provided thus far, the prior distribution for the unknown θ can be expressed as: $p(\theta = 1) = p(\theta = 0) = \frac{1}{2}$.

The new information which is used for inference is the status of the woman's sons. Suppose she has two sons, neither of whom is affected. Let $d_i = 1$ or

0 *mean affected or unaffected son, respectively. We assume that sons are not identical twins, therefore the outcomes are independent. The likelihood takes the following form:*

$$p(d_1 = 0, d_2 = 0|\theta = 1) = (0.5)(0.5) = 0.25$$
$$p(d_1 = 0, d_2 = 0|\theta = 0) = (1)(1) = 1.$$

Bayes' theorem can now be used to update our beliefs in whether the woman is a carrier of a haemophilia gene. Using D to denote the joint data (d_1, d_2), the posterior probability that the woman is a carrier is:

$$\begin{aligned} p(\theta = 1|D) &= \frac{p(D|\theta = 1)p(\theta = 1)}{p(D|\theta = 1)p(\theta = 1) + p(D|\theta = 0)p(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = \frac{0.125}{0.625} = 0.20. \end{aligned}$$

Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes' theorem provides a formal mechanism for determining the extent of the correction.

In Example 4.1 we were in a (very rare) position, when we were able to iterate through the finite space of the available options, and compute the likelihoods and the priors by simple enumeration of possible outcomes.

Analytical inference of parameter posteriors is also possible in some special cases when the likelihood belongs to the exponential family, e.g. Normal, Bernoulli, Poisson) and a conjugate prior is used. This approach is not considered in this thesis as it is not applicable to the class of models used. A detailed description of the conjugate priors approach can be found in (Bernardo and Smith 1994).

In this chapter we consider several complex models as our case studies. These are formulated using non-linear ordinary differential equations. We argue that, in the majority of realistic applications within Systems Biology, it is not possible to perform inference analytically due to the complexity of the integrals involved. In such cases we need some numerical methods to be able to evaluate the posteriors. A large family of such numerical methods is called Monte Carlo methods, which we consider in Section 4.1.3.

4.1.2 Deterministic Approximations to the Posterior

In this section we consider an analytical method of inferring parameter posteriors using a Taylor series expansion of the likelihood around the maximal likelihood estimate of the parameter.

This approach is justified for the cases when the posterior is almost multivariate normal. This, however, is rare when the models are defined using nonlinear differential equations, and only a few variables can be observed.

In some cases, analytical methods cannot be applied for inference mainly due to complexity of evaluating an integral of the form:

$$E\left[g(\theta)|D\right] = \int_{\theta\in\Theta} g(\theta)p(\theta|D)d\theta, \tag{4.1}$$

where $p(\theta|D)$ is derived from a predictive model, and $g(\theta)$ is some real-valued function of interest. Often, $g(\theta)$ is a first or second moment, and $p(\theta|D)$ is given by

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int_{\theta\in\Theta} p(D|\theta)p(\theta)d\theta}.$$

In order to evaluate (4.1) using Laplace's method (see Tierney and Kadane 1986) we first express the integrand in the form $exp\left\{\log(g(\theta)p(\theta|D))\right\}$ and then expand $\log(g(\theta)p(\theta|D))$ as a function of θ in a quadratic Taylor series around its mode θ_0:

$$\log(g(\theta)p(\theta|D)) \approx \log(g(\theta_0)p(\theta_0|D)) - \frac{1}{2}(\theta-\theta_0)^T\mathbf{A}(\theta-\theta_0) + \ldots,$$

where

$$\mathbf{A} = -\nabla^2 \log\left(g(\theta)p(\theta|D)\right)\big|_{\theta=\theta_0}, \tag{4.2}$$

$$\mathbf{A}_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log(g(\theta)p(\theta|D))\bigg|_{\theta=\theta_0}. \tag{4.3}$$

This method assumes that θ_0, as a mode of a multivariate Gaussian distribution, is the maximum a posteriori estimate of the model parameters.

$g(\theta)p(\theta|D)$ can then be approximated by an unnormalised Gaussian

$$q^*(\theta) = g(\theta_0)p(\theta_0|D)exp\left\{-\frac{1}{2}(\theta-\theta_0)^T\mathbf{A}(\theta-\theta_0)\right\}, \tag{4.4}$$

and the normalising constant Z_q for this Gaussian is

$$Z_q = g(\theta_0)p(\theta_0|D)\sqrt{\frac{(2\pi)^n}{\det \mathbf{A}}}, \tag{4.5}$$

where n is the dimensionality of the parameter space.

For the cases when matrix $\mathbf{A}$ cannot be evaluated analytically, numerical derivation methods can be used. Though, this approximation is valid, as already mentioned, it is only valid in the cases when the posterior is unimodal and almost multivariate normal. For more complex cases, for example, when the likelihood is multimodal or differs significantly from a normal distribution (as in the case studies considered in Sections 4.3.1 and 4.3.2), the Laplace approximations are not valid. In more complex cases, other methods, for example, Monte Carlo methods, must be applied.

4.1.3 Monte Carlo Methods

As we stated above, we need to evaluate complex probability distributions when performing inference over model parameters from observed data. A straightforward evaluation of these probabilities is problematic, as the likelihood function can involve complex nonlinear forms. Monte Carlo methods are computational techniques developed to generate samples from a desired probability distribution $p(x)$ and to compute integrals of the form (4.1). The generated samples can be used to estimate the probability densities of interest. We start with a concise introduction to Monte Carlo integration, and then consider three Monte Carlo methods in this chapter: rejection sampling, Metropolis-Hastings sampling and Gibbs sampling.

Monte Carlo Integration

Monte Carlo methods are computational techniques developed to generate samples $\left\{\theta^{(r)}\right\}_{r=1}^{R}$ from a desired probability distribution $p(\theta|D)$ and to compute integrals of the form (4.1):

$$E\left[g(\theta)|D\right] = \int_{\theta\in\Theta} g(\theta)p(\theta|D)d\theta,$$

The probability distribution $p(\theta|D)$, which in this case is called *target den-*

sity, might be a distribution of model parameters arising in biological modelling, for example, the posterior distribution of model parameters given some observed data.

If the first part of the problem, generating a sample from the target density, is solved, then the value of the integral can be estimated as

$$E\left[g(\theta)|D\right] = \int_{\theta\in\Theta} g(\theta)p(\theta|D)d\theta, \approx \frac{1}{R}\sum_{r=1}^{R} g(\theta^{(r)}). \tag{4.6}$$

As the number of samples R increases, the variance of this estimate will decrease as σ^2/R, where σ^2 is the variance of $g(\theta)$,

$$\sigma^2 = \int_{\theta\in\Theta} p(\theta|D)(g(\theta) - E\left[g(\theta)|D\right])^2 d\theta.$$

An important property of Monte Carlo integration methods is that the accuracy of the Monte Carlo estimate (4.6) depends only on the variance of $g(\theta)$ and not on the dimensionality of the space sampled. So, regardless of the dimensionality of the parameter space, it may be that as few as a dozen samples $\left\{\theta^{(r)}\right\}$ suffice to estimate $E\left[g(\theta)|D\right]$ satisfactorily.

We consider three popular Monte Carlo methods of generating a sample from a target density: rejection sampling, the Metropolis-Hastings method and Gibbs sampling.

Rejection Sampling

The general background for this method is described in (Robert and Casella 2004) and (MacKay 2003) as the following:

Assume that we need to generate a sample from a univariate probability density $p(x) = p^*(x)/Z$, and it is difficult to sample from such distribution directly. We assume that we have a simpler proposal density $q(x)$ which we can evaluate (within some multiplicative factor Z_q, such that $q(x) = q^*(x)/Z_q$), and from which we can generate samples. We further assume that we know the value of a constant c such that

$$\forall x(cq^*(x) > p^*(x)). \tag{4.7}$$

A schematic picture of such functions is depicted in Figure 4.1.

To produce a sample from $p(x)$ we generate two random numbers. The first,

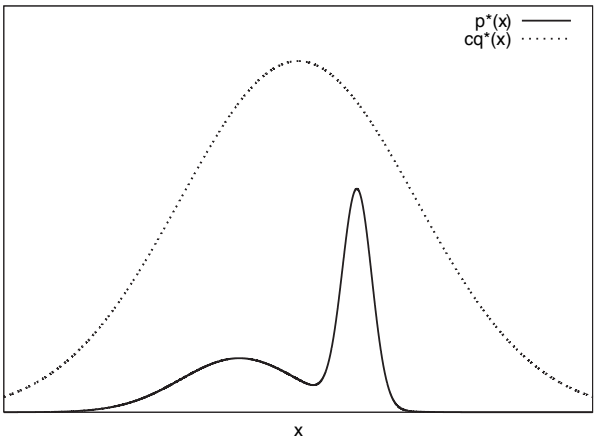


Figure 4.1: The functions involved in rejection sampling. $p^*(x)$ is too complex to sample from directly, so some simple distribution $cq^*(x)$ is chosen in a way that its density is always larger than $p^*(x)$.

x, is generated from the proposal density $q(x)$. We then evaluate $cq^*(x)$ and generate the second random number, u, from a uniform distribution over the interval $[0, cq^*(x)]$. At the next stage we evaluate $p^*(x)$ and accept or reject sample x by comparing the value of u with the value of $p^*(x)$. If $u > p^*(x)$, then the sample x is rejected; otherwise it is accepted, which means that we add x to our set of samples $\{x^{(r)}\}$.

This procedure generates samples from $p(x)$ because the proposed point (x, u) comes uniformly from the area underneath the curve $cq^*(x)$ (see Figure 4.1), and the rejection rule rejects all the points that lie above the curve $p^*(x)$. So the points (x, u) which are accepted are uniformly distributed in the area under the curve $p^*(x)$. This implies that the probability density of the x-coordinates of the accepted points must be proportional to $p^*(x)$, so the samples must be independent samples from $p(x)$.

Rejection sampling will work best if the proposal distribution $q(x)$ is close to $p(x)$. However, in the cases when $q(x)$ is significantly larger (in terms of the area under the curve $cq^*(x)$) than $p(x)$, then rejection sampling will be very inefficient, as the majority of the proposed points (x, u) will be rejected. This problem becomes a significant drawback when applying this method to sampling from multidimensional distributions, as the acceptances become very rare indeed.

The Metropolis-Hastings Method

We stated that rejection sampling is very inefficient when sampling from complex multidimensional distributions. The Metropolis-Hastings algorithm employs a different approach which overcomes the efficiency problems of rejection sampling. This algorithm originates in statistical physics (see Metropolis et al. 1953) where it is used to investigate properties of large two- and three-dimensional systems of interacting particles. Later, this method was employed as a Bayesian inference machine by Grenander (1983) and Geman and Geman (1984) in the context of image analysis.

The Metropolis-Hastings algorithm uses a proposal density Q which depends on the current sample $x^{(r)}$. This proposal density $q(x', x^{(r)})$ can be any fixed density from which we can draw samples.

As before, we assume that we can evaluate $p^*(x)$ for any proposed x. A tentative new x' is generated from the proposal density $q(x', x^{(r)})$, to decide whether to accept this value into our sample, we have to compute the value:

$$a = \frac{p^*(x')}{p^*(x^{(r)})} \cdot \frac{q(x^{(r)}, x')}{q(x', x^{(r)})}. \tag{4.8}$$

If $a \geq 1$ then x' is accepted into the sample. Otherwise, x' has to be accepted with probability a. If x' is accepted into the sample, then it will be taken as the base for the proposal distribution of $x^{(r+1)}$ at the next sampling iteration. Otherwise, the last accepted value $x^{(r)}$ has to be duplicated: $x^{(r+1)} = x^{(r)}$.

The Metropolis-Hastings algorithm is an example of *the Markov Chain Monte Carlo* (*MCMC*) method (see Gilks et al. 1995 Robert and Casella 2004 Gamerman 2006). In contrast to rejection sampling where the accepted values $\left\{x^{(r)}\right\}$ are independent samples from the desired distribution, MCMC methods use a partial realisation $\left\{x^{(r)}\right\}_{r=0}^{N}$ from a Markov chain with stationary distribution $p(x)$ (see below for definitions).

We now give some basic background from the theory of Markov chains relevant to MCMC.

Definition 4.1 (Markov chain)**:** *A Markov chain is a sequence of random variables $X^0, X^1, \ldots$ taking values in $\mathcal{X}$ with the Markov property:*

$$Pr\left(X^{t+1} = x \,\middle|\, X^t = x_t, \ldots X^1 = x_1, X^0 = x_0\right) = Pr\left(X^{t+1} = x \,\middle|\, X^t = x_t\right).$$

$\mathcal{X}$ is called the state space of a Markov chain.

For simplicity, we focus on Markov chains with countable state space $\mathcal{X}$. For a deeper theoretical background, especially with regard to continuous state spaces, see, for example, (Tierney 1994) or (Athreya et al. 1992).

Definition 4.2: *A Markov chain is called time homogeneous if*

$$Pr\left(X^{t+1} = x \,\middle|\, X^t = y\right) = Pr\left(X^t = x \,\middle|\, X^{t-1} = y\right)$$

for all t.

While considering MCMC methods we use time homogeneous Markov chains, thus we assume this property for all the subsequent definitions.

Definition 4.3: *The transition kernel $p(x \to x')$ of a Markov chain is defined by*

$$p(x \to x') = Pr\left(X^{t+1} = x' \,\middle|\, X^t = x\right).$$

Definition 4.4: *The n-step transition kernel $p^n(x \to x')$ of a Markov chain is defined by*

$$p^n(x \to x') = Pr\left(X^{t+n} = x' \,\middle|\, X^t = x\right).$$

Definition 4.5: *A Markov chain is irreducible if there is a positive probability to get from any state to any state.*

Definition 4.6: *A period d_x of the state x is*

$$d_x = \gcd\left\{n \geq 0 : p^n(x \to x') > 0\right\}.$$

A Markov chain is aperiodic if $\forall x \in \mathcal{X} \quad (d_x = 1)$.

Definition 4.7: *$\pi(x)$ is a stationary distribution of a Markov chain if*

$$\forall x' \in \mathcal{X} \quad \left(\sum_{x \in \mathcal{X}} \pi(x) p(x \to x') = \pi(x')\right).$$

Definition 4.8: *Let $T_x = inf\left\{t \geq 1 : X^t = x \,|X^0 = x\right\}$, a Markov chain is called*

positive recurrent if, and only if, $\forall x \in \mathcal{X}(E[T_x] < \infty)$*, where* $E[T_x]$ *is the expected value of* T_x*.*

Theorem 4.1 (Ergodic theorem)**:** *A Markov chain which is aperiodic, irreducible, and positive recurrent has a unique stationary distribution.*

More details about ergodicity and the ergodic theorem can be found in (Feller 1968) or (Iosifescu 1980).

The key feature of the Markov chain theory for MCMC is that the empirical distribution of an aperiodic, irreducible, positive recurrent Markov chain converges to its stationary distribution (see Levental 1988).

It was demonstrated (see e.g. Hastings 1970 Neal 1993) that the Metropolis-Hastings algorithm builds an aperiodic, irreducible, positive recurrent Markov chain with the stationary distribution $p(x)$ (or more precisely, performs a random walk in such Markov chain), which means that after a large number of initial steps it produces a sample from $p(x)$.

Example 4.2 (Demonstration of the Metropolis-Hastings algorithm.)

The Metropolis-Hastings algorithm is often used for multidimensional problems, as it avoids the common problems of rejection samplers. Many implementations of this algorithm employ a proposal distribution with a length scale ϵ *which is short relative to the scale* L *of the desired distribution (see Figure 4.2). The reason for such choice is that for multivariate problems a large random step from a typical point is very likely to end in a state which has very low probability; such steps are unlikely to be accepted.*

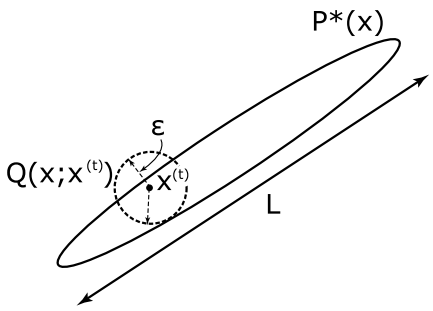


Figure 4.2: Traditional proposal density for a Metropolis-Hastings algorithm in two dimensions (see MacKay 2003).

For a demonstration we use the Metropolis-Hastings algorithm to draw samples from a two-dimensional normal distribution:

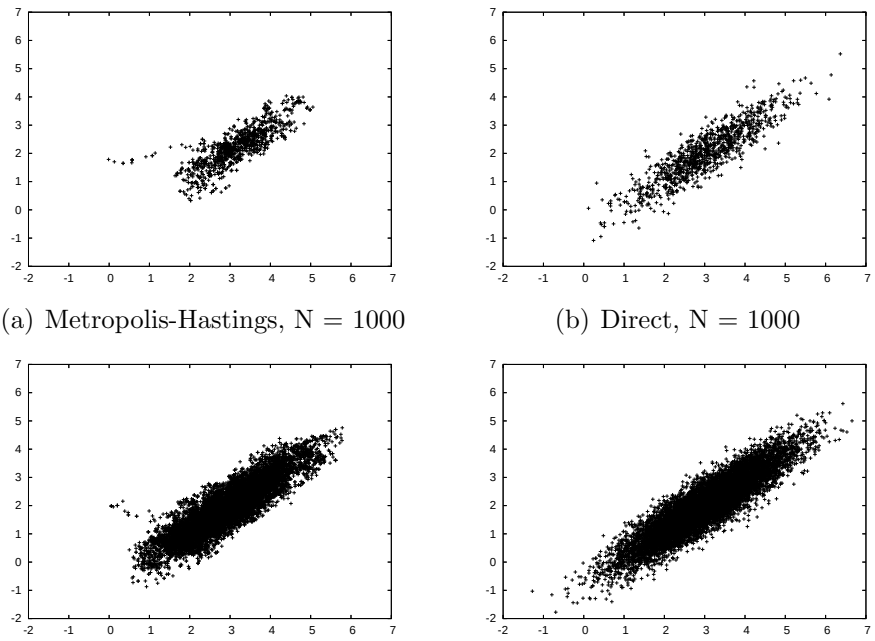
$$\mathcal{N}\left(x \middle| \mu = (3,2)\,, \Lambda = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right).$$

We use

$$N\left(x^{(r+1)} \middle| \mu = x^{(r)}, \Lambda = \begin{bmatrix} 0.15 & 0 \\ 0 & 0.15 \end{bmatrix}\right)$$

as the proposal distribution. Figure 4.3 depicts a comparison of samples generated using Metropolis-Hastings algorithm to the samples drawn directly from the target distribution.

The smaller sample produced with the Metropolis-Hastings algorithm (see Figure 4.3(a)) has not converged enough to the target; as a result the initial random walk (a trail in the left part of Figure 4.3(a)) significantly distorts the sample. At the same time the larger sample produced with Metropolis-Hastings algorithm with the same parameters (see Figure 4.3(c)) is much closer to the target.



(a) Metropolis-Hastings, N = 1000

(b) Direct, N = 1000

(c) Metropolis-Hastings, N = 10000

(d) Direct, N = 10000

Figure 4.3: Metropolis-Hastings sampling for a toy problem. Samples (a) and (c) were produced with the Metropolis-Hastings algorithm, while samples (b) and (d) were drawn directly from the target distribution.

Gibbs Sampling

Gibbs sampling (see e.g. MacKay 2003 Thomas et al. 1992) is a method of sampling from distributions of at least two dimensions. This method is used when conditional distributions of the joint target distribution can be easily evaluated. It is assumed that, while $p(x)$ is too complex to draw samples from directly, its conditional distributions $P\left(x_i \middle| \{x_j\}_{j\neq i}\right)$ are tractable to work with. This holds for many simple statistical models, but is generally not applicable to models of most biological systems described with nonlinear differential equations, since the likelihood function in such cases cannot be reduced to conditional probabilities.

The iterative procedure of the Gibbs sampler is similar to the Metropolis-Hastings algorithm described above, the only difference is how the proposal step is performed. In a general case of a system with K parameters, a single proposal step is performed by sampling each parameter separately, conditioned on the values of the rest of parameters:

$$\begin{aligned} x_1^{(r+1)} &\sim P\left(x_1 \middle| x_2^{(r)}, x_3^{(r)}, \dots x_K^{(r)}\right) \\ x_2^{(r+1)} &\sim P\left(x_2 \middle| x_1^{(r+1)}, x_3^{(r)}, \dots x_K^{(r)}\right) \\ &\dots \\ x_K^{(r+1)} &\sim P\left(x_1 \middle| x_1^{(r+1)}, x_2^{(r+1)}, \dots x_{K-1}^{(r+1)}\right) \end{aligned} \tag{4.9}$$

The convergence of the Gibbs sampler to the target distribution follows from the fact that this sampler is a special case of the Metropolis-Hastings algorithm, and therefore the probability distribution of $x^{(r)}$ tends to $p(x)$ as $t \to \infty$.

Both the basic Metropolis-Hastings algorithm and the Gibbs sampler suffer from the same defect that the state space is explored by a slow random walk. This can be easily observed when some of the parameters are strongly correlated, in such cases the acceptance rate of both algorithms reduces usually to very small values (only few proposed steps become accepted).

Adaptive Proposals

To overcome the problem with inefficient proposals in a general Metropolis-Hastings algorithm, Gelman et al. (1995) proposes to use an adaptive proposal distribution, resetting the Markov chain several times until a good acceptance rate is achieved.

Suppose there is a model with K parameters, and the posterior distribution of the parameters $\theta = (\theta_1, \ldots, \theta_K)$. Gelman et al. (1995) propose to take draws using the Metropolis-Hastings algorithm with a symmetric normal proposal distribution of the same shape as the current approximation of the target distribution: that is

$$Q\left(\theta^{(r)} \middle| \theta^{(r-1)}\right) = N\left(\theta^{(r)} \middle| \mu = \theta^{(r-1)}, \Lambda = c^2\Sigma\right),$$

where Σ is an estimate of the posterior's variance-covariance matrix. In practice, among this class of proposal distributions, the most efficient[1] has scale $c \approx 2.4/\sqrt{K}$ (see Gelman et al. 1995). The optimal acceptance rate for multivariate problems is about 0.23. This improved proposal suggests the following *adaptive* sampling algorithm:

1. Start the simulation with a fixed proposal distribution using a standard version of the Metropolis-Hastings or Gibbs algorithm.

2. After some number of iterations, update the proposal distribution as follows:

   (a) Adjust the covariance of the proposal distribution to be proportional to the posterior covariance matrix estimated from the simulated sample.

   (b) Increase or decrease the scale of the jumping distribution if the acceptance rate of the simulations is much too high or low, respectively. The goal is to bring this acceptance rate to the approximate optimal value between 0.44 and 0.23.

Note, that when using this adaptive algorithm, the simulation of the Markov chain has to be restarted when the proposal distribution is updated, the sample can be expected to converge to the target distribution only once the proposal is fixed after a number of updates.

Convergence monitoring

As we already stated in the overview of Markov chain theory, the chains built with MCMC algorithms converge to some stationary distribution, and the Metropolis-

[1] Gelman et al. (1995) justify this choice by their practical experience with many statistical problems.

Hastings algorithm builds the chains in a way that the desired parameter posterior is this stationary distribution. We immediately can see the main technical problem for MCMC sampling: how many samples are required to be sure that the sample is produced from the stationary distribution?

The actual number of required "burn-in"[2] samples can not be precisely defined. Some methods, however, have been developed to detect whether the sampler reached a stationary distribution. In our practice, we use a method published by Gelman et al. (1995). This method suggests running several similar Markov Chains in parallel, and assessing the level of mixing[3] of the samples produced by them.

The method of Gelman et al. (1995) is used to monitor the mixing of parameter samples by computing a mixing statistic for each parameter separately, and when the values of this statistic is small enough for each of the parameters, we assume that the Markov chains have converged.

