UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 358

Numerical Methods for Stochastic Modeling of Genes and Proteins

PAUL SJÖBERG

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007

ISSN 1651-6214
ISBN 978-91-554-7009-8
urn:nbn:se:uu:diva-8293

Dissertation presented at Uppsala University to be publicly examined in room 2446, Building 2, Pollacksbacken, Lägerhyddsvägen 2, Uppsala, Friday, November 30, 2007 at 13:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Sjöberg, P. 2007. Numerical Methods for Stochastic Modeling of Genes and Proteins. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 358. 42 pp. Uppsala. ISBN 978-91-554-7009-8.

Stochastic models of biochemical reaction networks are used for understanding the properties of molecular regulatory circuits in living cells. The state of the cell is defined by the number of copies of each molecular species in the model. The chemical master equation (CME) governs the time evolution of the the probability density function of the often high-dimensional state space. The CME is approximated by a partial differential equation (PDE), the Fokker-Planck equation and solved numerically. Direct solution of the CME rapidly becomes computationally expensive for increasingly complex biological models, since the state space grows exponentially with the number of dimensions. Adaptive numerical methods can applied in time and space in the PDE framework, and error estimates of the approximate solutions are derived. A method for splitting the CME operator in order to apply the PDE approximation in a subspace of the state space is also developed. The performance is compared to the most widely spread alternative computational method.

Keywords: master equation, Fokker-Planck equation, stochastic models, biochemical reaction networks

Paul Sjöberg, Department of Information Technology, Box 337, Uppsala University, SE-75105 Uppsala, Sweden

© Paul Sjöberg 2007

ISSN 1651-6214
ISBN 978-91-554-7009-8
urn:nbn:se:uu:diva-8293 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-8293)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I Elf, J., Lötstedt, P., Sjöberg, P. (2003) Problems of High Dimension in Molecular Biology. *Proceedings of the* 19^{th} *GAMM-Seminar*, Leipzig, 21–30.

II Ferm, L., Lötstedt, P., Sjöberg, P. (2006) Conservative Solution of the Fokker-Planck Equation for Stochastic Chemical Reactions *BIT Numerical Methematics*, 46:S61–S83.

III Sjöberg, P., Lötstedt, P., Elf, J. (2008) Fokker-Planck Approximation of the Chemical Master Equation in Molecular Biology, *Comput. Visual. Sci.*, 11 *published electronically in February 2007*, DOI:10.1007/s00791-006-0045-6.

IV Sjöberg, P. (2007) Partial Approximation of the Master Equation by the Fokker-Planck Equation, In Kågström, B., Elmroth, E., Dongarra, J., and Waśniewski, J., editors, *Applied Parallel Computing: State of the Art in Scientific Computing*, *Lecture Notes in Comput. Sci.* 4999:631–644, Springer-Verlag. [1]

V Sjöberg, P. (2007) PDE and Monte Carlo Approches to Solvning of the Master Equation Applied to Gene Regulation *Technical Report 2007-028*, Department of Information Technology, Uppsala University.

Reprints were made with permission from the publishers.

[1] With kind permission of Springer Science and Business Media.

Contents

1. Introduction

The foundation of all life on Earth is the cell. It contains an intricate machinery for reproducing itself and for improving the odds of doing so. Success is measured by survival and survival is a competition for resources. It is essential to make efficient use of available resources and adapt to changes in the environment to make use of any advantage over the competition. A complex system of chemical reactions have evolved for coping with this task. The complexity is a great challenge for the scientists pursuing the search for understanding the fundamental principles of life, but the impact of success to do so is hard to overestimate.

The cellular control system has a huge number of components and it is necessary to identify subsystems which can be examined. Such a subsystem is sometimes referred to as a circuit. The circumstances inside the cell put certain demands on the control system. The numbers of molecules are often small, for instance, which make random events influential and it is therefore important that the circuit functions well in a noisy system. A mathematical description that does not capture that aspect is not accurate enough for examining the circuit [2] and a model of the circuit that does not function under these circumstances is not credible [1] [5].

This thesis treats computational methods for simulating biochemical reaction networks in a stochastic sense and thereby introduces application of methods in numerical analysis to a new field. Computational methods are an irreplaceable tool for examining the consequences of