For each scalar parameter p, we label the draws from J parallel samplers of length n as p_{ij} $(i = 1, \ldots, n; j = 1, \ldots, J)$. For each scalar parameter we compute its between- and within-sequence variances, B and W correspondingly.

$$B = \frac{n}{J-1} \sum_{j=1}^{J} (\bar{p}_{\cdot j} - \bar{p}_{\cdot\cdot})^2, \quad W = \frac{1}{J} \sum_{j=1}^{J} s_j^2,$$

where

$$\bar{p}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}, \quad \bar{p}_{\cdot\cdot} = \frac{1}{J} \sum_{j=1}^{J} \bar{p}_{\cdot j}, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (p_{ij} - \bar{p}_{\cdot j})^2.$$

The between-sequence variance, B, contains a factor of n because it is based on the variance of the within-sequence means, $\bar{p}_{\cdot j}$, each of which is an average of n values p_{ij}. If only one sequence is simulated, B cannot be calculated.

The marginal posterior variance of parameter p can be estimated by a weighted average of W and B:

$$\widehat{var}(p|D) = \frac{n-1}{n} W + \frac{1}{n} B,$$

which overestimates the marginal posterior variance assuming the starting dis-

[2]Burn-in is an initial part of the sample which is discarded to give the Markov chain some time to converge to the target distribution.

[3]Mixing of samples means that the samples are produced from the same probability distribution.

tribution is overdispersed[4].

For any finite n, the within-sequence variance W should be an underestimate of $var(p|D)$ because the individual sequences have not had time to range over all the target distribution and, as a result, will have less variability; in the limit as $n \to \infty$, the expectation of W approaches $var(p|D)$.

Gelman et al. (1995) propose to monitor convergence of the iterative simulation by estimating the factor by which the scale of the current distribution for p might be reduced if the simulations were continued in the limit $n \to \infty$. This potential scale reduction is estimated by

$$\hat{R} = \sqrt{\frac{\widehat{var}(p|D)}{W}},$$

which declines to 1 as $n \to \infty$. If the potential scale reduction is high, then we have reason to believe that proceeding with further simulations may improve our sample from the target distribution.

4.1.4 Model Comparison and Bayes Factors

The methodology presented in this section allows one to rank competing hypotheses by the evidential support from experimental data, and therefore evaluate relative confidence values for such hypotheses. A complete comprehensive overview of Bayes factors and model comparison can be found in e.g. (Kass and Raftery 1995).

In the cases when a discrete set of competing hypotheses is considered, the hypotheses can be ranked by the ratio of their posterior probabilities. For a pair of hypotheses H_1 and H_2 represented with models M_1 and M_2 the ratio is

$$\frac{p\,(M_1|D)}{p\,(M_2|D)}. \tag{4.10}$$

Taking a prior distribution of beliefs in preference of each hypotheses π into account, and in the case when hypotheses are represented by parametric models, this ratio is:

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{\pi\,(M_1)}{\pi\,(M_2)} \times \frac{p\,(D|M_1)}{p\,(D|M_2)} = \frac{\pi\,(M_1)}{\pi\,(M_2)} \times \frac{\int p(D|M_1,\theta_1)\cdot p(\theta_1|M_1)d\theta_1}{\int p(D|M_2,\theta_2)\cdot p(\theta_2|M_2)d\theta_2} \tag{4.11}$$

[4]Overdispersion in this context means that the variance of the initial population of parallel chains is significantly higher than the variance of the posterior distribution.

Definition 4.9: *The ratio of the marginal likelihoods for two competing hypotheses:*

$$\frac{\int p(D|M_1, \theta_1) \cdot p(\theta_1|M_1) d\theta_1}{\int p(D|M_2, \theta_2) \cdot p(\theta_2|M_2) d\theta_2}$$

is called the Bayes factor.

Bayes factors are used to test competing hypotheses, and update corresponding beliefs using formula (4.11).

Example 4.3 (Bayes factors applied to Example 4.1.)

The genetics example given as Example 4.1 can be reintroduced in terms of Bayes factors, considering two competing hypotheses:

H_1**:** *the woman is affected,*

H_2**:** *the woman is not affected.*

That is $\theta = 1$ *and* $\theta = 0$ *in terms of Example 4.1. The prior odds are*

$$\frac{\pi(H_2)}{\pi(H_1)} = 1,$$

as we consider both hypotheses to be equally probable: $\pi(H_1) = \pi(H_2)$.

The Bayes factor of data that the woman has two unaffected sons is

$$\frac{p(y|H_2)}{p(y|H_1)} = \frac{1.0}{0.25}.$$

The posterior odds are thus

$$\frac{p(H_2|y)}{p(H_1|y)} = 4.$$

Which means that it is four times more likely that the woman is unaffected if she has two unaffected sons. This result matches the one obtained by applying Bayes' theorem directly in Example 4.1.

We propose that reasoning employing Bayes factors is much more comprehensible and more natural than that given in Example 4.1.

Example 4.3 demonstrates how Bayes factors are calculated when only discrete options are considered. When using models with continuous parameter space the problem becomes more complex, as Bayes factors have to be evalu-

ated by integration. In the vast majority of practical problems these integrals cannot be evaluated analytically, and therefore numerical methods are required to estimate them. These integrals are called marginal likelihoods, and we give a brief overview of some numerical methods to estimate them in the next section (see Section 4.1.5).

The Bayes factor is a summary of the evidence provided by the data in favour of one hypothesis, represented by a model, as opposed to another. Jeffreys (1961) suggested interpreting Bayes factors in half-units on the $\log_{10}$ scale. Pooling two of his categories together for simplification we demonstrate his scale in Table 4.1.

$\log_{10}(B)$	B	Evidence support
0 to 1/2	1 to 3.2	Not worth more than a bare mention
1/2 to 1	3.2 to 10	Substantial
1 to 2	10 to 100	Strong
> 2	> 100	Decisive

Table 4.1: Interpretation of the Bayes factor as evidence support categories according to Jeffreys (1961)

These categories are not a calibration of the Bayes factor, as it already provides a meaningful interpretation as probability, but rather a rough descriptive statement about standards of evidence in scientific investigation.

Kass and Raftery (1995) propose a slight modification to this scale, and use natural logarithms instead. This modified scale is demonstrated in Table 4.2.

$2\log_e(B)$	B	Evidence support
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Table 4.2: Interpretation of the Bayes factor as evidence support categories according to Kass and Raftery (1995)

There are a number of publications on the controversy between Bayesian and non-Bayesian testing procedures. The following four issues are usually considered:

1. P values used in non-Bayesian significance testing are not similar to the posterior probability that the null hypotheses is correct. Jeffreys (1961)

demonstrates this problem and considers the results obtained with both approaches.

2. Non-Bayesian tests tend to reject null hypotheses in very large samples, whereas Bayes factors do not. This has been a problem in sociology, where the data sets can contain thousands of cases. Facing this problem, sociologists have taken to ignoring significance tests and using other criteria and informal methods when comparing models. An example with number of samples $n = 113,566$ was discussed by Raftery (1986), where a meaningful model that explained 99.7% of the deviance was rejected by a standard chi-squared test with a P value of about 10^{-120} but was nevertheless favoured by the Bayes factor. Bayes factors are now widely used in sociology, usually with BIC (Bayesian Information Criterion) as an approximation.

3. Bayes factors can be applied to both nested[5] and non-nested models, while application of non-Bayesian significance tests to non-nested models is difficult. This problem is briefly discussed in (Kass and Raftery 1995).

4. Non-Bayesian significance tests were designed for comparison of two models, but practical data analysis often involves more than two models, at least implicitly. In such a case, performing multiple significance tests to guide a search for the best model can give very misleading results (e.g. Freedman 1983). This problem can be avoided by taking model uncertainty into account and employing Bayes factors (e.g. Raftery et al. 1993).

Arkinson (1978) has noted some examples when Bayes factors favoured the simpler model H_0 even when a more complex model H_1 was correct. Smith and Spiegelhalter (1980) demonstrated that this occurs only when the models are so close that there is almost no loss in predictive power when cutting back to the simpler model, so that Bayes factors can be considered as *a fully automatic Occam's razor*[6].

Akaike (1973) proposed yet another criterion for model comparison, which also takes the complexity of the models into account. This criterion suggests to

[5]Nested models are statistical models with model parameters arranged in a hierarchical structure.

[6]Occam's razor is a principle which states that the explanation of any phenomenon should make as few assumptions as possible. Thus, the simplest model which explains the evidence sufficiently should be chosen as the most appropriate one.

choose the model which minimises AIC (Akaike information criterion):

$$AIC = -2(\text{log maximum likelihood}) + 2(\text{number of parameters}). \tag{4.12}$$

There are two main justifications for this criterion. The first one is based on the predictive argument. Suppose that, given current data and a set of possible models, we are seeking for a predictive distribution of a future datum. Then, if the predictive distribution is conditional on a single model and on its estimated parameters, the AIC picks the model that gives the best approximation, asymptotically, in the Kullback-Leibler sense. AIC tends to overestimate the number of parameters needed, even asymptotically. The second main justification for the AIC is Bayesian. Akaike (1983) claimed that model comparisons based on the AIC are asymptotically equivalent to those made with Bayes factors. But this is true only in the situations when predictions of the prior are compatible to those of the likelihood, and not in the more usual situation when prior information is small in comparison to the information provided by the data. In the latter (and more usual situation) the Bayesian Information Criterion (BIC), also known as Schwartz criterion, indicates that the model with the highest posterior probability is the one which minimises

$$BIC = -2(\text{log maximum likelihood}) + (logN)(\text{number of parameters}), \tag{4.13}$$

where N is a number of observations.

Comparing Equations (4.12) and (4.13) indicates that BIC tends to favour simpler models than those chosen by the AIC criterion.

The deviance information criterion (DIC) (see Spiegelhalter et al. 2002) is sometimes used for model comparison when parameter posteriors were obtained with Markov chain Monte Carlo simulation. This criterion allows one to avoid marginal likelihood estimation and is more convenient than AIC and BIC as it does not require maximum likelihood estimation. Like AIC and BIC it is an asymptotic approximation. It is only valid when the posterior distribution is approximately multivariate normal. DIC indicates that the model which minimises

$$DIC = 2\bar{D} - D(\bar{\theta}) \tag{4.14}$$

should be preferred over the rest. In Equation (4.14)

$$D(\bar{\theta}) = -2\log(p(D|M,\bar{\theta})),$$

where $\bar{\theta}$ is the expectation of the parameter posterior; and

$$\bar{D} = \mathbf{E}_{p(\theta|D,M)}[-2\log(p(D|M,\theta))]$$

is the expectation of $-2log(p(D|M,\theta))$ on the posterior sample.

Using hypotheses testing approaches based on asymptotic approximations (AIC, BIC, DIC) sometimes provides unreliable results, as these methods are justified only for the cases when parameter posteriors are unimodal and almost multivariate normal. This is a very rare case in modelling biological systems, as the majority of the involved models are nonlinear. Employing asymptotic methods with such models often provides confusing and incorrect results. For example, the definition of DIC assumes that the likelihood estimate of the average parameter value produces the maximum likelihood value, which is incorrect in the case of multimodal or nonlinearly shaped posteriors.

In our Case Studies we rely on the hypotheses testing results obtained with Bayes factors. However, computing such Bayes factors is a challenging problem, as the marginal likelihoods for nonlinear models have to be evaluated to obtain these. In the following section we discuss alternative methods for estimation of the marginal likelihoods.

4.1.5 Estimation of the Marginal Likelihoods

Evaluation of marginal likelihoods can be successfully avoided in estimation of the posterior parameter distribution by employing Metropolis-Hastings sampling algorithm. However, in Section 4.1.4 we demonstrated that marginal likelihoods are required to perform hypotheses testing and model comparison with Bayes factors. A review of different methods for evaluating marginal likelihoods can be found in (Newton and Raftery 1994 Kass and Raftery 1995 Chib 1995).

The main problem is that the marginal likelihood

$$p(D|M) = \int p(D|M,\theta) \cdot p(\theta|M) d\theta$$

can be evaluated *analytically* only in very special cases, e.g. when the likelihood belongs to the exponential family, and conjugate priors are used. The majority of the mechanistic biological models considered in this thesis are based on non-linear ordinary differential equations that contribute to the likelihood. In such cases analytical integration of the marginal likelihood is impossible, and therefore we will not consider this method in our case studies. Brute force *numerical integration* can be applied to low-dimensional problems. This approach, however, becomes computationally intractable for more complicated applications (this estimate becomes impractical in more than two-dimensional case, as its computational complexity depends exponentially on the dimensionality of the parameter space). The artificially constructed models considered in Case Study 1 are defined in 5 to 8 dimensions, and the realistic models from Case Study 2 are defined in over 100 dimensions each. Brute force numerical integration cannot be performed effectively in such parameter spaces, and therefore this method will not be considered.

The reasons of complexity discussed above leave us with the only practical option of considering methods for approximate evaluation of marginal likelihoods. Many of these approximate methods are limited by very strong conditions. For example, *Laplace approximations* or DIC are large sample approximations around the maximum a posteriori estimate, which can be difficult to find in some cases of complex problems. Moreover, such asymptotic approximations rely on almost-normality of the target distribution, which is often wrong for non-linear problems. For example, see Figure 4.14(b) from Case Study 1 (page 132), strong interaction of the model parameters causes significant curvature of the posterior distribution density.

The *reversible Jump MCMC* approach (Green 1995), where Markov chains are constructed in a special way, allow jumps between alternative models in accordance with Metropolis-Hastings ratio, and in principle, can be tuned for any problem. Such an approach, however, creates significant technical difficulties, as the rate of jumps between alternative models has to be maintained at some acceptable level, to obtain a satisfactory result. Friel and Pettitt (2006) investigated the problem of estimates stability for the Reversible Jump MCMC approach, they suggested to introduce a correction to the relative prior preference between the alternative models to achieve acceptable precision. Such correction sacrifices the possibility to use uneven priors over alternative models,

and consequently limits the analysis possibilities.

Two more methods which can be applied in a general case are *importance sampling estimators* (Newton and Raftery 1994) and *thermodynamic integration* or *path sampling* (Ogata 1989 Gelman 1998).

In this section, we give a detailed description and practical comparison of three estimators from the above classes: *the prior arithmetic mean estimator*, *the posterior harmonic mean estimator*, and *annealing-melting integration*.

Importance Sampling Estimators

Importance sampling estimation consists of generating a sample

$$\left\{\theta^{(i)}; i = 1, \ldots, m\right\}$$

from an unnormalised density $\pi^*(\theta)$. Under quite general conditions, an estimate of integral

$$I = \int p(D|M, \theta) p(\theta|M) d\theta$$

is

$$\hat{I} = \frac{\sum_{i=1}^{m} \omega_i \cdot p(D|M, \theta^{(i)})}{\sum_{i=1}^{m} \omega_i}, \tag{4.15}$$

where $\omega_i = p(\theta^{(i)}|M) \big/ \pi^*(\theta^{(i)})$; the function $\pi^*(\theta)$ is known as the *importance sampling function*.

The simplest application of this method is to use the prior as the importance sampling function $\pi^*(\theta) = p(\theta|M)$, in which case (4.15) produces the *prior arithmetic mean estimator* (see McCulloch and Rossi 1991):

$$p(D|M) \simeq \frac{1}{m} \sum_{i=1}^{m} p(D|M, \theta^{(i)}); \quad \theta^{(i)} \sim p(\theta|M). \tag{4.16}$$

A well known problem with this estimator is that the high-likelihood region can be very small. Therefore, unless m is very large, the sample drawn from the prior will contain virtually no points from the high-likelihood region, resulting in a very poor estimate of the marginal likelihood. Lewis and Raftery (1997) reference a study in which to reduce the standard error to an acceptable level, it was necessary to use a sample of roughly 50 million draws from the prior distribution. We investigated the effectiveness of this estimate using linear regression

models in Example 4.4. The relative error of the Bayes factor, which relies on two marginal likelihood estimates, when using 500,000 samples from the prior is 25%, is significantly worse than the estimates achieved with thermodynamic integration methods.

An alternative application of importance sampling estimation, proposed by Newton and Raftery (1994), is to use the parameter posterior as the importance sampling function $\pi^*(\theta) = p(\theta|D, M)$. A sample from the parameter posterior can be obtained using MCMC sampling. Such a sample should be significantly better in covering the high-likelihood region. Substituting the parameter posterior into (4.15) results in the *posterior harmonic mean estimator,* we obtain:

$$\begin{aligned}
p(D|M) \simeq \hat{I} &= \frac{\sum_{i=1}^m \omega_i \cdot p(D|M, \theta^{(i)})}{\sum_{i=1}^m \omega_i} = \frac{\sum_{i=1}^m \frac{p(\theta^{(i)}|M)}{\pi^*(\theta^{(i)})} \cdot p(D|M, \theta^{(i)})}{\sum_{i=1}^m \frac{p(\theta^{(i)}|M)}{\pi^*(\theta^{(i)})}} \\
&= \frac{\sum_{i=1}^m \frac{p(\theta^{(i)}|M)p(D|M)}{p(D|M,\theta^{(i)})p(\theta^{(i)}|M)} \cdot p(D|M, \theta^{(i)})}{\sum_{i=1}^m \frac{p(\theta^{(i)}|M)p(D|M)}{p(D|M,\theta^{(i)})p(\theta^{(i)}|M)}} = \frac{\sum_{i=1}^m p(D|M)}{\sum_{i=1}^m \frac{p(D|M)}{p(D|M,\theta^{(i)})}} \\
&= \frac{m \cdot p(D|M)}{p(D|M) \cdot \sum_{i=1}^m \frac{1}{p(D|M,\theta^{(i)})}} = \frac{m}{\sum_{i=1}^m \frac{1}{p(D|M,\theta^{(i)})}} \\
&= \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{p(D|M, \theta^{(i)})} \right)^{-1} ; \ \theta^{(i)} \sim p(\theta|D, M). \quad (4.17)
\end{aligned}$$

The main problem with this estimate is that, in many practical situations, its variance is infinite, because of the occasional occurrence of a value of $\theta^{(i)}$ with a small likelihood and hence a large effect on the final result. As we demonstrate in Example 4.4 on simple regression models, this estimate is very unstable because of the reasons described above.

There exists a number of modifications of this approach proposed by Newton and Raftery (1994) which propose to combine sampling from the posterior with sampling from the prior, for example, by using a mixture $\delta p(\theta|M) + (1 - \delta)p(\theta|D, M)$ as the importance sampling function, where δ is small. The forms for an estimate for this case can be found in (Newton and Raftery 1994).

Thermodynamic Integration

The method of thermodynamic integration originates in Statistical Physics (for an overview see Neal 1993), where the marginal likelihood is equivalent to the

so-called partition function and its logarithm to the free energy. The computations required to perform thermodynamic integration are computationally more intensive, but the results are usually more stable (Gelman 1998).

This method is based on the following principles: suppose that there are two unnormalised distributions $q_0(\theta)$ and $q_1(\theta)$, defined on the same parameter space Θ. We can normalise these densities dividing them by normalisation constants.

$$p_i(\theta) = \frac{1}{Z_i} q_i(\theta),\ i = 0, 1,$$

where

$$Z_i = \int_\Theta q_i(\theta) d\theta,\ i = 0, 1$$

To perform the evaluation of log-ratio

$$\mu = \ln\left(\frac{Z_1}{Z_0}\right) = \ln Z_1 - \ln Z_0$$

a continuous and differentiable path $(q_\beta)_{0\leq\beta\leq 1}$ can be defined in the space of unnormalised densities, joining q_0 and q_1. Similarly,

$$p_\beta(\theta) = \frac{1}{Z_\beta} q_\beta(\theta),$$

$$Z_\beta = \int_\Theta q_\beta(\theta) d\theta.$$

Taking the derivative of $\ln Z_\beta$ with respect to β:

$$\begin{aligned}\frac{\partial \ln Z_\beta}{\partial \beta} &= \frac{1}{Z_\beta}\frac{\partial Z_\beta}{\partial \beta} = \frac{1}{Z_\beta}\frac{\partial}{\partial \beta}\int_\Theta q_\beta(\theta) d\theta \\ &= \frac{1}{Z_\beta}\int_\Theta \frac{\partial q_\beta(\theta)}{\partial \beta} d\theta = \int_\Theta \frac{1}{q_\beta(\theta)}\frac{\partial q_\beta(\theta)}{\partial \beta}\frac{q_\beta(\theta)}{Z_\beta} d\theta \\ &= \int_\Theta \frac{\partial \ln q_\beta(\theta)}{\partial \beta} p_\beta(\theta) d\theta = E_{p_\beta(\theta)}\left[\frac{\partial \ln q_\beta(\theta)}{\partial \beta}\right], \quad (4.18)\end{aligned}$$

where $E_{p_\beta(\theta)}[\cdots]$ is the expectation with respect to $p_\beta(\theta)$. Defining the *potential*

$$\mathcal{U}(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta},$$

we obtain

$$\frac{\partial \ln Z_\beta}{\partial \beta} = E_{p_\beta(\theta)}\left[\mathcal{U}\right].$$

Integrating over $[0,1]$ yields the log-ratio μ:

$$\mu = \ln Z_1 - \ln Z_0 = \int_0^1 \frac{\partial \ln Z_\beta}{\partial \beta} d\beta = \int_0^1 E_{p_\beta(\theta)}\left[\mathcal{U}\right] d\beta.$$

To compute this integral, a Markov chain Monte Carlo simulation is usually run for particular values of β, in which q_β is used as an unnormalised density in the Metropolis-Hastings ratio. By definition, this produces a sample from p_β. Expectations of the potential can then be estimated as averages on this sample. This computation is repeated for a series of values of β spaces between 0 and 1, which implies running a separate chain for each value of β.

The log-ratio μ can then be estimated by numerical integration using trapezoidal (as in Friel and Pettitt 2006) or Simpson's scheme (as in Lartillot and Philippe 2006).

A particular integration scheme employed in this thesis is called *annealing-melting integration* according to Lartillot and Philippe (2006) or *power-posteriors integration* according to Friel and Pettitt (2006).

Assuming that $q_0(\theta)$ above is the prior $p(\theta|M)$, and $q_1(\theta)$ is the unnormalised posterior $p(D|M,\theta)p(\theta|M)$, and the corresponding normalisation constants are $Z_0 = 1$ (as the prior is already normalised) and $Z_1 = p(D|M)$, the resulting log-ratio μ is the logarithm of the marginal likelihood.

Defining $q_\beta(\theta)$ as a path in the probability densities space which connects the prior and the posterior:

$$q_\beta(\theta) = p(D|M,\theta)^\beta p(\theta|M),$$

the potential takes a simple form:

$$\mathcal{U}(\theta) = \frac{\partial \ln q_\beta(\theta)}{\partial \beta} = \ln p(D|M,\theta).$$

And the logarithm of the marginal likelihood we are seeking an estimate for is

$$\ln p(D|M) = \mu = \ln Z_1 - \ln Z_0 = \int_0^1 E_{p_\beta(\theta)}\left[\ln p(D|M,\theta)\right] d\beta.$$

The stability improvement was explained by Gelman (1998) by using "bridge" densities to effectively shorten the distances between target densities Z_0 and Z_1, distances that are responsible for large errors with the standard importance sampling methods.

There are a number of ways to select a schedule for β to estimate this integral. In the case studies considered in this chapter, we use the schedule proposed by Friel and Pettitt (2006), and select these values as

$$\beta_i = a_i^c,\ a_i = \frac{i}{N},\ i = 0, \ldots, N.$$

Good results can usually be achieved with $N \in [20, 100]$ and $c = 4$ or $c = 5$.

In Example 4.4 we compare four methods of estimating the marginal likelihoods by calculating the error of the Bayes factor estimate for two simple linear regression models.

Example 4.4 (Model comparison for linear regression models.)

Williams (1959) described a linear regression example which has traditionally been used as a benchmark for model comparison approaches. Table 4.3 describes the maximum compression strength parallel to the grain y_i, the density x_i, and the resin-adjusted density z_i for 42 specimens of radiata pine. This dataset has been examined in (Han and Carlin 2001), (Carlin and Chib 1995), (Bartolucci and Scaccia 2004) and (Friel and Pettitt 2006), where they compared several methods to estimate the Bayes factor between two non-nested competing models.

Two competing models are the following:

$$\begin{aligned} M_1: &\quad y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2). \\ M_2: &\quad y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i,\ \eta_i \sim N(0, \tau^2). \end{aligned}$$

The following priors have been used in the above studies: $N(3000, 10^6)$ for α and γ, $N(185, 10^4)$ for β and δ, and $IG(3, 1/(2 \cdot 300^2))$ for σ^2 and τ^2, where $IG(a, b)$ is an inverse Gamma distribution with density

$$f(x) = \frac{1}{exp(1/bx)\Gamma(a)b^a x^{a+1}}.$$

Green and O'Hagan (1998) computed the Bayes factor $B_{21} = 4862$ by brute force numerical integration.

i	y_i	x_i	z_i	i	y_i	x_i	z_i	i	y_i	x_i	z_i
1	3040	29.2	25.4	15	2250	27.5	23.8	29	1670	22.1	21.3
2	2470	24.7	22.2	16	2650	25.6	25.3	30	3310	29.2	28.5
3	3610	32.3	32.2	17	4970	34.5	34.2	31	3450	30.1	29.2
4	3480	31.3	31.0	18	2620	26.2	25.7	32	3600	31.4	31.4
5	3810	31.5	30.9	19	2900	26.7	26.4	33	2850	26.7	25.9
6	2330	24.5	23.9	20	1670	21.1	20.0	34	1590	22.1	21.4
7	1800	19.9	19.2	21	2540	24.1	23.9	35	3770	30.3	29.8
8	3110	27.3	27.2	22	3840	30.7	30.7	36	3850	32.0	30.6
9	3160	27.1	26.3	23	3800	32.7	32.6	37	2480	23.2	22.6
10	2310	24.0	23.9	24	4600	32.6	32.5	38	3570	30.3	30.3
11	4360	33.8	33.2	25	1900	22.1	20.8	39	2620	29.9	23.8
12	1880	21.5	21.0	26	2530	25.3	23.1	40	1890	20.8	18.4
13	3670	32.2	29.0	27	2920	30.8	29.8	41	3030	33.2	29.4
14	1740	22.5	22.0	28	4990	38.9	38.1	42	3030	28.2	28.2

Table 4.3: Radiata pine dataset from (Williams 1959): y_i – maximum pine wood compression strength parallel to the grain, x_i – wood density, z_i – resin-adjusted wood density.

In this example we compare four marginal likelihood estimators on the example described above. We estimated the marginal likelihoods for alternative models M_1 and M_2 and computed the Bayes factor B_{21} using the prior arithmetic mean estimator, the posterior harmonic mean estimator, the annealing-melting integration and the Laplace approximation based estimator. Each estimate was evaluated 100 times, so the standard and relative errors of the estimates can be computed.

$$Mean = \hat{B}_{21} = \frac{1}{100}\sum_{i=1}^{100}\hat{B}_{21,i}$$

$$Standard\ error = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{B}_{21,i} - B_{21})^2}$$

$$Relative\ error = \frac{1}{B_{21}}\sqrt{\frac{1}{100}\sum_{i=1}^{100}(\hat{B}_{21,i} - B_{21})^2}$$

where B_{21} is the true value of the Bayes factor. The comparative overview of the estimates is demonstrated in Table 4.4.

MCMC sampling was performed in all of the compared estimators by the Metropolis-Hastings algorithm. The initial burn-in period was 1,000,000 samples. The length of the utilised sample was 500,000 samples.

True value for $\mathbf{B_{21}}$	4862
Prior arithmetic mean	**$\hat{\mathbf{B}}_{\mathbf{21}}$ estimate**
mean:	5052.58
standard error:	1229.93
relative error:	25.30%
Posterior harmonic mean	**$\hat{\mathbf{B}}_{\mathbf{21}}$ estimate**
mean:	6412.43
standard error:	6094.93
relative error:	125.36%
Annealing-Melting integration	**$\hat{\mathbf{B}}_{\mathbf{21}}$ estimate**
mean:	5007.63
standard error:	203.86
relative error:	4.19%
Laplace approximation	**$\hat{\mathbf{B}}_{\mathbf{21}}$ estimate**
mean:	3215.17
standard error:	2568.55
relative error:	52.83%

Table 4.4: Error comparison for different marginal likelihood estimators.

For the annealing-melting estimator we used 101 different values of β_i distributed within $[0, 1]$ as the following:

$$\beta_i = a_i^5, \quad a_i = \frac{i}{100}, \; i = 0, 1, \ldots, 100.$$

The true Bayes factor value (4862) lies within the error for each of the compared estimates, however using the annealing-melting integration provides significantly smaller estimate error.

We also computed AIC, BIC and DIC values for both of the alternative regression models to demonstrate that in this case the results achieved using large sample based criteria are consistent with the results obtained using Bayes factors:

$$\begin{aligned}
\mathrm{AIC}(M_1) &= 612.58064 \pm 5.5 \times 10^{-4} \\
\mathrm{BIC}(M_1) &= 617.79365 \pm 5.5 \times 10^{-4} \\
\mathrm{DIC}(M_1) &= 1522.8535 \pm 3.01 \times 10^{-2} \\
\\
\mathrm{AIC}(M_2) &= 595.52311 \pm 6.7 \times 10^{-4} \\
\mathrm{BIC}(M_2) &= 600.73612 \pm 6.7 \times 10^{-4} \\
\mathrm{DIC}(M_2) &= 1480.5757 \pm 3.27 \times 10^{-2}
\end{aligned}$$

Table 4.5: AIC, BIC and DIC values for the comparison of regression models.

The values of all the information criterions listed in Table 4.5 suggest preferring the second model over the first one, which is consistent with the result obtained using Bayes factor. The success of applying AIC, BIC and DIC is caused by the fact that the parameter posteriors for both of the alternative models are distributed almost normally.

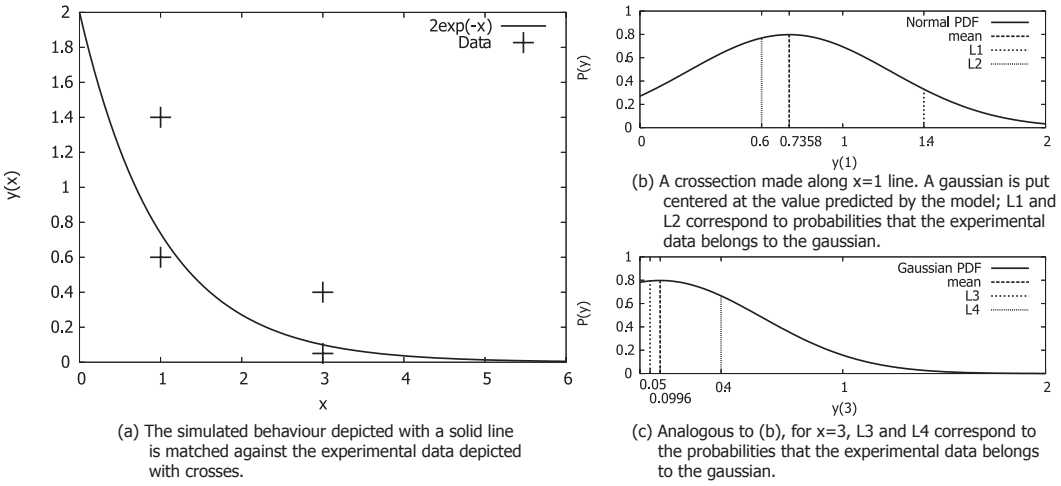
This example justifies the use of the annealing-melting integration in the case studies.

In this section we presented the fundamental concepts of Bayesian theory and introduced main methods that will be employed to implement an inferential framework for Systems Biology. In the next section we describe how such a framework can be built by making certain practical decisions on model implementation and MCMC sampling methods.

4.2 A Bayesian Inference Framework for Systems Biology

The Bayesian inferential methodology can be applied to model identification and model comparison (and therefore hypotheses testing) of any kind of models, as long as the likelihood function is provided.

Evaluating the likelihood is crucial for parameter inference and it is also required for estimation of marginal (integrated over parameter space) likelihoods for model comparison. The likelihood functions can be provided in a quite straightforward way for ODE models by using a normalised metric which provides the distance from the predicted behaviour to the observed data. For example, consider Figure 4.4. The solid line in Figure 4.4 (a) depicts a system behaviour predicted with a model, while points (depicted with crosses) corre-



(a) The simulated behaviour depicted with a solid line is matched against the experimental data depicted with crosses.

(b) A crossection made along x=1 line. A gaussian is put centered at the value predicted by the model; L1 and L2 correspond to probabilities that the experimental data belongs to the gaussian.

(c) Analogous to (b), for x=3, L3 and L4 correspond to the probabilities that the experimental data belongs to the gaussian.

Figure 4.4: Likelihood evaluation. The predicted behaviour is compared to the experimental data by putting a Gaussian centred at the predicted value at each time point and then evaluating the probability for the experimental data points under this Gaussian. L1, L2, L3, and L4 are such probabilities for four data points in this example. Multiplying these values together produces the overall likelihood value $L1 \cdot L2 \cdot L3 \cdot L4 = 0.1343$.

spond to the measured data. At each of the time points, where the experimental data is available, we put a Gaussian centred at the value predicted with the model and measure probabilities of the corresponding experimental data under this Gaussian. In Figures 4.4 (b) and (c) these probabilities are marked L1, L2, L3, L4. Multiplying these values together produces the overall likelihood value. For the general case a definition of such likelihood function for N data points is:

$$p(D|M,\theta) = \prod_{i=1}^{N} N_{D_i}(\phi(M,\theta,x_i),\sigma_i), \tag{4.19}$$

where x_i is the time when D_i was measured, σ_i is the variation of experimental error, and $\phi(M,\theta,x_i)$ produces the value predicted with model M using parameters θ for time point x_i where D_i was measured.

The likelihood function in Equation 4.19 assumes a normal distribution of errors. In some cases, e.g. when data is produced using microarrays, when the measurement error is not normally distributed, this distribution has to be replaced with a different probability distribution. Khanin et al. (2007) proposed to use the log-normal distribution to define likelihood function in the cases when experimental data were produced using microarrays. This decision was made to

model the microarray saturation effect which causes larger experimental errors when larger data values are observed.

Example 4.5 (The likelihood based on Equation (4.19))

Consider an example depicted in Figure 4.4. The experimental data is measured at two time points, when $x = 1$ and $x = 3$. For each of the time points we have two data:

$x = 1$	$x = 3$
$D_1 : y(x) = 0.6$	$D_3 : y(x) = 0.05$
$D_2 : y(x) = 1.4$	$D_4 : y(x) = 0.4$

We model the observed system with function

$$y(x) = 2e^{-x}. \tag{4.20}$$

Our model M is defined using equation 4.20 which has no parameters, so the set of model parameters θ is empty. The overall statistical model has four noise parameters $\sigma_1, \sigma_2, \sigma_3, \sigma_4$. The likelihood, according to (4.19)*, is:*

$$p(D|M, \theta) = \prod_{i=1}^{4} N_{D_i}(\phi(M, \theta, x_i), \sigma_i)$$

We chose $\sigma_i = 0.5$ for all $i \in \{1, 2, 3, 4\}$. And therefore:

$$p(D|M, \theta) = N_{D_1}(2e^{-1}, 0.5) \cdot N_{D_2}(2e^{-1}, 0.5) \cdot N_{D_3}(2e^{-3}, 0.5) \cdot N_{D_4}(2e^{-3}, 0.5) =$$
$$= L1 \cdot L2 \cdot L3 \cdot L4 = 0.7690 \cdot 0.3302 \cdot 0.7940 \cdot 0.6661 = 0.1343,$$

where L1, L2, L3, L4 are the values defined in Figures 4.4 (b) and (c).

Such a likelihood function can be defined for any models which are capable of producing deterministic quantitative system behaviour predictions, e.g. models based on ordinary or partial differential equations. Without restricting the generality of the approach, in this thesis we consider case studies which employ systems of ordinary differential equations (see Section 2.1) as models of biological systems. We discuss the technical problems experienced with ODE models of biological systems, and the solutions to tackle these problems in Section 4.2.2.

4.2.1 Biochemical Data Interpretation

The next problem is how the experimental data is interpreted, and how it can be used in a form suitable to formulate a likelihood function as in (4.19). The solution to this problem largely depends on a technique used to acquire the experimental data. We suggest that each particular case study has to include a dedicated discussion on how the laboratory measurements are interpreted. For some common techniques there already exist traditional ways of quantifying and interpreting the data.

In the case studies we consider in this chapter, the experimental technique employed is Western blotting (see Voet and Voet 1995). This technique is used to measure protein concentrations in studied cells. The main characteristics of this technique are:

- It only provides a relative measure of concentration of interest within a range of experiments. For example, if a range of experiments corresponds to different time points (e.g. 1 min, 2 min, 5 min after cells stimulation) it will only be possible to say the relative concentrations of the measured protein y, e.g. $y(2)/y(1)$, $y(5)/y(1)$, etc.

- It is only possible to compare the concentrations for the same protein, as for different proteins different antibodies with different affinity are used, and the results cannot be compared properly.

- It is difficult to compare the values from different experimental runs (different gels), as environmental conditions and experimental protocols can be slightly different. Measurement scaling is usually performed in such a case. The same experiment is performed in both experimental runs (alongside other experiments), assuming that it should provide the same outcome, the measured values are then scaled in a way that both results (from different runs) are matched. However, it introduces a significant degree of imprecision and uncertainty in the experimental data.

- Despite the fact that quantified values of the blots can be obtained with very high precision, the blots usually contain significant experimental errors. So, the data produced using this technique usually demonstrate significant variability.

- This experimental technique is a low throughput one: production of the data is slow and expensive. Therefore, we cannot expect large amounts of data to be available.

In our main case study (Case Study 2, Section 4.3.2) the methods for an appropriate data quantification were discussed with biologists. Starting with a table of relative values for protein concentration (concentrations of only one protein were measured in the laboratory) in different experimental conditions for different time points, all the relative values were scaled in the way that the biggest one became a 100% of the available protein, while the rest were expressed as a smaller percentage. When the models for the studied systems were built using ODEs, we introduced a likelihood function based on such relative values. E.g. if the model predicts that a concentration of the measured protein is 1 Molar out of 10 Molars theoretically achievable, we assign a value of 10% to the model variable used as a predictive output.

4.2.2 Technical Problems and Solutions

We experienced a number of technical problems while investigating the applications of Bayesian inference to ODE models of biological systems. All these problems are mainly due to the performance of initial value problem solvers. MCMC sampling requires the likelihood to be evaluated many times. In our case studies it was required to perform likelihood evaluations several million times. As a result, solving the initial value problems became a bottleneck of the sampling process.

The performance problems can be separated into two general classes:

1. Performance problems due to stiffness of a model.

2. Performance problems due to the number of initial value problems.

Consider both performance problem classes with possible technical solutions.

Stiffness

As soon as one deals with more than one first-order differential equation, the possibility of a *stiff* set of equations arise.

Definition 4.10: *An initial value problem is called stiff, if it causes numerical problems while solving it using non-specialised solvers.*

This definition is very vague, but there is no better definition. Stiffness occurs in a problem where there are two or more very different scales of the independent variable on which the dependent variables are changing. A conventional solver (such as Runge-Kutta method, see Press et al. (2002)) will require the integration step size to be reduced to very small values, therefore making very little progress through the large time intervals required for the solution. Since this problem was discovered, a number of specialised solvers have been developed to overcome the performance issues with stiff problems. The most widely used ones are Rosenbrock methods (see Press et al. 2002), and Bader-Deuflhard method (see Bader and Deuflhard 1983).

We use a specialised solver for stiff problems developed in the Lawrence-Livermore National Laboratory, USA (see Hindmarsh et al. 2005). This solver is publicly available and demonstrates performance compatible with proprietary patented algorithms (e.g. ones implemented in Matlab).

Number of Initial Value Problems

While computing the likelihood as in (4.19), it is sometimes possible to compute the predicted values $\phi(M, \theta, x_i)$ for different experimental conditions x_i in parallel.

For the model M and the set of model parameters θ, the values of $\phi(M, \theta, x_i)$ can be computed independently for different values of x_i. For the case studies presented in this chapter, we compute such predictions simultaneously using a distributed computing cluster. The results for separate values of x_i are then combined to produce the overall value of the likelihood function.

Our distributed algorithm runs a number of initial value problem solvers in parallel on a cluster, submitting the parameter values θ, the models M, and experimental conditions x_i to each of the solvers (see Figure 4.5). The results from the solvers are then substituted into equation (4.19) to compute the value of the likelihood.

In a case study considered in Section 4.3.2 this distributed computations algorithm allowed us to produce the inference results within reasonable time, while a single threaded version of the inference algorithm was too slow. For

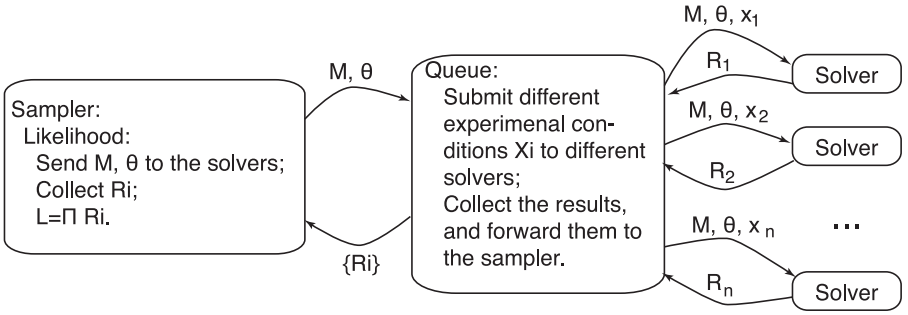


Figure 4.5: Distributed algorithm for the likelihood evaluation. The central sampler submits a proposed set of model parameters to several initial value problem solvers to evaluate the likelihood in different experimental conditions. The results are then returned to the central sampler where they are multiplied to produce the overall likelihood. $R_i = N_{D_i}(\phi(M, \theta, x_i), \sigma_i)$

each proposed value of model parameters we had to solve eleven initial value problems to compute the likelihood function. The distributed algorithm allowed us to solve these initial value problems in parallel. The required communication overhead, however, impacts the theoretical speedup factor and the practical benefit of this distributed algorithm is slightly lower than eleven times. In the case study in Section 4.3.2, the MCMC sampling took about three weeks on a distributed cluster of 42 workstations able to sustain the average performance of 5.267 GFLOP/s[7].

4.2.3 Parameter Inference

The first and core application of the Bayesian inferential methodology is the inference of model parameters from observed data. Given a parametric model M with parameters θ, and a set of experimental observations $D = \{d_1, d_2, \ldots, d_n\}$ in experimental conditions $X = \{x_1, x_2, \ldots, x_n\}$, the goal is to infer the distribution of parameters θ to reproduce the experimental observations.

We start with the prior distribution of the parameter values $p(M, \theta)$. Then we employ MCMC sampling to produce a sample from the parameter posterior. The posterior is a parameter distribution *learned* from the experimental data taking into account, and updating the prior knowledge.

Consider the details of the inference process:

[7]A GFLOP/s is a measure of computational system performance equal to a billion floating point arithmetical operations per second.

Prior Probabilities

In the absence of any experimental evidence we note that there is a large amount of existing literature and subjective knowledge based around studied systems. This information can be employed in defining the prior probabilities for the models and the associated parameter values. In the case when the priors are defined by $p(M, \theta)$, we can assume that the choice of the model structure or systems of equations describing the physical process will not affect *a priori* the range of feasible values, and thus the prior, on the associated parameters in this case $p(M, \theta) = p(M)p(\theta)$.

Structured priors can now be designed to best reflect the levels of subjective knowledge and uncertainty in the prior beliefs regarding the parameter values. In many cases such structured priors can be suggested based on the physical laws involved. In our case studies, we use Gamma distributions to define the priors of model parameters, as the laws employed in our models do not allow negative parameter values. Moreover, these distributions have non-zero density on all $\mathbb{R}^+$, which corresponds to our ignorance about the parameter values.

Posterior Distributions and Predictions

Having made our experimental observations and collected data D, we are in a position to update our strength of belief in the values of the model parameters *a posteriori*. Bayes' theorem provides us with the required posterior distribution in terms of our likelihood and prior distributions

$$p(\theta|X, D, M) = \frac{p(D|X, M, \theta)p(M_i, \theta)}{\int p(D|X, M, \theta)p(M_i, \theta)d\theta}. \tag{4.21}$$

This is an important level of inference as now the distribution $p(\theta|X, D, M)$ will provide us with valuable insights into the plausible range of the parameter values within the model. This improves vastly on the maximum likelihood estimates (see Section 2.8) as the posterior distribution over parameter values indicates how informative the experimental data has been in reducing our uncertainty in parameter values.

It also means that predictions on new experimental conditions can be made from model M, and again we can now integrate over our levels of posterior

uncertainty such that

$$\begin{aligned} p(D_{new}|X_{new}, M) &= \int p(D_{new}|X_{new}, M, \theta)p(\theta|X, D, M)d\theta \simeq \\ &\simeq \frac{1}{N}\sum_{i=1}^{N} p(D_{new}|X_{new}, M, \theta_i) \quad \theta_i \sim p(\theta|X, D, M). \end{aligned} \tag{4.22}$$

Now we have a distribution for the predictions and thus consistent levels of confidence can be assigned based on the above distribution with takes into account the prior beliefs and how these have been updated in the light of experimental evidence $\{X, D\}$.

The problem with multidimensional integrals can be resolved by employing MCMC sampling from the posterior distributions (see Section 4.1.3).

Parameter Identifiability Issues

Non-linear models are motivated by the laws of chemical kinetics, but the increased complexity means that such models, in some cases, become overparameterised. In such situations, some parameters of the model are unidentifiable. In other words, different combinations of unidentifiable parameters lead to the same likelihood, making it impossible to select single maximum likelihood estimates of the parameter values. This, however, does not cause a methodological problem for the Bayesian inferential approach, as all alternative combinations of non identifiable parameters are considered as a part of the parameter posterior.

Considering the hypotheses testing methodology, a consistent framework of Bayesian inference, as introduced in this chapter, demonstrates a significant advantage over maximum likelihood ratio methods due to considering such alternative combinations of parameter values. The identifiability issues are properly addressed when marginal likelihoods are evaluated, as information about the whole parameter distribution is taken into account. This contributes to the capability of Bayes factors to implement Occam's razor concept, as simple models with less unidentifiable parameters will be preferred to unnecessarily more complex models.

MCMC convergence problems, which are often observed when sampling from parameter posteriors of non identifiable models, can be resolved by adopting an adaptive proposal distribution idea as discussed in Section 4.1.3.

We conclude that all identifiability issues are naturally addressed in the

Bayesian inferential framework by considering full distributions of parameter posteriors. The corresponding uncertainty is naturally carried through Bayesian analysis methods, and can be resolved at the final stage if proper informative prior on all the parameters is specified. For detailed discussion of this issue, see (Hills 1987) and (Section 4.5 of Florens et al. 1990).

4.2.4 Hypotheses Testing

Consider the case where there are two competing hypotheses described by models M_1 and M_2. We employ Bayes factors (see Section 4.1.4) for model comparison, i.e.

$$\frac{p(M_1|D,X)}{p(M_2|D,X)} = \frac{p(D|X,M_1)p(M_1)}{p(D|X,M_2)p(M_2)} \tag{4.23}$$

In the case when no preference over competing hypotheses is suggested, we assign equal prior probabilities for both of the alternative models considered $p(M_1) = p(M_2)$. These probabilities can, however, be different, if the prior knowledge suggests preference of one hypothesis over another.

Now we can see that to compare our models the marginal likelihood under each model is required, i.e. for a kth indexed model

$$p(D|X,M_k) = \int_{\theta_k \in \Theta} p(D|X,M_k,\theta_k)p(M_k,\theta_k)d\theta_k. \tag{4.24}$$

Different methods of computing these marginal likelihoods are discussed in Section 4.1.5. We will demonstrate some of them in the case studies considered in Section 4.3.

4.3 Applications to Signal Transduction Pathways

In this section we develop our contribution with two case studies which demonstrate the applications of the Bayesian inference framework implementation. The first case study (Section 4.3.1) investigates alternative hypotheses formulated with artificially generated models and uses simulated data for parameter inference. This case study demonstrates how the Bayesian inference methods can be applied to ODE models. The second case study (Section 4.3.2) investigates the

complex problem of testing hypotheses about the structure of a realistic signal transduction pathway.

4.3.1 Case Study 1: Artificial Biochemical Networks

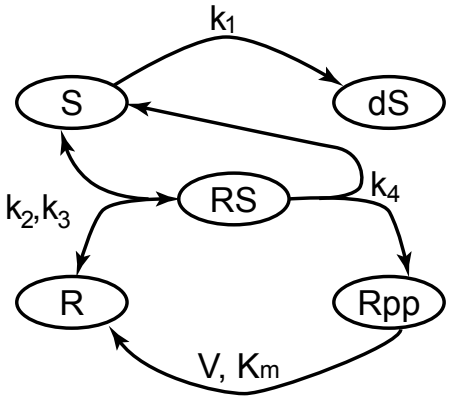
In this case study we consider four alternative models of a biochemical system. The models are artificially constructed to demonstrate the essence of the proposed methodology and demonstrate its main points and advantages on an example with a known result.

The schematic diagrams for the models are depicted in Figure 4.6. The entities in circles represent proteins involved in a biochemical network, while arrows correspond to biochemical reactions. Enzymatic behaviour is indicated by an arrow with a circle as head. For example, see Figure 4.6(b) where S in an enzyme for activation of R. Kinetic parameters of the reactions are depicted as text beside the arrows e.g. k_1, V_1. These networks represent realistic networks, and they all have a structure which is very common in nature.

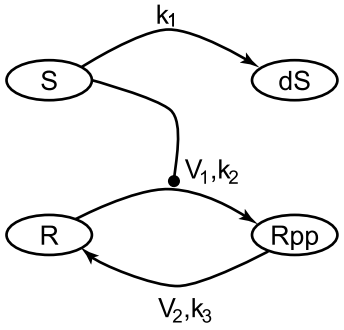
Model 1 This model defines a common motif of signalling pathways that is a stage in a signal transduction cascade. The input signal is represented by the concentration of protein S depicted in the top left of the diagram (see Figure 4.6(a)). This protein activates the next stage of the cascade by binding to protein R forming complex RS, and activating R into its phosphorylated form Rpp. Protein Rpp can then be deactivated. Model 1 also defines input signal degradation by converting protein S into its degraded form dS. This is added to demonstrate a nontrivial quantitative behaviour of the cascade stage.

To perform quantitative analysis of this system we formally describe a model using ordinary differential equations. All the proteins used in this model (depicted with ellipses in Figure 4.6(a)) will be represented as dependent variables in our ODE model. As we are interested in modelling and analysis of temporal behaviour, the independent variable is time.

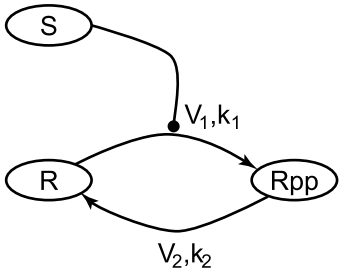
The dephosphorylation reaction $Rpp \rightarrow R$ is defined using the Michaelis-Menten kinetic law, while the rest of the reactions (arrows in Figure 4.6(a)) are defined using the Mass Action kinetic law with parameters depicted as textual remarks beside the arrows in model diagram (e.g. k_1, k_4). The following system



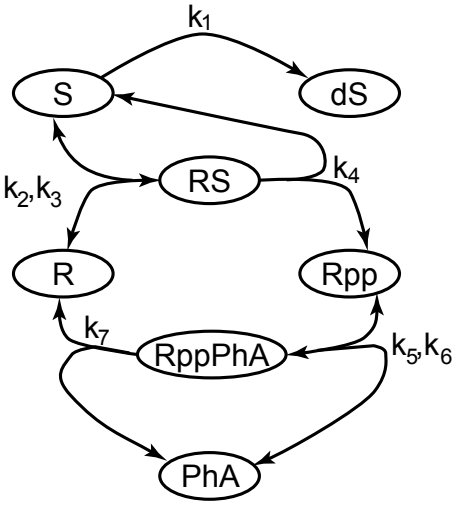
(a) Model 1: A model of a signal transduction cascade. Protein S represents the input signal. S can degrade to dS. At the same time S activates protein R from its inactive state, to an active state Rpp by binding and activation. Protein Rpp can then be deactivated. This model was used to generate the experimental data.



(b) Model 2: A simplified version of a signal transduction cascade. It represents the same process as described by Model 1, but a mechanistic description of the activation process is omitted and replaced with more general functions.



(c) Model 3: A model of a signalling cascade which is significantly different to the rest of the models in this case study. This model does not describe degradation of protein S. Our goal is to demonstrate that this model will gain much smaller evidential support at the hypotheses testing stage.



(d) Model 4: An overcomplicated version of Model 1. This model mechanistically describes how protein Rpp is deactivated by phosphotase PhA.

Figure 4.6: Models constructed for Case Study 1.

of ODEs defines this model:

$$\begin{cases} \dot{S} = -k_1 \cdot S - k_2 \cdot S \cdot R + k_3 \cdot RS + k_4 \cdot RS \\ \dot{dS} = k_1 \cdot S \\ \dot{R} = -k_2 \cdot S \cdot R + k_3 \cdot RS + \dfrac{V \cdot Rpp}{Km + Rpp} \\ \dot{RS} = k_2 \cdot S \cdot R - k_3 \cdot RS - k_4 \cdot RS \\ \dot{Rpp} = k_4 \cdot RS - \dfrac{V \cdot Rpp}{Km + Rpp} \end{cases}$$

We fix the initial values for the variables in this model. We could have added these initial values as additional parameters of our model if inferring these was required. For demonstration purposes we avoid introducing these as additional parameters, and define them as the following:

$$S|_{t=0} = 1 \qquad dS|_{t=0} = 0 \qquad Rpp|_{t=0} = 0$$
$$R|_{t=0} = 1 \qquad RS|_{t=0} = 0$$

This model has six kinetic parameters: $k_1 \ldots k_4, V, Km$.

Model 2 The model depicted in Figure 4.6(b) was constructed as a simplified representation of the signal transduction cascade stage. It essentially represents the same system as defined with Model 1, but uses different kinetic laws to define reactions.

The system of ODEs used in this model is

$$\begin{cases} \dot{S} = -k_1 \cdot S \\ \dot{dS} = k_1 \cdot S \\ \dot{R} = -\dfrac{V_1 \cdot R \cdot S}{k_2 + R} + \dfrac{V_2 \cdot Rpp}{k_3 + Rpp} \\ \dot{Rpp} = \dfrac{V_1 \cdot R \cdot S}{k_2 + R} - \dfrac{V_2 \cdot Rpp}{k_3 + Rpp} \end{cases}$$

The following initial values were chosen for this model

$$S|_{t=0} = 1 \qquad R|_{t=0} = 1$$
$$dS|_{t=0} = 0 \qquad Rpp|_{t=0} = 0.$$

The model has five kinetic parameters: k_1, k_2, k_3, V_1, V_2.

Model 3 The model depicted in Figure 4.6(c) is a version of Model 2 with degradation of protein S removed. As protein S cannot degrade, we would not observe any signal decrease in system behaviours, and therefore we would expect to see behaviours which are significantly different to the ones produced with other models in this case study. Our goal for this model is to demonstrate through hypotheses testing, that this model gains significantly smaller evidential support from data than the rest of the models.

The system of ODEs used in this model is

$$\begin{cases} \dot{S} = 0 \\ \dot{R} = -\dfrac{V_1 \cdot R \cdot S}{k_1 + R} + \dfrac{V_2 \cdot Rpp}{k_2 + Rpp} \\ \dot{Rpp} = \dfrac{V_1 \cdot R \cdot S}{k_1 + R} - \dfrac{V_2 \cdot Rpp}{k_2 + Rpp} \end{cases}$$

The following initial values were chosen for this model

$$S|_{t=0} = 1 \quad R|_{t=0} = 1 \quad Rpp|_{t=0} = 0.$$

The model has four kinetic parameters: k_1, k_2, V_1, V_2.

Model 4 The model depicted in Figure 4.6(d) is an overcomplicated version of Model 1. Phosphatase PhA depicted in the bottom of the diagram deactivates protein R. This model was constructed to demonstrate how it would be penalised for complexity according to Occam's razor concept in Bayesian hypotheses testing.

The system of ODEs used in this model is

$$\begin{cases} \dot{S} = -k_1 \cdot S - k_2 \cdot S \cdot R + k_3 \cdot RS + k_4 \cdot RS \\ \dot{dS} = k_1 \cdot S \\ \dot{R} = -k_2 \cdot S \cdot R + k_3 \cdot RS + k_7 \cdot RppPhA \\ \dot{RS} = k_2 \cdot S \cdot R - k_3 \cdot RS - k_4 \cdot RS \\ \dot{Rpp} = k_4 \cdot RS - k_5 \cdot Rpp \cdot PhA + k_6 \cdot RppPhA \\ \dot{PhA} = -k_5 \cdot Rpp \cdot PhA + k_6 \cdot RppPhA + k_7 \cdot RppPhA \\ \dot{RppPhA} = k_5 \cdot Rpp \cdot PhA - k_6 \cdot RppPhA - k_7 \cdot RppPhA \end{cases}$$

The following initial values were chosen for this model

$$\begin{aligned} S|_{t=0} &= 1 \qquad dS|_{t=0} = 0 \qquad Rpp|_{t=0} = 0 \\ R|_{t=0} &= 1 \qquad RS|_{t=0} = 0 \qquad PhA|_{t=0} = 1 \\ RppPhA|_{t=0} &= 0 \end{aligned}$$

This model has seven kinetic parameters: $k_1 \ldots k_7$.

Data Generation

In this case study we use data artificially generated from Model 1. To generate the data we simulated the behaviour of Model 1 with the following values for kinetic parameters:

$$\begin{aligned} k_1 &= 0.07 \qquad k_2 = 0.6 \qquad k_3 = 0.05 \\ k_4 &= 0.3 \qquad V = 0.017 \qquad Km = 0.3 \end{aligned}$$

by solving an initial value problem, and generated the time series of variable values (protein concentrations).

We decided to generate a data set for further experiments as a time series of *Rpp* values measured at the following time points: $t \in \{2s, 5s, 10s, 20s, 40s, 60s, 100s\}$. We added observation noise with variance 0.01 to the simulated values at each of the time points. The data set D contains twenty one samples. The obtained values are depicted in Figure 4.7.

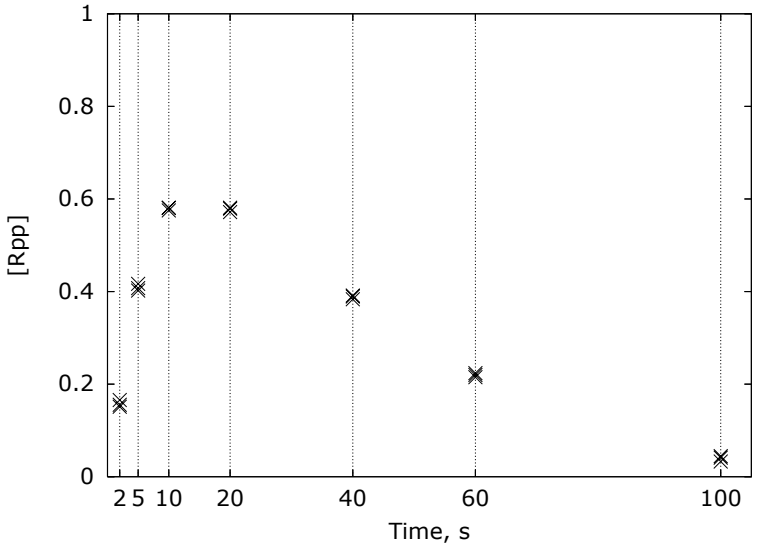


Figure 4.7: Data set generated from Model 1.

Overall Statistical Models

Performing Bayesian inference of model parameters and testing corresponding hypotheses involves likelihood estimation as defined in Section 4.2. We added only one additional noise parameter σ to each of the models, because we used the same noise value when the data was generated. The same noise parameter will be substituted to the normal distributions placed at each of the time points when evaluating the likelihood (4.19). The resulting statistical models contain the parameters given in Table 4.6.

Model	**Number of Parameters**	**Parameters**
Model 1	7	σ, $k_1 \dots k_4$, V, Km
Model 2	6	σ, $k_1 \dots k_3$, V_1, V_2
Model 3	5	σ, k_1, k_2, V_1, V_2
Model 4	8	σ, $k_1 \dots k_7$

Table 4.6: Parameters used in models for Case Study 1.

We discarded any information about the parameter values used to generate data for our experiments with Model 1; and defined the prior for model parameters. As none of the parameters can take a negative value, we defined the priors for all of the parameters of all the models to be distributed according to Gamma distribution $\Gamma(1, 3)$. This prior is depicted in Figure 4.8.

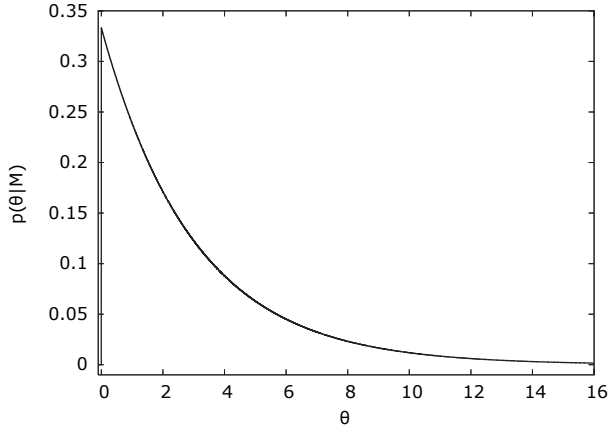


Figure 4.8: The shape of the prior distribution density used for all the parameters in Case Study 1.

Parameter Inference

MCMC sampling from parameter posteriors for all the models was performed. We used 20 instances of the Metropolis-Hastings sampler in parallel to ensure that the chains have mixed well before producing the actual sample for the parameter posteriors. Chain mixing was assessed using a method proposed by Gelman et al. (1995)[8]. An adaptive proposal distribution referenced in Section 4.1.3 was used to achieve faster convergence of Markov chains to the posterior distribution.

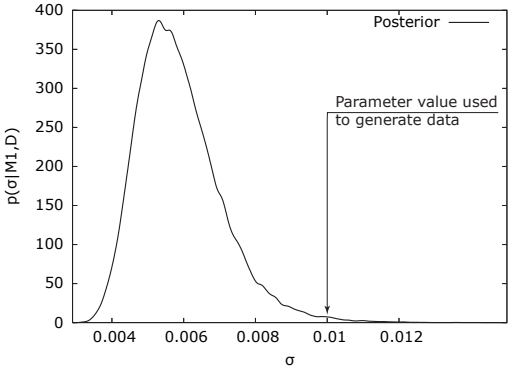
200,000 samples were generated from parameter posteriors after convergence of the chains was achieved[9]. The posterior for parameters of Model 1 is plotted in Figure 4.9. The parameter values used for data generation are also indicated in Figure 4.9, this demonstrates that the parameter posteriors were inferred correctly.

We can see that the posterior has diverged from the prior, which corresponds to the fact that some new information has been extracted from the experimental evidence. This additional information allowed us to update our beliefs, quantitative characteristics of which are expressed with these distributions.

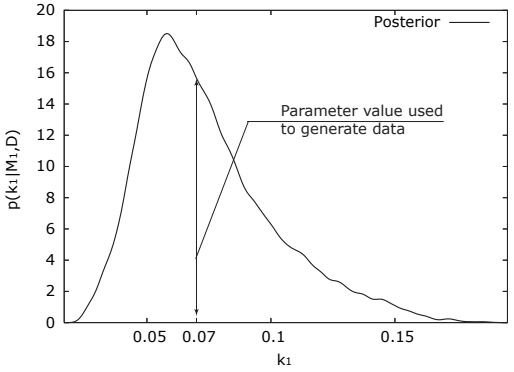
Notice, that the identified amount of observational noise was somewhat underestimated (posterior mean is at about 0.005, while the value used for data generation is 0.01). This, however, is not an inference problem, but a characteristic of this particular data set. As at each time point we generated only

[8]See Section 4.1.3 for more details.

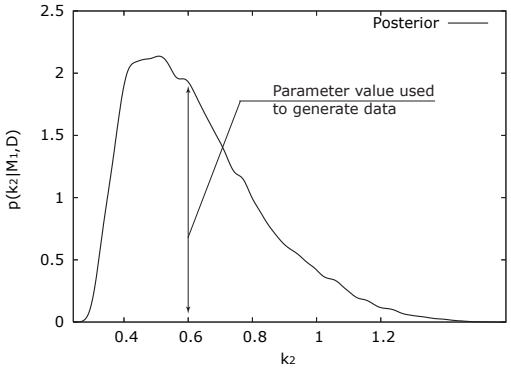
[9]Convergence of the chains was achieved after about 1,400,000 burn-in samples.



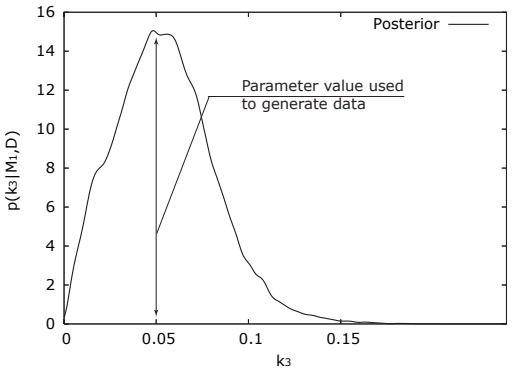
(a) Parameter σ of Model 1



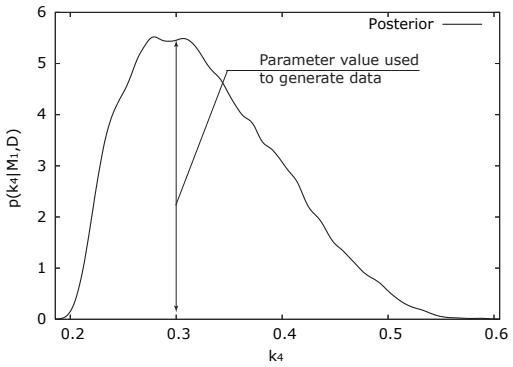
(b) Parameter k_1 of Model 1



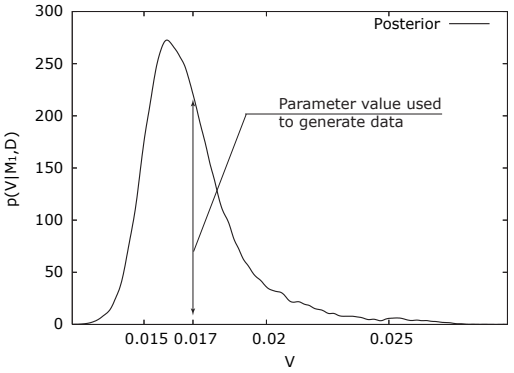
(c) Parameter k_2 of Model 1



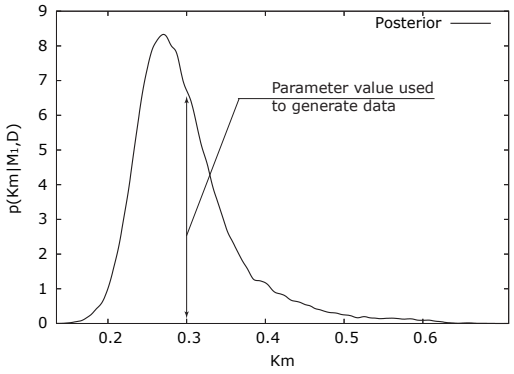
(d) Parameter k_3 of Model 1



(e) Parameter k_4 of Model 1



(f) Parameter V of Model 1



(g) Parameter Km of Model 1

Figure 4.9: Posterior distribution for kinetic parameters of Model 1.

three experimental values, the amount of real noise is not necessary equal to the true standard deviation of the noise generator due to the small sample size. To prove this we computed the amount of the observational noise, and our results are listed in Table 4.7. The average amount of the true observational noise is 0.005394 which agrees well with the infered results depicted in Figure 4.9(a).

Time Point	**Computed standard deviation of the dataset**
2	0.0075045
5	0.007388366
10	0.003265294
20	0.005031623
40	0.004247883
60	0.004336995
100	0.005981922

Table 4.7: The true amount of the observational noise in our dataset.

The distributions for parameters in Figure 4.9 are marginalised posterior distributions, and do not demonstrate the complete picture about the full posterior. In our case with 7 parameters for Model 1[10], the full posterior is a distribution in 7 dimensions. Nevertheless, we can investigate some properties of this posterior, such as its correlation. The resulting correlation matrix indicates possible dependencies between different model parameters. Correlation matrixes[11] for all four of our models are depicted with heat maps in Figures 4.10 – 4.13.

In Figure 4.10 one can see that parameters V and Km, and also k_2 and k_4 of Model 1 have high covariance values, and are likely to be dependent. We have plotted their joint posterior distributions (marginalised from the rest of model parameters) in Figure 4.14. Indeed, parameters V and Km are two parameters of the same reaction, and the found dependency is a well known characteristic of the Michaelis-Menten kinetic law. In the second case, for smaller values of k_2 larger values of k_4 have been proposed and vice versa. Consider the diagram for Model 1 in Figure 4.6(a). Parameter k_2 is the forward rate of binding protein R to S to form complex RS, and parameter k_4 is the rate of dissociation reaction for complex RS that produces Rpp and S. Maintaining the inferred dependency between parameter values allows the system to maintain the same flux from R

[10]The number of parameters for other models can be found in Table 4.6.

[11]Absolute values of the correlation coefficients were taken for these matrices, as we are not interested in the sign of correlation.

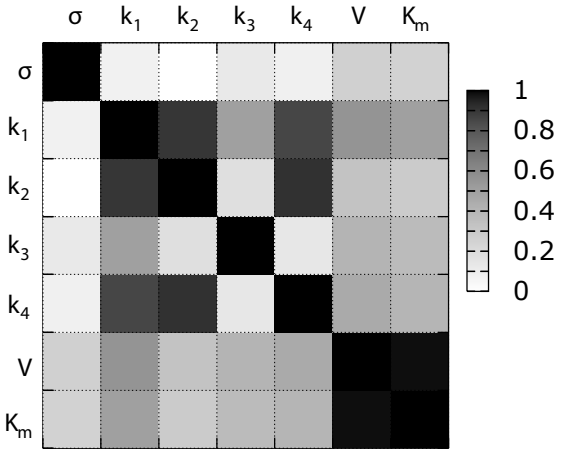


Figure 4.10: A heat map of the correlation matrix of the parameter posterior for Model 1.

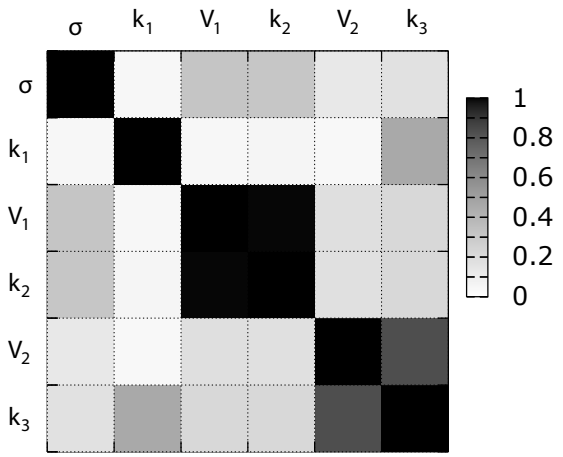


Figure 4.11: A heat map of the correlation matrix of the parameter posterior for Model 2.

to *Rpp*, that means the inferred dependency can be explained by the laws of chemical kinetics.

Note that the values for these parameters used for data generation:

$$k_2 = 0.4 \qquad k_4 = 0.4 \qquad V = 0.017 \qquad Km = 0.3$$

lie within the area of high posterior probability density in Figure 4.14.

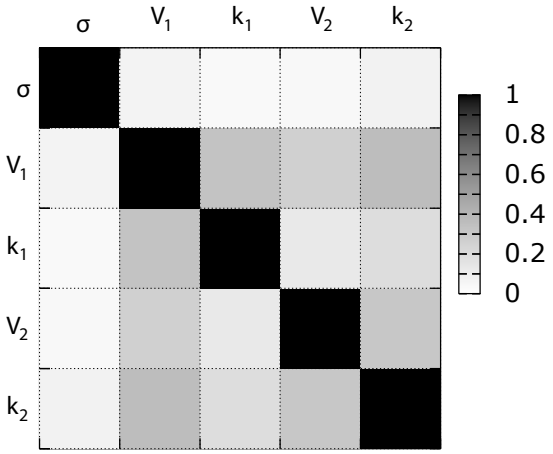


Figure 4.12: A heat map of the correlation matrix of the parameter posterior for Model 3.

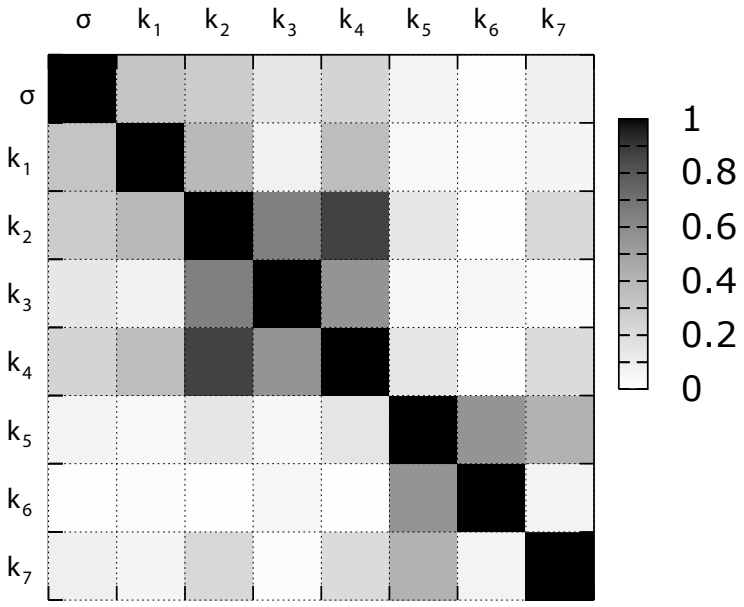
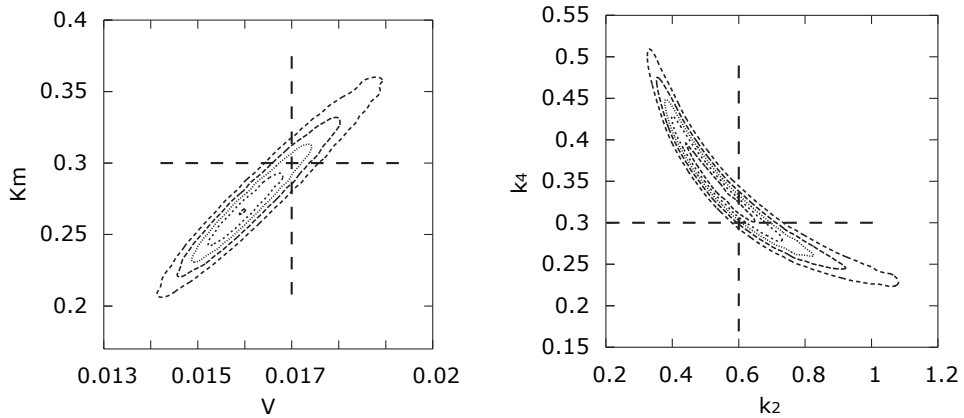


Figure 4.13: A heat map of the correlation matrix of the parameter posterior for Model 4.

Predictions

Having obtained samples from the posterior distribution of model parameters we now can generate predictions for the system behaviour. We consider two initial conditions for our prediction experiments. The first one is exactly the same condition which was used for data generation. Using the initial values for variables as defined in model description and parameter values from the identified



(a) Joint posterior distribution of V and Km

(b) Joint posterior distribution of k_2 and k_4

Figure 4.14: Joint posterior distribution for parameters of Model 1 inferred from simulated data.

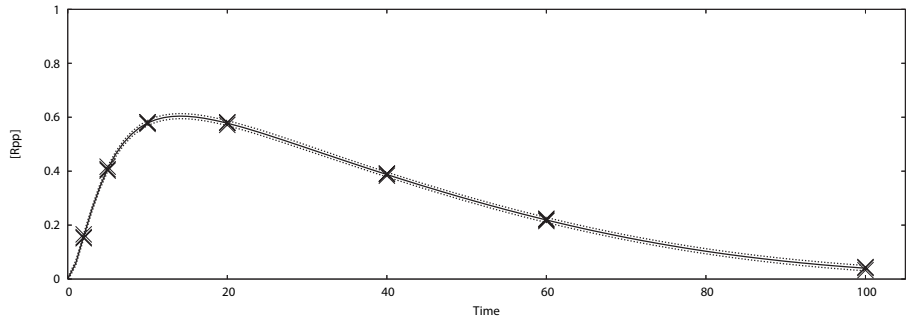
parameter posterior we produce a family of solutions for the concentration of *Rpp*, as we consider it to be the output of the models. The family of predicted behaviours is depicted in Figure 4.15. The solid line in the middle of each plot corresponds to the mean of the predicted behaviours family, while two bounding lines correspond to the standard deviation of the identified observation noise. Note, that different models predict different amounts of such noise.

The predictions made with Models 1, 2 and 4 (see Figures 4.15(a), 4.15(b) and 4.15(d)) match the data well, while the predictions made with the incorrect model (Model 3) in Figure 4.15(c) are not capable of reproducing the data properly. We will perform model comparison using Bayes factors in Section 4.3.1 to decide which one of the models is necessary for faithful modelling.

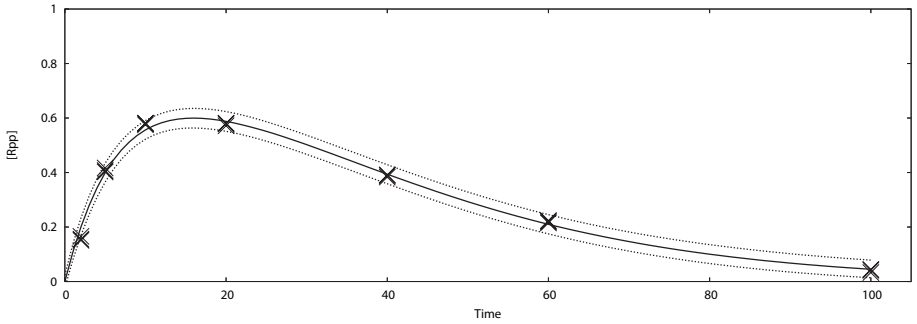
Additionally, we produce system behaviour predictions for a new experimental condition. Assume, we are interested in how the system behaves in the case when we use double the concentration of protein S is used, which corresponds to two times stronger input signal into our system. To perform these prediction experiments, we substituted a new initial value for variable S in our initial value problem:

$$S|_{t=0} = 2,$$

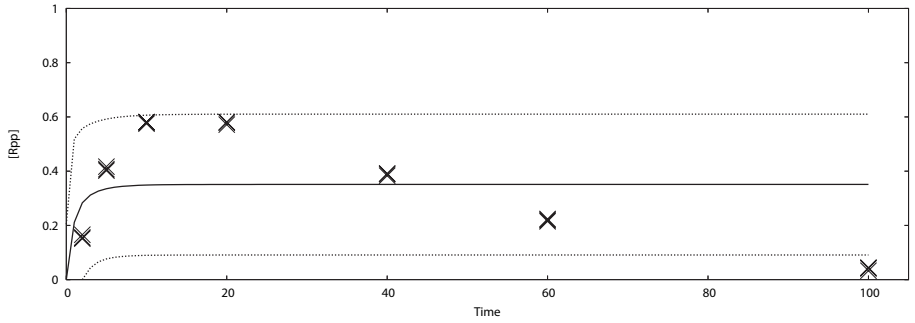
took the parameter values from the identified parameter posterior, and solved the new initial value problem 200,000 times producing the predictions for the concentrations of *Rpp*. The results of this prediction experiment is depicted in



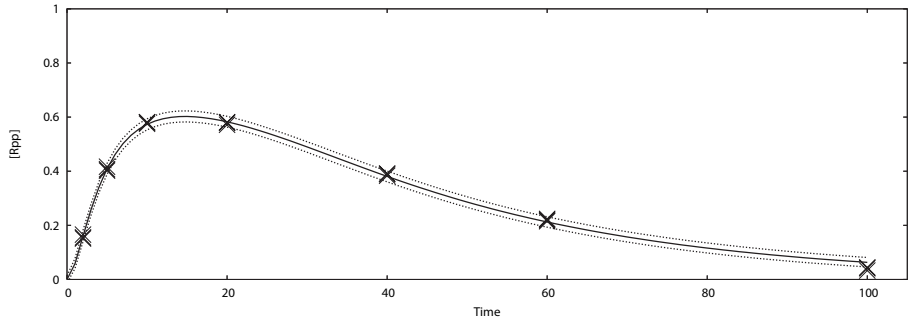
(a) Predictions made with Model 1 (originally used for data generation)



(b) Predictions made with Model 2 (simplified one)



(c) Predictions made with Model 3 (incorrect one)



(d) Predictions made with Model 4 (overcomplicated one)

Figure 4.15: Predictions for the original experiment used to produce data set D plotted against original data from data set D.

Figure 4.16.

Model Comparison

In the final stage of our case study, we compute estimates for marginal likelihoods of each of our models, and perform model comparison (and corresponding hypotheses testing) using Bayes factors.

We compare different estimators for marginal likelihoods by computing required Bayes factors using posterior harmonic means estimator, prior arithmetic means estimator, Laplace approximations and the annealing-melting integration (see Section 4.1.5). We use large (about 600,000 each) samples from each distribution $q_\beta(\theta)$. The chosen schedule for β in annealing-melting integration is defined as

$$\beta_i = a_i^4, \quad a_i = \frac{i}{40}, \quad i = 0 \ldots 40.$$

The obtained estimates for the marginal likelihoods for each of the models using each of the methods are given in Table 4.8.

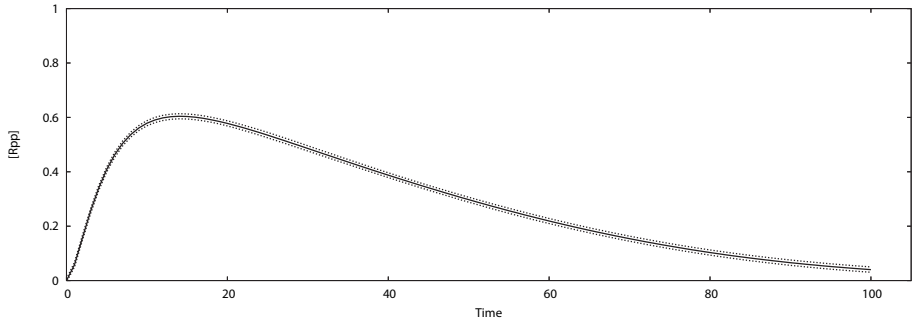
Once again, an estimator based on annealing-melting integration principles demonstrates significant superiority in the stability of obtained estimates. These estimates produce the following pairwise Bayes factors for model comparison:

$$\begin{aligned}
\frac{p(M_1|D)}{p(M_2|D)} &= 1.636 \times 10^{7} & \frac{p(M_2|D)}{p(M_1|D)} &= 6.114 \times 10^{-8} \\
\frac{p(M_1|D)}{p(M_3|D)} &= 2.377 \times 10^{20} & \frac{p(M_2|D)}{p(M_3|D)} &= 1.454 \times 10^{13} \\
\frac{p(M_1|D)}{p(M_4|D)} &= 5.862 \times 10^{4} & \frac{p(M_2|D)}{p(M_4|D)} &= 3.584 \times 10^{-3}
\end{aligned}$$

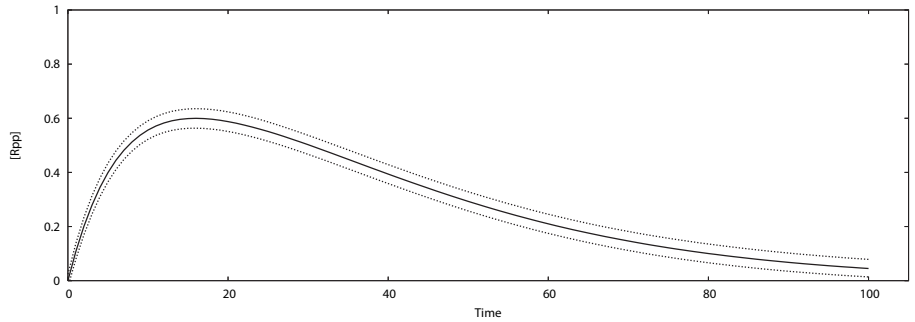
$$\begin{aligned}
\frac{p(M_3|D)}{p(M_1|D)} &= 4.206 \times 10^{-21} & \frac{p(M_4|D)}{p(M_1|D)} &= 1.706 \times 10^{-5} \\
\frac{p(M_3|D)}{p(M_2|D)} &= 6.880 \times 10^{-14} & \frac{p(M_4|D)}{p(M_2|D)} &= 2.790 \times 10^{2} \\
\frac{p(M_3|D)}{p(M_4|D)} &= 2.466 \times 10^{-16} & \frac{p(M_4|D)}{p(M_3|D)} &= 4.056 \times 10^{15}
\end{aligned}$$

these correspond to the following relative ranking of the four competing models:

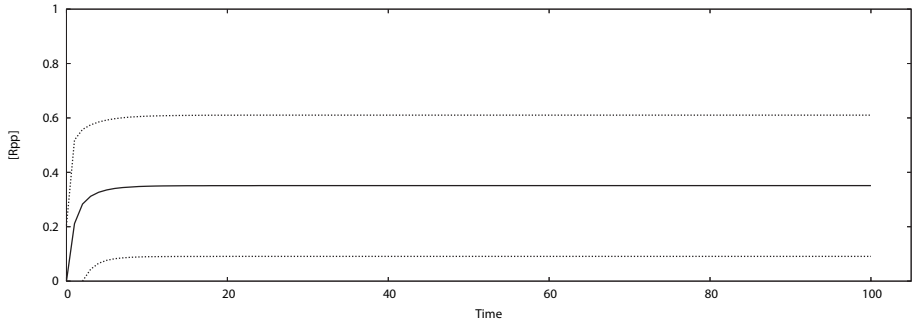
$$p(M_1|D) > p(M_4|D) > p(M_2|D) > p(M_3|D)$$



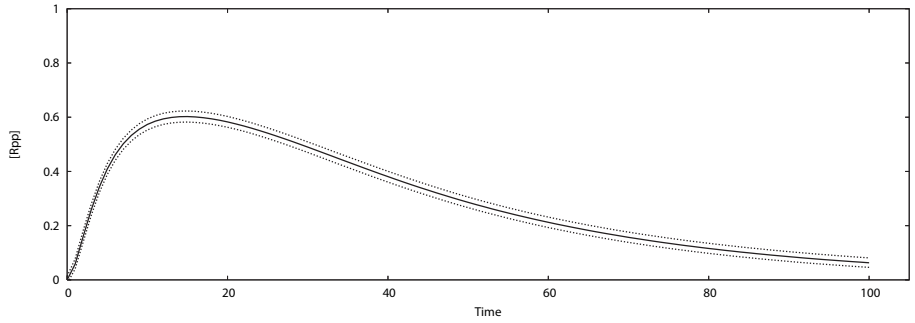
(a) Predictions made with Model 1 (originally used for data generation)



(b) Predictions made with Model 2 (simplified one)



(c) Predictions made with Model 3 (incorrect one)



(d) Predictions made with Model 4 (overcomplicated one)

Figure 4.16: Predictions for a new experimental condition.

Prior arithmetic means estimator		
Model	**Estimate mean**	**Standard deviation of the estimate**
Model 1	2.89×10^{6}	5.64×10^{6}
Model 2	2.40×10^{6}	3.78×10^{6}
Model 3	3.35×10^{-1}	3.21×10^{-2}
Model 4	2.68×10^{10}	7.11×10^{10}

Posterior harmonic means estimator		
Model	**Estimate mean**	**Standard deviation of the estimate**
Model 1	2.03×10^{31}	2.46×10^{31}
Model 2	2.80×10^{19}	1.70×10^{19}
Model 3	8.11	4.46
Model 4	1.64×10^{24}	1.29×10^{24}

Annealing-melting integration		
Model	**Estimate mean**	**Standard deviation of the estimate**
Model 1	7.98×10^{19}	1.49×10^{19}
Model 2	4.87×10^{12}	2.82×10^{11}
Model 3	3.36×10^{-1}	4.47×10^{-3}
Model 4	1.36×10^{15}	1.27×10^{14}

Laplace approximations		
Model	**Estimate mean**	**Standard deviation of the estimate**
Model 1	5.85×10^{18}	1.91×10^{18}
Model 2	1.14×10^{11}	2.30×10^{10}
Model 3	1.76×10^{-3}	1.12×10^{-4}
Model 4	4.67×10^{13}	4.69×10^{13}

Table 4.8: Estimated marginal likelihoods for models in Case Study 1.

The incorrect model (Model 3) gained the smallest evidential support and its marginal likelihood is dwarfed by the marginal likelihoods of other models. Model 1, which was used for data generation, has the maximal marginal likelihood, and therefore should be preferred over the rest of the models. Model 4, which was constructed to be an overcomplicated version of Model 1, has a smaller marginal likelihood value, and therefore is rated second. This demonstrates that Bayesian hypotheses testing accounts for the complexity of models, and implements Occam's razor principle.

According to the evidence support categories by Kass and Raftery (1995) defined in Table 4.2, the evidence suggests *"very strong"* preference of the original Model 1 over the rest of the models.

An interesting detail is that Laplace approximations based estimate produces the correct ordering of the models while maintaining second best relative error. The error, however, grows with the complexity of the model which corresponds to the known limitations of this estimator.

Applying information criteria approach for hypotheses testing in this case study produces correct model order when using AIC or BIC. However, DIC fails dramatically (producing completely inconsistent results with infinite variance of DIC score) due to the fact that posterior distributions are not normal. DIC is designed using an assumption that the mode of the posterior distribution matches its mean, which does not hold in this case.

Summary of Case Study 1

In the first case study we considered four alternative models. All the models were artificially constructed to allow testing of the proposed methodology on an example with a known result. We generated the "experimental" data used for parameter inference from one of the suggested models; selecting the original model as the most probable one on the model comparison stage, was a crucial result to demonstrate the correctness of our approach. We demonstrated how the principle of Occam's razor works in this framework, as the more complicated model was not preferred to the original one despite it being capable of reproducing the experimental data precisely enough. At the same time, the framework does not blindly select the simplest model, as the simplified alternative model was not preferred to the original due to poor likelihood of reproducing the experimental data. The control experiment using a structurally different model which was not capable of reproducing the general trends of system behaviour was also successful, as we demonstrated that this model was rated significantly lower than the rest of the alternatives.

Additionally to model comparison (and the underlying hypotheses testing) we performed parameter inference and behaviour predictions using all of the alternative models. Precisely as we expected, the parameter values used for initial data simulation were identified to belong to the high probability parts of the parameter posterior. The predictions made with the models using the inferred parameter posteriors demonstrate how model simulation traces reproduce the experimental data, but at the same time allow us to make predictions and plans for new experiments in yet untested experimental conditions.

In the next case study, we apply this inferential approach to a systems biology problem. The topic is a current problem in signal transduction pathways study about the structure of the ERK signalling network.

4.3.2 Case Study 2: The ERK Signal Transduction Pathway

In the second case study we consider an application of Bayesian inference to an ongoing research investigation. For this research we collaborated with the research group of Professor M. D. Houslay. Members of Prof. Houslay's group performed biochemical experiments in a laboratory, while we conducted the analysis of the experimental data and assessment of competing hypotheses.

The aim of this case study is to analyse the ERK signal transduction pathway using different biochemical interventions to decide which of the alternative hypotheses (presented below) about the pathway topology is better supported by the experimental evidence. This will allow us to test the alternative hypotheses about the pathway structure, make better predictions for future experiments, and will contribute to a better understanding of the underlying pathway.

Biological Background and Pathway Description

Epidermal Growth Factor (EGF) and Nerve Growth Factor (NGF) mediate the different biological processes of cellular proliferation and differentiation (Marshall 1995). It is known that NGF stimulation produces long term activation of Extracellular Signal-Regulated Kinase (ERK) whilst EGF provides a transient activation of ERK and both effects are mediated through the same ERK pathway (Marshall 1995 Kao et al. 2001). Both growth factors clearly employ ERK in a different manner to produce either cell differentiation or proliferation but the biochemical mechanisms underlying this diversity are unknown. The receptors for NGF and EGF are different and thus may explain why NGF and EGF mediate different biological responses. Indeed there are various Ras and Raf isoforms which may also lead to the observed differences in response and as both MEK and ERK are well conserved, they do not contribute to this difference. We are motivated to consider the role that the isoforms of Ras and Raf play in these different responses.

It is known that Ras activates c-Raf and generally accepted that Rap1 acti-

vates B-Raf (York et al. 1998 Wang et al. 2005), each of which directly activates MEK and subsequently ERK. In (Dhillon et al. 2002) it is observed that c-Raf is transiently active whilst B-Raf is constantly active. To explain the difference between NGF and EGF responses, our hypothesis is that NGF employs c-Raf and B-Raf to sustain the activation of ERK, whilst EGF employs only c-Raf. Therefore, if c-Raf is inhibited then we would be able to confirm this hypothesis. It is known that Protein Kinase A (PKA), which is activated by cyclic adenosine monophosphate (cAMP), inhibits c-Raf (Wu et al. 1993 Dhillon et al. 2002).

To investigate the ERK pathway our collaborators performed experiments on PC12 cells generated by Greene and Tischler (1976) from a transplantable rat adrenal pheochromocytoma line.

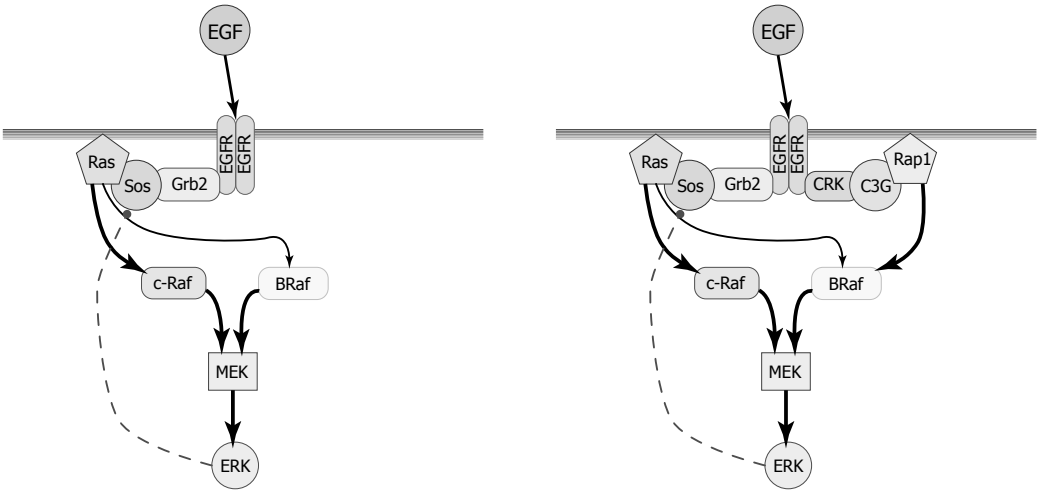
It is now appreciated that regulation of degradation of cAMP by cAMP phosphodiesterases (PDEs) plays a pivotal role in controlling intracellular cAMP concentrations and crosstalk with signalling pathways such as ERK (Houslay and Kolch 2000). Eight PDE families are involved in cAMP regulation, with the PDE3 and PDE4 families performing a dominant role in many cell types. Here we use cilostamide, a specific inhibitor of PDE3, to evaluate the effect of inhibiting cAMP degradation in PC12 cells by PDE3 on the activity status of ERK.

For this case study, we focused our interest on the activation of the ERK pathway by EGF which triggers specifically the proliferation of the cell. The mechanisms through which EGF activates cell proliferation are not fully understood and other pathways may also be responsible for such regulation. In this context we consider the cAMP pathway, which activates molecules that have been shown to be involved in ERK pathway regulation (Houslay and Kolch 2000). These molecules are PKA and a guanine nucleotide exchange factor (EPAC or cAMP-GEF). The biologists used a crosstalk between the ERK and the cAMP pathways to introduce biochemical interventions into the pathway dynamic behaviour, and thus were able to collect data which might be useful for model inference of the ERK signalling pathway.

Working Hypotheses

There are two alternative hypotheses on how the ERK pathway mediates the phosphorylation of ERK. The first one, supported by Brown et al. (2004) and Schoeberl et al. (2002), considers a single path of ERK activation by EGF sig-

nalling (see Figure 4.17(a)): EGF → EGFR → Grb2 → SOS → Ras → c-Raf → MEK → ERK. An alternative hypothesis, proposed in the paper by Kao et al. (2001), considers two ways of ERK activation by EGF signalling (see Figure 4.17(b)). This second hypothesis consider the second way of ERK activation through the EGF → EGFR → Crk → C3G → Rap1 → B-Raf → MEK → ERK cascade (right hand path in Figure 4.17(b)). Kao et al. (2001), Brown et al. (2004), and Schoeberl et al. (2002), we also consider the possibility that Ras activates B-Raf, as it was demonstrated by Marais et al. (1997).



(a) Pathway model supported by Brown et al. (2004) and Schoeberl et al. (2002). ERK (Extracellular Signal-Regulated Kinase) is at the bottom of the diagram. We study regulation of this kinase activity by the pathway depicted above it. EGF (Epidermal Growth Factor) is at the top of the diagram, as it initiates the activation of this particular pathway. This model (and the corresponding hypothesis) considers only one way of passing the signal – through the left and only branch (EGF, EGFR, Grb2, Sos, Ras, c-Raf or BRaf, MEK and ERK).

(b) Pathway model supported by Kao et al. (2001). As in Figure 4.17(a) ERK is at the bottom of the diagram, and EGF is at the top. There are two ways of passing the signal from EGF down to ERK: the first one is through the left branch (EGF, EGFR, Grb2, Sos, Ras, c-Raf or BRaf, MEK and ERK) and the second one is through the right branch (EGF, EGFR, CRK, C3G, Rap1, BRaf, MEK and ERK).

Figure 4.17: Hypotheses about the topology of the ERK signalling pathway

Crosstalk of the EGF signalling pathway with the cAMP pathway is achieved through small molecules activated by cAMP. These molecules are PKA and EPAC (Houslay and Kolch 2000 Baillie and Houslay 2005), and the structure of the crosstalk is depicted in Figure 4.18. The nature of this crosstalk can take a variety of different forms that are selectively utilised in different cell types and

may lead to cAMP either activating or inhibiting the ERK signalling or both. Here we model both inhibitory and stimulatory cAMP inputs that are mediated through c-Raf and B-Raf respectively. Experimentally, biologists demonstrated that PDE3 activity regulates cAMP input into the ERK signalling pathway in PC12 cells.

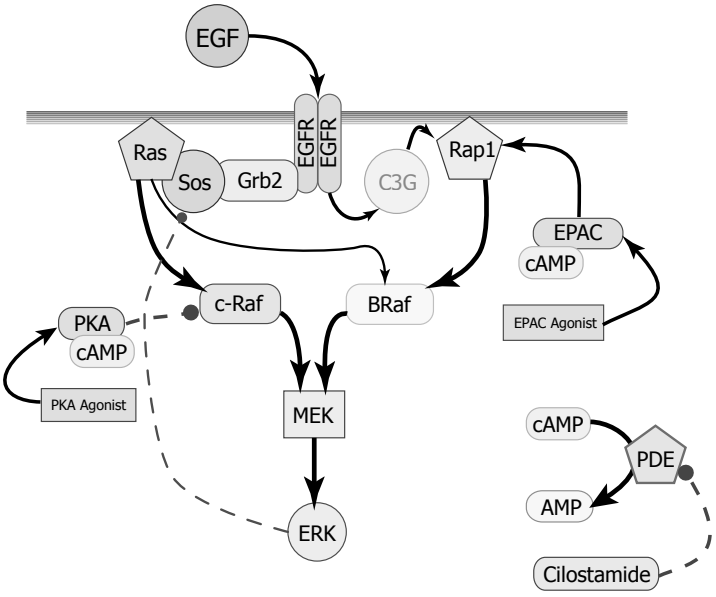


Figure 4.18: Other processes taking place in the cell can impact signalling through the ERK pathway. We particularly consider a case when another network of biochemical reactions, called the cAMP pathway, interacts with the signalling processes in the ERK pathway. cAMP (cyclic adenosine monophosphate) is an important second messenger involved in many biological processes. If the cell is stimulated with specific drugs targeted to regulate the levels of cAMP then the dynamics of signalling through the ERK pathway changes. We use three of such drugs: cilostamide, EPAC agonist and PKA agonist. The schematic interactions of these drugs with cAMP and the ERK pathway are depicted in this diagram.

A series of experiments have then been designed to assess the validity of the single and dual path hypotheses and the experimental data measured is then employed in devising single and dual path models to objectively assess the support of each hypothesis.

Experimental Data

The biologists performed a number of experiments using PC12 cells generated by Greene and Tischler (1976) from a transplantable rat adrenal pheochromocytoma line to investigate the dynamic behaviour of the ERK pathway. To do so, they measured the activation of ERK by quantifying the phosphorylation of this molecule using the Western blotting technique (see Voet and Voet 1995). Cells were starved for 3 hours in serum-free medium prior to any of the experiments. Then they were pre-treated for 10 minutes with cAMP analogues that activate specifically EPAC or PKA. Cilostamide was also used, which is a phosphodiesterase inhibitor that increases the level of cAMP within the cell. Next, the cells were stimulated with EGF for 0, 2, 5, 10, 20 and 40 minutes to activate the ERK signalling pathway. At the next stage the cells were lysed in lysis buffer containing protease and phosphatase inhibitors. Cell debris was removed by centrifugation. The protein concentration of each cell lysate was measured and normalised to the same concentration in each experiment to load the same amount of protein in each gel. Proteins were separated by NuPAGE®Novex 4-12% Bis-Tris gels electrophoresis and transferred on nitrocellulose membranes to perform Western blots. The membranes were immunoblotted with specific antibody directed against phosphorylated or non-phosphorylated ERK. These primary antibodies were detected using fluorescent secondary antibodies that emit at different wavelengths. The membranes were analysed using an infrared scanner (Licor, ODYSSEY) that detects the fluorescent secondary antibodies. As the infrared scanner is able to scan two bands of the spectrum at the same time, two different antibodies were used on the same gel. The first scanned band (green) corresponds to the total amount of ERK in the cell lysade, and the second one (red) corresponds to the amount of the phosphorylated from of ERK. Consequently, the ratio of the phosphorylated form to the total amount of ERK can be calculated.

EGF stimulation A number of experiments was performed by stimulating PC12 cells with EGF only. The cells were stimulated for 0, 2, 5, 10, 20 and 40 minutes after being starved for 3 hours in serum-free medium. 100 ng/ml of EGF was used for stimulation. This experiment was replicated 4 times. The results are given in Table 4.9.

Time	Experiment 1	Experiment 2	Experiment 3	Experiment 4
0	9.03%	3.78%	3.29%	3.34%
2	7.03%	4.77%	3.78%	2.75%
5	34.49%	34.49%	30.66%	24.31%
10	23.71%	20.26%	34.49%	34.49%
20	9.03%	6.67%	7.41%	5.09%
40	7.90%	5.10%	4.85%	3.91%

Table 4.9: Data for Case Study 2: control set, stimulation with EGF only.

Cilostamide stimulation PC12 cells were used after being starved for 3 hours in serum-free medium. The cells were then stimulated with Cilostamide (10 μM) for ten minutes. At the second stage the cells were stimulated with EGF (100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. This experiment was replicated 3 times. The results are given in Table 4.10.

Time	Experiment 1	Experiment 2	Experiment 3
0	5.04%	4.47%	3.74%
2	6.72%	3.63%	2.22%
5	61.72%	77.92%	52.82%
10	45.93%	87.52%	54.62%
20	16.96%	18.48%	11.11%
40	6.04%	5.96%	4.39%

Table 4.10: Data for Case Study 2: stimulation with Cilostamide and EGF.

A similar experiment was performed when no EGF stimulation was provided, and only Cilostamide (10 μM) was used. This experiment has shown no activation of ERK at all.

EPAC agonist stimulation The cells were stimulated with EPAC agonist (10 μM) for ten minutes. In the second stage the cells were stimulated with EGF (100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. This experiment was replicated 3 times. The results are given in Table 4.11.

An experiment when no EGF stimulation was provided was also performed, and only EPAC agonist (10 μM) was used. This experiment has shown no activation of ERK at all.

PKA agonist stimulation The cells were stimulated with PKA agonist (10 μM) for ten minutes. In the second stage the cells were stimulated with EGF

Time	Experiment 1	Experiment 2	Experiment 3
0	3.96%	2.84%	3.92%
2	7.63%	7.57%	6.60%
5	61.12%	56.63%	56.75%
10	39.83%	49.63%	51.54%
20	10.33%	9.00%	6.94%
40	5.40%	4.91%	4.14%

Table 4.11: Data for Case Study 2: stimulation with EPAC agonist and EGF.

(100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. This experiment was replicated 3 times. The results are given in Table 4.12.

Time	Experiment 1	Experiment 2	Experiment 3
0	4.16%	3.15%	3.40%
2	4.33%	3.19%	3.23%
5	24.69%	27.04%	16.98%
10	16.86%	21.82%	24.87%
20	8.03%	5.41%	4.82%
40	5.87%	4.51%	3.68%

Table 4.12: Data for Case Study 2: stimulation with PKA agonist and EGF.

A similar experiment was also performed when no EGF stimulation was provided, and only PKA agonist (10 μM) was used. This experiment has shown no activation of ERK at all.

Cilostamide and EPAC agonist stimulation The cells were stimulated with EPAC agonist and Cilostamide (10 μM each) for ten minutes. In the second stage the cells were stimulated with EGF (100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. This experiment was replicated 3 times. The results are given in Table 4.13.

Cilostamide and PKA agonist stimulation The cells were stimulated with PKA agonist and Cilostamide (10 μM each) for ten minutes. In the second stage the cells were stimulated with EGF (100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. This experiment was replicated 3 times. The results are given in Table 4.14.

Time	Experiment 1	Experiment 2	Experiment 3
0	5.89%	3.57%	4.02%
2	22.48%	22.99%	5.09%
5	73.63%	90.76%	66.43%
10	51.31%	46.28%	78.12%
20	10.88%	16.14%	11.64%
40	5.32%	6.84%	4.81%

Table 4.13: Data for Case Study 2: stimulation with EPAC agonist, Cilostamide, and EGF.

Time	Experiment 1	Experiment 2	Experiment 3
0	6.53%	4.16%	2.56%
2	9.73%	6.00%	5.64%
5	78.98%	80.19%	52.69%
10	53.42%	67.72%	49.20%
20	10.27%	12.86%	10.25%
40	8.17%	4.55%	4.23%

Table 4.14: Data for Case Study 2: stimulation with PKA agonist, Cilostamide, and EGF.

EPAC and PKA agonists stimulation The cells were stimulated with PKA and EPAC agonists (10 μM each) for ten minutes. In the second stage the cells were stimulated with EGF (100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. This experiment was replicated 3 times. The results are given in Table 4.15.

Time	Experiment 1	Experiment 2	Experiment 3
0	5.06%	4.11%	3.75%
2	10.36%	7.27%	4.77%
5	46.72%	38.85%	30.17%
10	50.00%	32.02%	27.09%
20	10.44%	6.52%	6.02%
40	6.65%	4.04%	4.55%

Table 4.15: Data for Case Study 2: stimulation with EPAC and PKA agonists; and EGF.

EPAC and PKA agonists, plus Cilostamide stimulation The cells were stimulated with PKA and EPAC agonists, and Cilostamide (10 μM each) for ten minutes. In the second stage the cells were stimulated with EGF (100 ng/ml) for 0, 2, 5, 10, 20, and 40 minutes. We replicated this experiment 3 times. The

results are given in Table 4.16.

Time	Experiment 1	Experiment 2	Experiment 3
0	6.92%	3.26%	3.51%
2	38.56%	26.06%	20.21%
5	100.00%	64.48%	63.24%
10	73.15%	58.24%	46.89%
20	9.12%	10.23%	9.41%
40	7.00%	6.16%	4.20%

Table 4.16: Data for Case Study 2: stimulation with EPAC and PKA agonists; Cilostamide and EGF.

Models

Three simplifications to the pathway models were made. These were done to reduce the size of the model by removing parts which are not covered by the experiments described above. The following simplifications of the models are adopted:

1. For simplification we consider that cilostamide is a direct activator of EPAC and PKA as it exerts its action solely by increasing cAMP;

2. cAMP itself (along with AMP, and PDE3) will not be considered as a part of the model, as we have no data available for this part of the pathway;

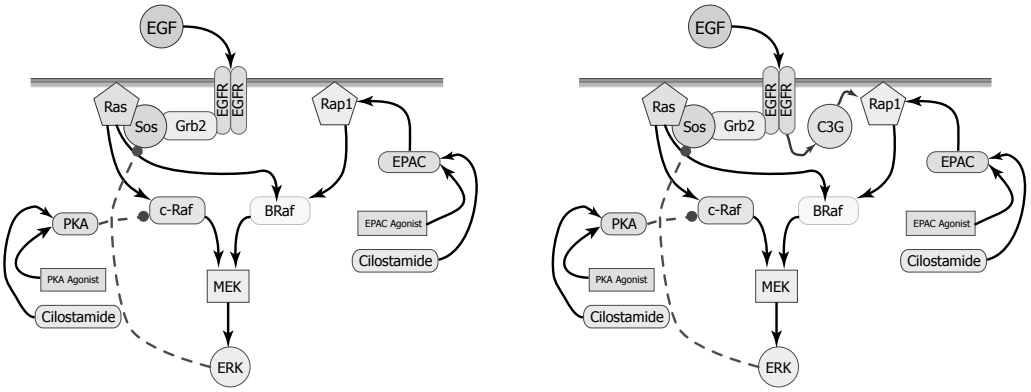
3. Receptor adaptor proteins activation process will be simplified (as in Brown et al.'s (2004) model), and defined as EGFR → Sos → Ras and EGFR → C3G → Rap1 pathways.

The goal of our analysis is to test the main hypothesis concerning the pathway: *Whether the pathway topology is best described by the utilisation of one or both branches.*

We consider two ODE models (see below: Model 1, Model 2) to test these hypotheses. Model 1 is defined using only one branch of the pathway, and Model 2 has both branches.

Model 1

This model considers only one path of ERK activation through the pathway stimulated with EGF. The model topology is depicted in Figure 4.19(a). The



(a) **Model 1:** The ERK pathway model with only one activation path defined. Rap1 can still be activated by EPAC, however there are no reactions which can activate Rap1 by EGFR receptor.

(b) Model 2: The ERK pathway model with two activation paths defined. In this case Rap1 can be activated by the receptor through C3G activation.

Figure 4.19: Models used in Case Study 2.

Rap1 → B-Raf → MEK → ERK path is also present in this model, but it cannot be activated by the receptor.

Model 2

This model considers both paths of ERK activation through the pathway stimulated with EGF. The model topology is depicted in Figure 4.19(b).

The overall statistical models used for inference also include noise parameters for the likelihood function depicted in equation 4.19. Following the previous definitions of the likelihood in Section 4.2, forty-seven noise parameters were added to each of the models. These noise parameters were inferred from the data in the same way as the rest of the model parameters.

Non-zero initial concentrations (measured in abstract concentration units, as no proper calibration is possible when using relative data) used in our models are:

$unboundEGFR\|_{t=0} = 500$	$inactiveSos\|_{t=0} = 1200$	$inactiveRas\|_{t=0} = 1200$
$inactivePKA\|_{t=0} = 1000$	$inactiveEPAC\|_{t=0} = 1000$	$inactiveRap1\|_{t=0} = 1200$
$BRaf\|_{t=0} = 1500$	$MEK\|_{t=0} = 3000$	$ERK\|_{t=0} = 10000$
$Gap\|_{t=0} = 2400$	$cRaf\|_{t=0} = 1500$	

We choose the concentrations of drugs used for the experiments as follows: the concentration of EGF is 1000, which corresponds to the fact that the cells

are treated with a large amount of EGF to ensure maximal activation. The concentrations for Cilostamide; EPAC and PKA agonists are chosen to be 100, which corresponds to the fact that the cells are treated with a tiny amount of those drugs to ensure that they do not cause signal activation by themselves.

The models are defined in SBML (Hucka et al. 2003) and in average contain 29 species and 28 reactions. They are omitted here due to their size, but electronically submitted with this theses as supplementary material. The models can also be obtained from the author.

An application of the Bayesian framework will now be demonstrated to perform posterior parameter inference of these models, which in its turn allows us to generate predictive posterior distributions of the system behaviour taking the overall underlying uncertainty into account. This analysis is then concluded with an application of Bayesian Inference for objective model comparison, which allows testing of the above formulated hypotheses about the pathway topology.

Analysis results

Wide Gamma priors were used for the parameters of our models which corresponds to our ignorance about the parameter values. The important properties of such Gamma priors are

1. Priors have only positive support, so negative values for the parameters will not appear in the posterior. This is quite reasonable, given the kinetic laws used to define our models.

2. Priors are not limited to the right (unlike uniform priors) and therefore vary large values for the parameters can be considered during the MCMC simulation.

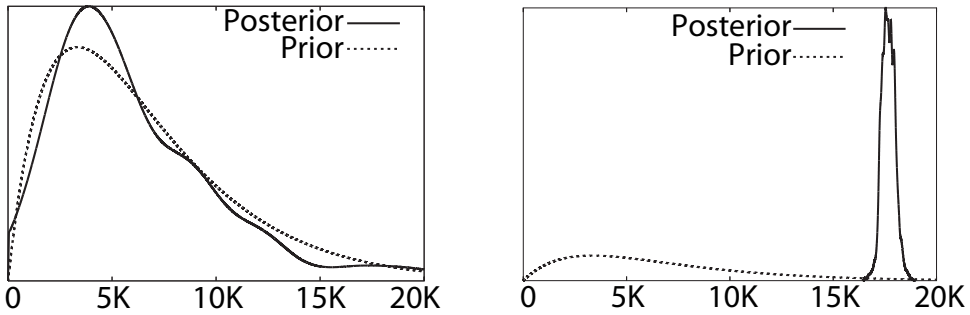
A description of the kinetic parameters and corresponding priors are listed in Appendix A.

We applied our implementation of the Bayesian inference algorithms on a case study of the ERK signal transduction pathway described above. One of the important features of the Bayesian inferential machinery is how knowledge is updated when new experimental evidence is considered. To illustrate this feature we performed posterior inference over parameters in two stages:

1. Wide Gamma priors for model parameters were used and posterior sampling was performed using a subset of the experimental protocols. The

experiments involving PKA agonist stimulation were intentionally omitted. Therefore, the process of PKA activation by the PKA agonist is not covered with experimental observations. The obtained posterior for the PKA activation parameter is depicted in Figure 4.20(a) which demonstrates that the posterior hardly diverged from the uninformative prior, so from a diagnostic perspective we can assess how informative a suite of experiments has been about certain parts of the network topology;

2. By including the data from PKA activation by the PKA agonist we see in Figure 4.20(b) a significant divergence of the posterior from the uninformed prior.



(a) Limited dataset. The posterior distribution is very similar to the prior distribution which means that there was no information in the dataset to update available knowledge about this parameter value.

(b) Complete dataset. In this case posterior diverged from the prior in a significant degree. This means that there was information which allowed us to identify the parameter more precisely.

Figure 4.20: Distributions for PKA activation by the PKA agonist parameter Km.

We also computed the correlation of the posterior samples for both models. Heat maps of the correlation matrices are depicted in Figure 4.21 and Figure 4.22.

Bayesian model comparison was performed to assess the proposed hypotheses about the pathway structure embodied in models 1 and 2. Bayes factors were computed for hypotheses testing.

The marginal likelihoods for the alternative models were estimated using annealing-melting integration (see Section 4.1.5):

$$\log(p(D|M_1)) = -1355.178 \quad \log(p(D|M_2)) = -1344.778445$$

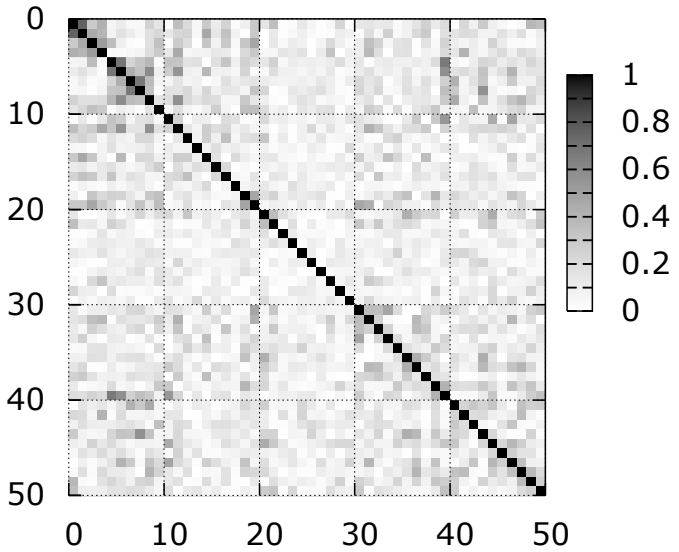


Figure 4.21: A heat map of the correlation matrix for the parameter posterior of Model 1.

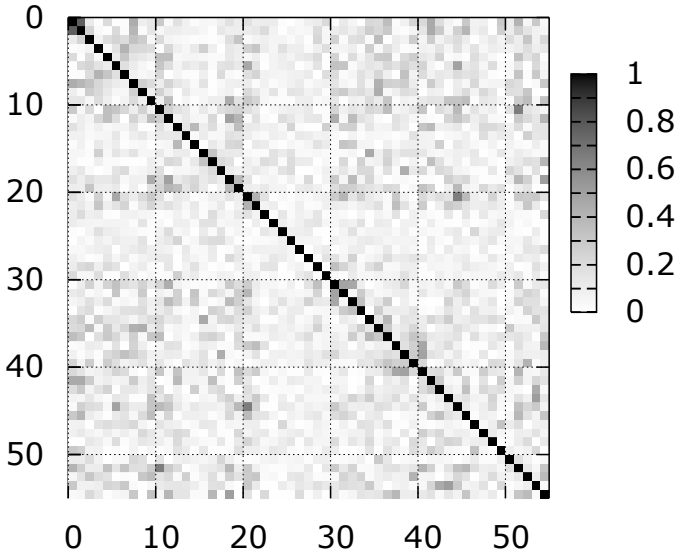


Figure 4.22: A heat map of the correlation matrix for the parameter posterior of Model 2.

The obtained Bayes factor:

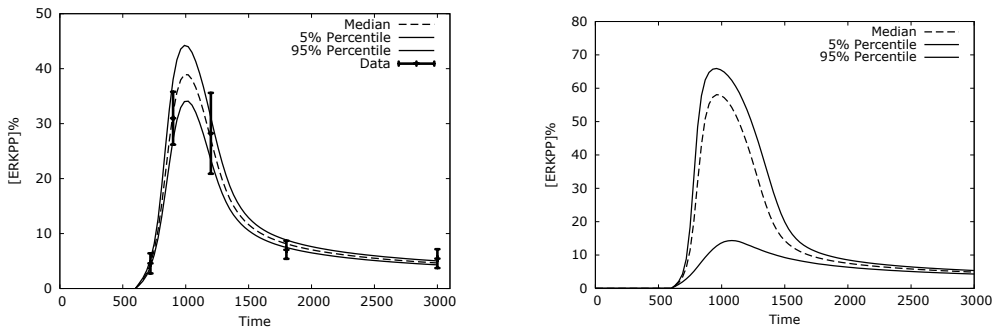
$$\frac{p(D|M_1)}{p(D|M_2)} \approx 3 \times 10^{-5}, \quad \frac{p(D|M_2)}{p(D|M_1)} \approx 32845.$$

The double logarithm of Bayes factor $p(D|M_2)/p(D|M_1)$ is approximately 20.8, which is bigger than 10. This, according to Kass and Raftery (1995) (see Table 4.2), suggests "Very Strong" preference of the two-branched pathway topology over the single-branched one.

After we identified two-branched Model 2 to be the best supported one, we resampled the model parameter values from the identified posterior distribution and generated some behavioural predictions using the preferred model. We simulated two experimental conditions for these predictions as following:

1. **Condition 1:** The cells are treated with PKA and EPAC agonists for 10 minutes and then stimulated with EGF (Epidermal growth factor). This corresponds to one of the performed experiments, so we will be able to compare model predictions to the real data.

2. **Condition 2:** C3G is knocked out using a specific siRNA; and the pathway is stimulated with Epidermal Growth Factor.



(a) Condition 1: PKA and EPAC agonists + EGF Stimulation. The widest part of the predicted behaviours distribution corresponds to a time interval where no data was collected, and therefore we still some uncertainty as to how the system behaves there.

(b) Condition 2: C3G knockout + EGF Stimulation. This is another example of how the identified model preserves the amount of uncertainty, as it predicts a very wide range of possible behaviours in such condition.

Figure 4.23: Behaviours predicted with Model 2 using the identified parameter values.

The predictions are depicted in Figure 4.23. The top line corresponds to the 95% percentile of the predicted behaviours distribution, and the bottom one to the 5% percentile. The median is depicted with a dashed line between the percentiles. The widest part of the envelope corresponds to the time intervals where the model behaviour cannot be predicted with enough certainty. These time intervals are located in the areas where model behaviour is transient and changes rapidly. When planning subsequent experiments these time intervals will be the best candidates in which to perform additional measurements.

We conclude that the proposed experimental condition when C3G knockout is simulated is a good candidate for subsequent experiments, as reducing the uncertainty in this condition should significantly improve our confidence.

Summary of Case Study 2

In this case study we applied the methods of Bayesian inference to an open problem in Systems Biology of a realistic size. We employed distributed computations system to produce samples from the parameter posteriors of the alternative models of the studied system. Each of the statistical models considered in this Case Study contains over 100 parameters, which causes performance difficulties when sampling from the parameter posteriors.

We managed to produce the posterior samples, behaviour predictions and compute Bayes factors for these models. We demonstrated that the experimental data, collected in the laboratory supervised by Prof. Houslay, supports the hypotheses of two signalling pathways involvement in this signalling network significantly stronger than the alternative one which considers only one pathway.

This case study, due to its complexity, also helped us to identify problems and bottlenecks of the proposed methodology. We discuss possible improvements of the methods in Chapter 5.

4.4 Discussion

In this chapter we have investigated how the methods of Bayesian inference can be applied to problems in Systems Biology.

We compared alternative methods to perform Bayesian inference and found that deterministic approximations to the parameter posterior around the maximum a posteriori estimate are quite simple to formulate, however, their appli-

cation is justified only when the parameter posterior is unimodal and approximately normal. This condition may not be satisfied when dealing with nonlinear ODE models. And ever though this method produces quite satisfactory results with some of the models, its correctness and reliability cannot be guaranteed in a general case.

We selected Markov chain Monte Carlo methods to perform Bayesian inference over ODE models, as these are the least constrained in applicability. The practical use of the MCMC samplers, however, exposed a number of technical challenges which had to be addressed to guarantee proper convergence of the Markov chains to the target distribution within reasonable time. We employed parallel sampling of multiple chains with convergence monitoring, and adaptive proposal distributions to address these technical challenges.

Evidential hypotheses testing in turn presented another methodological challenge as it requires estimation of the marginal likelihoods integrated over the whole parameter space. We had to employ the latest developments in applied statistics, such as thermodynamic integration methods built upon the path sampling ideas) to produce stable (repeatable with small variation of the result) estimates of the marginal likelihoods. At the same time we have demonstrated that alternative methods for model comparison, such as information criteria, cannot guarantee the correct result on a general case due to the methodological restrictions.

The selected methods were applied to actual problems in Systems Biology, demonstrating applicability, tractability and the value of the proposed approach. The results were produced taking the uncertainty of the data and available knowledge into account; this was not achievable using traditional methods based on maximum likelihood estimates.

Summary

In this chapter we developed a Bayesian inferential framework to enable quantitative plausible reasoning about models of biological systems. The methods of Bayesian inference were demonstrated, with a discussion about how model parameters can be inferred from the data, and how different models can be compared for evidence driven hypotheses testing. Through two case studies, we applied Bayesian inference methods to models of biochemical pathways and

demonstrated how this approach supports the scientific method in biological research.

The first case study describes four alternative models of a small biochemical network which were specifically built to demonstrate the proposed methods on a small example with a known solution. We generated the experimental data from the first of four proposed models by adding some noise to the selected parameter values. Then, we discarded all the information about the parameter values used for data simulation, and assigned wide Gamma priors to the model parameters. Parameter posteriors were then inferred using these priors and the data simulated from the first model. Inferred parameter posteriors for the first model essentially support the original values used for data simulation: this confirms the correctness of the chosen method. We also performed Bayesian model comparison to test alternative hypotheses expressed with the models. The correct model gained the most of evidence support from the simulated data.

In Case Study 2, we performed Bayesian inference and analysis of an open problem. We consider two alternative hypotheses about the topology of the ERK signalling pathway. The first hypothesis assumes that only one of the signalling branches is involved in passing the signal from the EGF receptor. The second hypothesis assumes that two parallel branches are used. Starting with wide uninformative parameter priors, we performed Bayesian inference of the model parameters; and then estimated a Bayes factor to test the hypotheses about the pathway topology. The two branched model has been found to be "significantly stronger" (in terms of Kass and Raftery 1995) supported by the experimental evidence than the single-branched one.

In both of our case studies we produced possible behaviour predictions from the inferred models.

Chapter 5

Conclusions and Further Work

Overview

In this chapter we summarise the contributions of our work, and describe its limitations. We also discuss the future of our work, including possible improvements and extensions to the proposed methodology as well as other work which may be inspired by it.

5.1 Conclusions

Reasoning based on available evidence is the foundation for consistent research in life sciences.

This thesis demonstrates how mathematical models can be used to describe hypotheses about the structure of biological systems, and how the methods of probabilistic reasoning and inference can be used to test alternative hypotheses and perform plausible reasoning in uncertain conditions.

A probabilistic reasoning methodology proposed in Chapter 3 enables quantitative logical analysis of models of biological systems. The proposed methodology suggests modelling biological systems using continuous time Markov chains. Probabilistic model checking is used to verify logical properties of biological models and to produce estimates of system behaviours. The consistency of the proposed modelling technique is demonstrated as an approximation of ODE solutions. A practical application of the reasoning methodology is demonstrated on an example involving the RKIP inhibited ERK pathway. The main contributions made in the area of probabilistic modelling and model-based reasoning

are the following. The problem of faithful stochastic modelling of large chemical populations is investigated, demonstrating that traditional modelling of individual molecules does not perform sufficiently well. Individual-based models either lead to enormous state spaces or largely misrepresent the dynamics of large populations. A population-based modelling approach is proposed, observational uncertainty is addressed by employing symbolic intervals to abstract from numerical values for model variables. Behaviour simulation and quantitative reasoning about the proposed models is performed using probabilistic symbolic model checking.

A Bayesian inferential approach to plausible reasoning about models of biological systems is considered in Chapter 4. A variety of algorithms which implement Bayesian inference methods are investigated and critically compared. Markov chain Monte Carlo methods are selected as the "gold standard" as these are based on the least constrained foundations. The most challenging problem is hypotheses testing using noisy experimental data. This requires the latest developments in applied statistics, such as path sampling methods, to be used to obtain stable results. We demonstrated the effectiveness of our methods with two case studies, one of which is an open problem in Systems Biology.

Our approach enables consideration of uncertain knowledge and data at all stages, this is not possible using traditional maximum likelihood methods.

Parameter inference for ODE models of biological systems is achieved by Markov chain Monte Carlo sampling using the Metropolis-Hastings algorithm. The identified parameter posteriors are then used to produce model behaviour predictions, taking the underlying uncertainty about the parameter values into account. Alternative models which correspond to competing hypotheses about the system structure are then systematically compared by the evidence support, evaluated as the marginal (integrated) likelihoods of the experimental data conditioned by each model. The employed hypotheses testing approach is consistent with the scientific method paradigm, and implements Occam's razor principle which allows the simplest model sufficiently explaining observed data to be selected.

We demonstrate applications of the proposed Bayesian inferential framework implementation on two case studies in the area of signal transduction pathways. The first case study operates on artificially designed models and generated data to demonstrate consistency of the proposed methodology and introduce main

analysis capabilities on simple models where the correct answers are known. The second case study demonstrates an application of the proposed inferential methodology to an open problem in Systems Biology. In this study we perform model parameter inference and consequent hypotheses testing to determine which of the alternative hypotheses about the structure of the ERK signal transduction pathway is better supported by the evidence collected in a biochemical laboratory. The models considered in the second case study are significantly more complex than the ones traditionally considered in this context. Many solutions to nontrivial scientific challenges were investigated and adopted to enable sufficient performance of the inference algorithms. The main solutions which allowed inference over realistically large models are simulating several Markov chains in parallel monitoring their convergence to the common target distribution, use of adaptive proposal distributions which adjust their shape to match the current approximation to the posterior, use of marginal likelihood estimator based on thermodynamic integration principles.

The case studies have also demonstrated the main problems and bottlenecks of the current methods. We will discuss some ideas for improvements and extensions in the next section.

5.2 Further Work

There are three main areas where the work presented in this thesis can be significantly improved or extended.

The first opportunity for future improvement is to apply the inferential methodology discussed in Chapter 4 to the stochastic models of biological systems proposed in Chapter 3.

As structural models of biological systems defined with the PRISM language are parametric, it would be beneficial to implement inference methods for such stochastic models. But since the method for model behaviour prediction described in Chapter 3 provides only the mean estimate for system behaviour, it would not be appropriate to substitute this estimate directly into equation (4.19) defined in Section 4.2 (see page 112), because this would not address the variance of such predictions properly. A more sophisticated method of likelihood definition is required to implement inference over stochastic models. If a sound likelihood is defined for the stochastic models, we then could infer model

parameters for these models from experimental data, and consequently evidence-based model comparison and a sound description of uncertainty for parameter values. On the other hand, the models identified with the inferential methodology can then be analysed deductively using the reasoning methods discussed in Chapter 3.

The second area for further improvement is performance of the Monte Carlo samplers proposed in Chapter 4. Producing a sample from the posterior distribution of model parameters of a nonlinear model in parameter space of many dimensions is indeed the biggest challenge and the toughest bottleneck for the methodology proposed in Chapter 4. As mentioned in Section 4.2.2, the analysis of Case Study 2 (Section 4.3.2), the problem of a realistic size, has taken three weeks of computing time on a large computer cluster.

We propose to investigate and apply two of the promising methods for improving the sampling performance: Hamiltonian MCMC sampling (MacKay 2003) and population Monte Carlo sampling (Iba 2001). The former method utilises gradient information to reduce random walk behaviour, and, consequently, to reduce the time required to obtain effectively independent samples from the posterior distribution. The population-based Monte Carlo methods are also designed to improve sampling performance by simulating several random walks through the parameter space in parallel and allowing such parallel samplers to share the information about the shape of the target distribution. Evolutionary Monte Carlo algorithm (Liang and Wong 2001), for example, simulates a population of parallel Markov chains updating such a population using genetic operations of mutation and crossover. The chains are embedded at different temperature ladders to incorporate the attractive features of parallel tempering.

The third option for future extensions is development of a methodology to utilise expert knowledge for formulation of informative priors which then can be utilised more effectively (than noninformative ones) in inferential analysis.

When performing Bayesian inference and evidential hypotheses testing it is required to formulate the prior knowledge first. In the work presented in this thesis we use only highly uninformed assumptions to formulate the initial knowledge and corresponding priors. However, using informative priors for the hypotheses enables utilisation of some additional knowledge when performing the inference. Such knowledge will then be updated using information from experimental ev-

idence while the inference is performed. An example of an informative prior can be some information about a plausible model parameter value which can be estimated using additional experimentation using a different technique. In the case when this kind of more particular information about the parameter value is available, it is possible to formulate a more complex hypothesis assignment to the confidence distribution for this parameter: for example, a narrow Gamma distribution located at the estimated value with variance related to the experimental error. In such a case, the value estimated with an additional technique is efficiently defined as the most likely one, but other values are still taken into consideration if the experimental evidence strongly suggests alternative options.

Another valuable source of initial knowledge is published biological literature. The main challenge here, however, is not only to find appropriate articles and formalise the knowledge contained therein, but also to assign proper confidence distributions to the found statements. This problem is quite complex as there is inconsistent knowledge published in different articles, so alternative hypotheses should be taken into account. In the case where several alternatives are described in the published literature, we can consider the number of publications supporting each of the hypotheses as an initial confidence assignment for the prior. For example, the number of articles found in PubMed[1] weighted with a journal impact factor can be utilised as a relative probability for each of the hypotheses.

Expert questionnaires can also be used to collect subjective confidence to formulate informative priors. The experts can be asked to rate the alternative hypotheses on a semantic scale (saying that they would, for example, strongly prefer hypotheses X over hypotheses Y). The relative probabilities then can be assigned to alternative hypotheses on a scale similar to the one proposed in Table 4.2.

New work which can be developed on the foundation established in this thesis can be done along the following two directions:

The first one is to build upon the predictive possibilities of the uncertain models, especially the ones with parameter distributions inferred with Bayesian methodology. The next step in this direction can be investigation of optimal experimental design methods to suggest the most promising experiments for future investigation to be performed first. For example, in the cases when two

[1] A literature database for biomedical publications

alternative hypotheses cannot be distinguished effectively using the available experimental evidence the predictions can be drawn using each of the alternative models for a number of future experiments. Then, the experiment which proposes the largest difference in the predicted outcomes with different models should be performed first as it is likely to provide more valuable data for hypotheses testing.

The second direction is utilisation of approximate methods either to perform approximate inference or to guide the MCMC sampler during the initial random walk, thus optimising sampling performance. For example, Laplace approximations can be used to perform approximate parameter posterior inference, and then the approximate posterior distribution can be used as a proposal distribution for the Metropolis-Hastings sampler to produce a sample from the true posterior more effectively. Another method we plan to consider in our future work is usage of Gaussian processes (see Rasmussen and Williams 2006 Gibbs and MacKay 2000) for likelihood approximation. Such likelihood approximation can serve either as a quick guide for the initial random walk, as a sampling distribution for a part of particle population in population-based Monte Carlo methods, or as a proposal distribution for the Metropolis-Hastings sampler.

We plan to investigate possible improvements and usage of approximate methods, and also to address the issues of optimal experiment design in follow-up projects to this work.

Summary

Probabilistic methods for model-based reasoning and inference proposed in this thesis address the problem of reasoning and hypotheses testing for Systems Biology in uncertain conditions.

The main scientific contributions of this work are

1. A novel population-based stochastic modelling approach is proposed to model biological systems which involve large chemical populations.
2. Probabilistic reasoning is performed over the proposed stochastic models enabling logical analysis of possible model behaviours.
3. A Bayesian inferential framework implementation is developed to enable parameter inference and evidential model comparison on ODE models of biological systems.

4. Different methods for estimation of the marginal likelihoods are compared, and the method based on annealing-melting integration is selected and implemented to enable hypotheses testing using realistically sized nonlinear ODE models to describe alternative hypotheses.
5. The proposed methods are applied to perform modelling, reasoning and inference in the area of signal transduction pathways.

The work described in this thesis opens several directions for further extensions and improvements.

Appendix A

Model Parameters for Case Study 2

The following table defines the parameters for Model 1 used in Case Study 2:

Table A.1: Parameters for Model 1 in Case Study 2

Parameter index	Description	Prior
1	Sos inhibition by ERKPP, Kcat	$\Gamma(1.1111, 9.0)$
2	Sos inhibition by ERKPP, Km	$\Gamma(2.0, 3333.0)$
3	Sos activation, Kcat	$\Gamma(1.1111, 9.0)$
4	Sos activation, Km	$\Gamma(2.0, 3333.0)$
5	Binding of the EGF to the receptor, forward, mass action k	$\Gamma(1.1111, 9.0)$
6	Binding of the EGF to the receptor, backward, mass action k	$\Gamma(2.0, 3333.0)$
7	Sos deactivation, Km	$\Gamma(2.0, 3333.0)$
8	Sos deactivation, V	$\Gamma(2.0, 3333.0)$
9	Ras activation, Kcat	$\Gamma(1.1111, 9.0)$
10	Ras activation, Km	$\Gamma(2.0, 3333.0)$
11	Ras deactivation by Gap, Kcat	$\Gamma(1.1111, 9.0)$
12	Ras deactivation by Gap, Km	$\Gamma(2.0, 3333.0)$
13	cRaf activation, Kcat	$\Gamma(1.1111, 9.0)$
14	cRaf activation, Km	$\Gamma(2.0, 3333.0)$

15	cRaf deactivation, Km	$\Gamma(2.0, 3333.0)$
16	cRaf deactivation, V	$\Gamma(2.0, 3333.0)$
17	MEK activation, Kcat	$\Gamma(1.1111, 9.0)$
18	MEK activation, Km	$\Gamma(2.0, 3333.0)$
19	MEK deactivation, Km	$\Gamma(2.0, 3333.0)$
20	MEK deactivation, V	$\Gamma(2.0, 3333.0)$
21	ERK activation, Kcat	$\Gamma(1.1111, 9.0)$
22	ERK activation, Km	$\Gamma(2.0, 3333.0)$
23	cRaf inhibition by PKA, Kcat	$\Gamma(1.1111, 9.0)$
24	cRaf inhibition by PKA, Km	$\Gamma(2.0, 3333.0)$
25	PKA activation by PKA agonist, Kcat	$\Gamma(1.1111, 9.0)$
26	PKA activation by PKA agonist, Km	$\Gamma(2.0, 3333.0)$
27	PKA activation by Cilostamide, Kcat	$\Gamma(1.1111, 9.0)$
28	PKA activation by Cilostamide, Km	$\Gamma(2.0, 3333.0)$
29	PKA deactivation, Km	$\Gamma(2.0, 3333.0)$
30	PKA deactivation, V	$\Gamma(2.0, 3333.0)$
31	EPAC activation by EPAC agonist, Kcat	$\Gamma(1.1111, 9.0)$
32	EPAC activation by EPAC agonist, Km	$\Gamma(2.0, 3333.0)$
33	EPAC activation by Cilostamide, Kcat	$\Gamma(1.1111, 9.0)$
34	EPAC activation by Cilostamide, Km	$\Gamma(2.0, 3333.0)$
35	EPAC deactivation, Km	$\Gamma(2.0, 3333.0)$
36	EPAC deactivation, V	$\Gamma(2.0, 3333.0)$
37	Rap1 activation by EPAC, Kcat	$\Gamma(1.1111, 9.0)$
38	Rap1 activation by EPAC, Km	$\Gamma(2.0, 3333.0)$
39	Rap1 deactivation by Gap, Kcat	$\Gamma(1.1111, 9.0)$
40	Rap1 deactivation by Gap, Km	$\Gamma(2.0, 3333.0)$
41	BRaf activation by Rap1, Kcat	$\Gamma(1.1111, 9.0)$
42	BRaf activation by Rap1, Km	$\Gamma(2.0, 3333.0)$
43	BRaf deactivation, Km	$\Gamma(2.0, 3333.0)$
44	BRaf deactivation, V	$\Gamma(2.0, 3333.0)$
45	MEK activation by BRaf, Kcat	$\Gamma(1.1111, 9.0)$
46	MEK activation by BRaf, Km	$\Gamma(2.0, 3333.0)$

47	BRaf activation by Ras, Kcat	$\Gamma(1.1111, 9.0)$
48	BRaf activation by Ras, Km	$\Gamma(2.0, 3333.0)$
49	ERK deactivation, Km	$\Gamma(2.0, 3333.0)$
50	ERK deactivation, V	$\Gamma(2.0, 3333.0)$

The following table defines the parameters for Model 2 used in Case Study 2:

Table A.2: Parameters for Model 2 in Case Study 2

Parameter index	Description	Prior
1	Sos inhibition by ERKPP, Kcat	$\Gamma(1.1111, 9.0)$
2	Sos inhibition by ERKPP, Km	$\Gamma(2.0, 3333.0)$
3	Sos activation, Kcat	$\Gamma(1.1111, 9.0)$
4	Sos activation, Km	$\Gamma(2.0, 3333.0)$
5	Binding of the EGF to the receptor, forward, mass action k	$\Gamma(1.1111, 9.0)$
6	Binding of the EGF to the receptor, backward, mass action k	$\Gamma(2.0, 3333.0)$
7	Sos deactivation, Km	$\Gamma(2.0, 3333.0)$
8	Sos deactivation, V	$\Gamma(2.0, 3333.0)$
9	Ras activation, Kcat	$\Gamma(1.1111, 9.0)$
10	Ras activation, Km	$\Gamma(2.0, 3333.0)$
11	Ras deactivation by Gap, Kcat	$\Gamma(1.1111, 9.0)$
12	Ras deactivation by Gap, Km	$\Gamma(2.0, 3333.0)$
13	cRaf activation, Kcat	$\Gamma(1.1111, 9.0)$
14	cRaf activation, Km	$\Gamma(2.0, 3333.0)$
15	cRaf deactivation, Km	$\Gamma(2.0, 3333.0)$
16	cRaf deactivation, V	$\Gamma(2.0, 3333.0)$
17	MEK activation, Kcat	$\Gamma(1.1111, 9.0)$
18	MEK activation, Km	$\Gamma(2.0, 3333.0)$
19	MEK deactivation, Km	$\Gamma(2.0, 3333.0)$
20	MEK deactivation, V	$\Gamma(2.0, 3333.0)$

21	ERK activation, Kcat	$\Gamma(1.1111, 9.0)$
22	ERK activation, Km	$\Gamma(2.0, 3333.0)$
23	cRaf inhibition by PKA, Kcat	$\Gamma(1.1111, 9.0)$
24	cRaf inhibition by PKA, Km	$\Gamma(2.0, 3333.0)$
25	PKA activation by PKA agonist, Kcat	$\Gamma(1.1111, 9.0)$
26	PKA activation by PKA agonist, Km	$\Gamma(2.0, 3333.0)$
27	PKA activation by Cilostamide, Kcat	$\Gamma(1.1111, 9.0)$
28	PKA activation by Cilostamide, Km	$\Gamma(2.0, 3333.0)$
29	PKA deactivation, Km	$\Gamma(2.0, 3333.0)$
30	PKA deactivation, V	$\Gamma(2.0, 3333.0)$
31	EPAC activation by EPAC agonist, Kcat	$\Gamma(1.1111, 9.0)$
32	EPAC activation by EPAC agonist, Km	$\Gamma(2.0, 3333.0)$
33	EPAC activation by Cilostamide, Kcat	$\Gamma(1.1111, 9.0)$
34	EPAC activation by Cilostamide, Km	$\Gamma(2.0, 3333.0)$
35	EPAC deactivation, Km	$\Gamma(2.0, 3333.0)$
36	EPAC deactivation, V	$\Gamma(2.0, 3333.0)$
37	Rap1 activation by EPAC, Kcat	$\Gamma(1.1111, 9.0)$
38	Rap1 activation by EPAC, Km	$\Gamma(2.0, 3333.0)$
39	Rap1 deactivation by Gap, Kcat	$\Gamma(1.1111, 9.0)$
40	Rap1 deactivation by Gap, Km	$\Gamma(2.0, 3333.0)$
41	BRaf activation by Rap1, Kcat	$\Gamma(1.1111, 9.0)$
42	BRaf activation by Rap1, Km	$\Gamma(2.0, 3333.0)$
43	BRaf deactivation, Km	$\Gamma(2.0, 3333.0)$
44	BRaf deactivation, V	$\Gamma(2.0, 3333.0)$
45	MEK activation by BRaf, Kcat	$\Gamma(1.1111, 9.0)$
46	MEK activation by BRaf, Km	$\Gamma(2.0, 3333.0)$
47	C3G activation, Kcat	$\Gamma(1.1111, 9.0)$
48	C3G activation, Km	$\Gamma(2.0, 3333.0)$
49	C3G deactivation, mass action k	$\Gamma(2.0, 3333.0)$
50	Rap1 activation by C3G, Kcat	$\Gamma(1.1111, 9.0)$
51	Rap1 activation by C3G, Km	$\Gamma(2.0, 3333.0)$
52	BRaf activation by Ras, Kcat	$\Gamma(1.1111, 9.0)$

53	BRaf activation by Ras, Km	$\Gamma(2.0, 3333.0)$
54	ERK deactivation, Km	$\Gamma(2.0, 3333.0)$
55	ERK deactivation, V	$\Gamma(2.0, 3333.0)$

Detailed semantics of these parameters and the kinetic laws used for modelling biochemical reactions can be found in a formal model definition using SBML format (see Hucka et al. 2003) which can be obtained from the author.

Bibliography

Adalsteinsson, D., McMillen, D. and Elston, T. C.: 2004, Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks, *BMC Bioinformatics* **5**(24).

Akaike, H.: 1973, Information theory and an extension of the maximum likelihood principle, *in* B. N. Petrox and F. Caski (eds), *Second International Symposium on Information Theory*, Akademiai Kiado, Budapest, p. 267.

Akaike, H.: 1983, Information measures and model selection, *Bulletin of the International Statistical Institute* **50**, 277–290.

Alur, R., Belta, C., Kumar, V., Mintz, M., Pappas, G. J., Rubin, H. and Schug, J.: 2002, Modeling and analyzing biomolecular networks, *Computing in Science and Engineering* **4**(1), 20–31.

Alur, R. and Henzinger, T.: 1999, Reactive modules, *Formal methods in system design* **15**(1), 7–48.

Arkinson, A. C.: 1978, Posterior probabilities for choosing a regression model, *Biometrika* **65**, 39–48.

Athreya, K. B., Doss, H. and Sethuraman, J.: 1992, A proof of convergence of the Markov chain simulation method, *Technical Report 868*, Department of Statistics, Florida State University.

Aziz, A., Sanwal, K., Singhal, V. and Brayton, R.: 1996, Verifying continuous time markov chains, *Proc. CAV96*, Vol. 1102 of LNCS, Springer, pp. 269–276.

Bader, G. and Deuflhard, P.: 1983, A Semi-Implicit Mid-Point Rule for Stiff Systems of Ordinary Differential Equations, *Numerische Mathematik* **41**, 373–398.

Baillie, G. S. and Houslay, M. D.: 2005, Arrestin times for compartmentalised cAMP signalling and phosphodiesterase-4 enzymes, *Curr Opin Cell Biol.* **12**(2), 129–134.

Bartolucci, F. and Scaccia, L.: 2004, A new approach for estimating the Bayes factor, *Technical report*, Università di Perugia.

Belta, C., Finin, P., Habets, L. C. G. J. M., Halász, A. M., Imieliǹski, M., Kumar, V. and Rubin, H.: 2004, Understanding the bacterial stringent response using reachability analysis of hybrid systems, *in* R. Alur and G. J. Pappas (eds), *Seventh International Workshop on Hybrid Systems: Computation and Control, HSCC 2004*, Vol. 2993 of *Lecture Notes in Computer Science*, Springer, pp. 111–125.

Bernardo, J. M. and Smith, A. F. M.: 1994, *Bayesian Theory*, Wiley, Chichester.

Box, G. E. P., Leonard, T. and Wu, C.-F. (eds): 1983, *Scientific Inference, Data Analysis and Robustness*, New York: Academic Press.

Brown, K. S., Hill, C. C., Calero, G. A., Myers, C. R., Lee, K. H., Sethna, J. P. and Cerione, R. A.: 2004, The statistical mechanics of complex signaling networks: nerve growth factor signaling, *Physical Biology* **1**, 184–195.

Burbeck, S. and Jordan, K. E.: 2006, An assessment of the role of computing in systems biology, *IBM J. RES & DEV.* **90**(6), 529–543.

Calder, M., Gilmore, S. and Hillston, J.: 2006, Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA, *Transactions on Computational Systems Biology VII* **4230**, 1–23.

Calder, M., Vyshemirsky, V., Gilbert, D. and Orton, R.: 2006, Analysis of signalling pathways using continuous time markov chains, *Transactions on Computational Systems Biology VI* **4220**, 44–67.

Calzone, L., Chabrier-Rivier, N., Fages, F., Gentils, L. and Soliman, S.: 2005, Machine learning bio-molecular interactions from temporal logic properties, *Proceedings of the Third International Workshop on Computational Methods in Systems Biology CMSB'05, associated to ETAPS'05.*

Calzone, L., Chabrier-Rivier, N., Fages, F. and Soliman, S.: 2006, Machine learning biochemical networks from temporal logic properties, *Transactions on Computational Systems Biology VI. Springer-Verlag. LNBI* **4220**, 68–94.

Carlin, B. P. and Chib, S.: 1995, Bayesian model choice via Markov chain Monte Carlo, *Journal of the Royal Statistical Society, Series B* **57**, 473–484.

Chabrier, N. and Fages, F.: 2003, Symbolic model checking of biochemical networks, *International Workshop on Computational Methods in Systems Biology CMSB'03, Roverto, Italy*, Springer-Verlag LNCS.

Chabrier-Rivier, N., Chiaverini, M., Danos, V., Fages, F. and Schächter, V.: 2004, Modeling and querying biomolecular interaction networks, *Theoretical Computer Science* **325**(1), 25–44.

Chib, S.: 1995, Marginal likelihood from the Gibbs output, *J. Amer. Statist. Assoc.* **90**(432), 1313–1321.

Cho, K.-H., Shin, S.-Y., Kim, H.-W., Wolkenhauer, O., McFerran, B. and Kolch, W.: 2003, Mathematical modeling of the influence of RKIP on the ERK signaling pathway, *Lecture Notes in Computer Science* **2602**, 127–141.

Clarke, E. M., Grumberg, O. and Long, D.: 1994, Verification tools for finite state concurrent systems, *A Decade of Concurrency – Reflections and Perspectives, LNCS 803.*

Clarke, E. M., Grumberg, O. and Peled, D. A.: 1999, *Model Checking*, MIT Press.

Clavel, M., Durán, F., Eker, S., Lincoln, P., Martí-Oliet, N., Meseguer, J. and Talcott, C.: 2003, The maude 2.0 system, *in* R. Nieuwenhuis (ed.), *Rewriting Techniques and Applications (RTA 2003)*, number 2706 in *Lecture Notes in Computer Science*, Springer-Verlag, pp. 76–87.

Cox, D. R. and Hinkley, D. V.: 1974, *Theoretical Statistics*, London: Chapman and Hall.

Crosby, J. L.: 1973, *Computer Simulation in Genetics*, John Wiley & Sons, London.

de Jong, H.: 2003, Modeling and simulation of genetic regulatory systems: a literature review, *Lecture Notes in Computing Science* **2602**, 149–162.

Dhillon, A. S., Pollock, C., Steen, H., Shaw, P. E., Mischak, H. and Kolch, W.: 2002, The cAMP dependent kinase regulates the Raf-1 kinase mainly by phosphorylation of serine, *Mol. Cell. Biol.* **22**(10), 3237–3246.

Doi, A., Fujita, S., Matsuno, H., Nagasaki, M. and Miyano, S.: 2004, Constructing biological pathway models with hybrid functional Petri nets, *In Silico Biology* **4**(3), 271–291.

Eker, S., Knapp, M., Laderoute, K., Lincoln, P. and Talcott, C.: 2002, Pathway logic: Executable models of biological networks, *Fourth International Workshop on Rewriting Logic and Its Applications (WRLA'2002)* **71**.

Emerson, E. A.: 1990, *Handbook of Theoretical Computer Science*, Vol. B, MIT Press, chapter Temporal and modal logic, pp. 955–1072.

Feller, W.: 1968, *An Introduction to Probability Theory and Its Applications*, Wiley, New York.

Florens, J.-P., Mouchart, M. and Rolin, J.-M.: 1990, *Elements of Bayesian Statistics*, New York: Marcel Dekker.

Freedman, D. A.: 1983, A note on screening regression equations, *The American Statistician* **37**, 152–155.

Friel, N. and Pettitt, A. N.: 2006, Marginal likelihood estimation via power posteriors, *Technical report*, Department of Statistics, University of Glasgow.

Gamerman, D.: 2006, *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Chapman&Hall/CRC.

Gelman, A.: 1998, Simulating normalising constants: From importance sampling to bridge sampling to path sampling, *Stat. Sci.* **13**, 163–185.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B.: 1995, *Bayesian Data Analysis*, Chapman & Hall.

Geman, S. and Geman, D.: 1984, Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattern Anal. Machine Intell.* **6**, 721–741.

Gibbs, M. N. and MacKay, D. J. C.: 2000, Variational Gaussian process classifiers, *IEEE Trans. on Neural Networks* **11**(6), 1458–1464.

Gilks, W., Richardson, S. and Spiegelhalter, D.: 1995, *Markov Chain Monte Carlo in Practice*, Chapman&Hall/CRC.

Gillespie, C. S., Wilkinson, D. J., Shanley, D. P., Proctor, C. J., Boys, R. J. and Kirkwood, T. B. L.: 2006, BASIS: an internet resourse for network modelling, *J. of Integrative Bioinformatics* **3**(2).

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

Gilmore, S. and Hillston, J.: 1994, The PEPA Workbench: A tool to support a process algebra-based approach to performance modelling, *Proceedings of the Seventh International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, Vol. 794 in Lecture Notes in Computer Science, pp. 353–368.

Glover, F. and Laguna, M.: 1997, *Tabu Search*, Kluwer, Norwell, MA.

Goss, P. J. and Peccoud, J.: 1998, Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets, *Proceedings of National Academy of Sciences* **95**, 6750–6755.

Green, P. J.: 1995, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika* **82**, 711–732.

Green, P. J. and O'Hagan, A.: 1998, Model choice with MCMC on product spaces without using pseudo-priors, *Technical report 98-13*, University of Nottingham.

Greene, L. A. and Tischler, A. S.: 1976, Establishment of a noradrenergic clonal line of rat adrenal pheochromocytoma cells which respond to nerve growth factor, *Proc Natl Acad Sci U S A* **73**(7), 2424–2428.

Grenander, U.: 1983, Tutorial in pattern theory, *Technical report*, Brown University.

Han, C. and Carlin, B. P.: 2001, Markov chain Monte Carlo methods for computing Bayes factors: A comparative review., *Journal of the American Statistical Association* **96**(455), 1122–1132.

Hastings, W. K.: 1970, Monte Carlo sampling methods using Markov chains and thier applications, *Biometrika* **57**, 97–109.

Heath, J., Kwiatkowska, M., Norman, G., Parker, D. and Tymchyshyn, O.: 2006, Probabilistic model checking of complex biological pathways, *in* C. Priami (ed.), *Proc. Computational Methods in Systems Biology (CMSB'06)*, Vol. 4210 of *Lecture Notes in Bioinformatics*, Springer, pp. 32–47.

Heiner, M., Koch, I. and Will, J.: 2004, Model validation of biological pathways using Petri nets – demonstrated for apoptosis, *Biosystems* **75**, 15–28.

Hills, S. E.: 1987, Reference priors and identifiability problems in non-linear models, *The Statistician* **36**, 235–240.

Hillston, J.: 1996, *A Compositional Approach to Performance Modelling*, Cambridge University Press.

Hindmarsh, A. C., Brown, P. N., Grant, K. E., Lee, S. L., Serban, R., Shumaker, D. E. and Woodward, C. S.: 2005, SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers, *ACM Transactions on Mathematical Software* **31**(3), 363–396.

Houslay, M. D. and Kolch, W.: 2000, Cell type specific integration of crosstalk between ERK and cAMP signaling, *Molecular Pharmacology* **58**, 659–668.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Novere, N. L., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J. and Wang, J.: 2003, The systems biology markup language (SBML): A

medium for representation and exchange of biochemical network models, *Bioinformatics* **19**(4), 524–531.

Iba, Y.: 2001, Population Monte Carlo algorithms, *Transactions of the Japanese Society for Artificial Intelligence* **16**(2), 279–286.

Iosifescu, M.: 1980, *Finite Markov Processes and Their Applications*, Wiley, Chichester.

Jaynes, E. T.: 2003, *Probability Theory: The Logic Of Science*, Cambridge University Press.

Jeffreys, H.: 1937, *Scientific inference*, Cambridge University Press.

Jeffreys, H.: 1961, *Theory of Probability (3rd ed.)*, Oxford University Press.

Kao, S., Jaiswal, R. K., Kolch, W. and Landreth, G. E.: 2001, Identification of the Mechanisms Regulating the Diferential Activation of the MAPK Cascade by Epidermal Growth Factor and Nerve Growth Factor in PC12 Cells, *The Journal of Biological Chemistry* **276**, 18169–18177.

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

Khanin, R., Vinciotti, V., Mersinias, V., Smith, C. and Wit, E.: 2007, Statistical reconstruction of transcription factor activity using Michaelis-Menten kinetics, *Biometrics* **Accepted for publication**, 2007.

Kholodenko, B. N.: 2000, Negative feedback and ultrasensitivity can bring about oscillations in the mitogen-activated protein kinase cascades, *Eur. J. Biochem.* **267**, 1583–1588.

Kholodenko, B. N., Demin, O. V., Moehren, G. and Hoek, J. B.: 1999, Quantification of short term signaling by the epidermal growth factor receptor, *The Journal of Biological Chemistry* **274**(42), 30169–30181.

Kierzek, A. M.: 2002, STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm, *Bioinformatics* **18**, 470–481.

Kirkpatrick, S., Gelatt, C. D. and Vecchi, M. P.: 1983, Optimization by simulated annealing, *Science* **220**(4598), 671–680.

Kozen, D.: 1983, Results on the Propositional Mu Calculus, *Theoretical Computer Science* **27**(3), 333–354.

Kwiatkowska, M.: 2003, Model checking for probability and time: From theory to practice, *Proc. 18th Annual IEEE Symposium on Logic in Computer Science (LICS'03)*, IEEE Computer Society Press, pp. 351–360. Invited Paper.

Kwiatkowska, M., Norman, G. and Parker, D.: 2002, PRISM: Probabilistic Symbolic Model Checker, *Lecture Notes in Computer Science* **2324**, 200–204.

Kwiatkowska, M., Norman, G. and Parker, D.: 2005, Quantitative analysis with the probabilistic model checker PRISM, *Electronic Notes in Theoretical Computer Science* **153**(2), 5–31.

Kwiatkowska, M., Norman, G. and Parker, D.: 2006, PRISM user's guide, Available from `www.cs.bham.ac.uk/~dxp/prism`.

Lartillot, N. and Philippe, H.: 2006, Computing Bayes factors using thermodynamic integration, *Systematic Biology* **55**(2), 195–207.

Levental, S.: 1988, Uniform limit theorems for Harris recurrent Markov chains, *Probab. Th. Rel. Fields* **80**, 101–118.

Lewis, S. M. and Raftery, A. E.: 1997, Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator, *J. Amer. Statist. Assoc.* **92**(438), 648–655.

Li, Z. and Chan, C.: 2004, Inferring pathways and networks with a bayesian framework, *The FASEB Journal* **Published online**, express article 10.1096/fj.03–0475fje.

Liang, F. and Wong, W. H.: 2001, Real-parameter Evolutionary Monte Carlo with applications to Bayesian mixture models, *Journal of the American Statistical Association* **96**, 653–666.

Lincoln, P. and Tiwari, A.: 2004, Symbolic systems biology: Hybrid modeling and analysis of biological networks, *in* R. Alur and G. J. Pappas (eds), *Seventh International Workshop on Hybrid Systems: Computation and Con-*

trol, HSCC 2004, Vol. 2993 of *Lecture Notes in Computer Science*, Springer, pp. 660–672.

Lindley, D. V.: 1965, *Introduction to Probability and Statistics from a Bayesian Viewpoint: Inference Pt. 2*, Cambridge University Press.

MacKay, D. J. C.: 2003, *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

Marais, R., Light, Y., Paterson, H. F., Mason, C. S. and Marshall, C. J.: 1997, Differential Regulation of Raf-1, A-Raf, and B-Raf by Oncogenic Ras and Tyrosine Kinases, *Journal of Biological Chemistry* **272**, 4378–4384.

Marshall, C. J.: 1995, Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation, *Cell* **80**, 179–185.

McCulloch, R. E. and Rossi, P. E.: 1991, Bayes factors for nonlinear hypotheses and likelihood distributions, *Technical Report 101*, Statistics Research Center, University of Chicago, Graduate School of Business.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E.: 1953, Equations of state calculations by fast computing machines, *J. Chemical Physics* **21**, 1087–1091.

Milner, R.: 1999, *Communicating and Mobile Systems: The π-calculus*, Cambridge University Press.

Moler, C.: 2004, *Numerical Computing with MATLAB*, SIAM Society for Industrial & Applied Mathematics.

Neal, R.: 1993, Probabilistic inference using Markov Chain Monte Carlo methods, *Technical Report CRG-TR-93-1*, Dept. of Computer Science, University of Toronto.

Newton, M. A. and Raftery, A. E.: 1994, Approximate Bayesian inference by the weighted likelihood bootstrap, *JRSS Ser. B* **3**, 3–48.

Ogata, Y.: 1989, A Monte Carlo method for high dimensional integration, *Num. Math.* **55**, 137–157.

Pinney, J., Westhead, D. and McConkey, G.: 2003, Petri Net representations in systems biology, *Biochem. Soc. Trans.* **31**, 1513–1515.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P.: 2002, *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge University Press.

Priami, C.: 1995, Stochastic pi-calculus, *Computer journal* **38**, 578–589.

Priami, C., Regev, A., Silverman, W. and Shapiro, E.: 2001, Application of a stochastic name passing calculus to representation and simulation of molecular processes, *Information Processing Letters* **80**, 25–31.

Raftery, A. E.: 1986, Choosing models for cross-classifications, *American Sociological Review* **51**, 145–146.

Raftery, A. E., Madigan, D. M. and Hoeting, J. A.: 1993, Model selection and accounting for model uncertainty in linear regression models, *Technical Report 262*, University of Washington, Dept. of Statistics.

Ramsey, F. P.: 1931, *The Foundations of Mathematics: and other logical essays*, Harcourt, Brace, and Co.

Rasmussen, C. E. and Williams, C. K. I.: 2006, *Gaussian Processes for Machine Learning*, MIT Press.

Regev, A., Silverman, W. and Shapiro, E.: 2001, Representation and simulation of biochemical processes using the π-calculus process algebra, *Proceedings of the Pacific Symposium of Biocomputing 2001 (PSB2001)*, Vol. 6, pp. 459–470.

Robert, C. P. and Casella, G.: 2004, *Monte Carlo Statistical Methods (second edition)*, New York: Springer-Verlag.

Sachs, K., Gifford, D., Jaakkola, T., Sorger, P. and Lauffenburgerm, D. A.: 2002, Bayesian network approach to cell signaling pathway modeling, *Sci. STKE* **148**, 38.

Sachs, K., Perez, O., Pe'er, D., Lauffenburgerm, D. A. and Nolan, G. P.: 2005, Causal protein-signaling networks derived from multiparameter single-cell data, *Science* **308**, 523–529.

Sackmann, A., Heiner, M. and Koch, I.: 2006, Application of Petri net based analysis techniques to signal transduction pathways., *BMC Bioinformatics* **7:482**, 1–17.

Schoeberl, B., Eichler-Jonsson, C., Gilles, E. D. and Muller, G.: 2002, Computational modelling of the dynamics of the MAP kinase cascade activated by surface and internalised EGF receptors, *Nature Biotechnology* **20**, 370–375.

Smith, A. F. M. and Spiegelhalter, D. J.: 1980, Bayes factors and choice criteria for linear models, *JRSS Ser. B* **42**, 213–220.

Spall, J. C.: 2003, *Introduction to Stochastic Search and Optimization*, John Wiley & Sons.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A.: 2002, Bayesian measures of model complexity and fit (with discussion), *JRSS Ser. B* **64**(4), 583–639.

Stryer, L.: 1995, *Biochemistry*, W. H. Freeman and Company.

Talcott, C., Eker, S., Knapp, M., Lincoln, P. and Laderoute, K.: 2004, Pathway logic modeling of protein functional domains in signal transduction, *Proceedings of the Pacific Symposium on Biocomputing*, pp. 568–580.

Thomas, A., Spiegelhalter, D. J. and Gilks, W. R.: 1992, *Bayesian Statistics 4*, Clarendon Press, chapter BUGS: A program to perform Bayesian inference using Gibbs sampling, pp. 837–842.

Tierney, L.: 1994, Markov chains for exploring posterior distributions, *Annals of Statistics* **22**, 1701–1762.

Tierney, L. and Kadane, J. B.: 1986, Accurate approximations for posterior moments and marginal densities, *J. Amer. Statist. Assoc.* **81**, 82–86.

Voet, D. and Voet, J. G.: 1995, *Biochemistry*, John Wiley & Sons, Inc.

Voit, E. O.: 2000, *Computational Analysis of Biochemical Systems*, Cambridge University Press.

Wang, Z., Dillon, T. J., Pokala, V., Mishra, S., Labudda, K., Hunter, B. and Stork, P. J. S.: 2005, Rap1-mediated activation of extracellular signal-regulated kinases by cyclic AMP is dependent on the mode of Rap1 activation, *Mol. Cell. Biol.* **26**, 2130–2145.

Werhli, A. V., Grzegorcyk, M. and Husmeier, D.: 2006, Comparative evaluation of reverse engineering gene regulatry networks with relevance networks, graphical Gaussian models and Bayesian networks, *Bioinformatics* **22**(20), 2523–2631.

Wilkinson, D. J.: 2006, *Stochastic Modelling for Systems Biology*, Chapman & Hall/CRC.

Williams, E.: 1959, *Regression Analysis*, Wiley.

Woolf, P. J., Prudhomme, W., Daheron, L., Daley, G. Q. and Lauffenburger, D. A.: 2005, Bayesian analysis of signaling networks governing embryonic stem cell fate decisions, *Bioinformatics* **21**(6), 741–753.

Wu, J., Dent, P., Jelinek, T., Wolfman, A., Weber, M. J. and Sturgill, T. W.: 1993, Inhibition of the EGF-activated MAP kinase signaling pathway by adenosine 3',5'-monophosphate, *Science* **262**, 1065–1069.

York, R. D., Yao, H., Dillon, T., Ellig, C. L., Eckert, S. P., McCleskey, E. W. and Stork, P. J. S.: 1998, Rap1 mediates sustained MAP kinase activation induced by nerve growth factor, *Nature* **391**, 622–626.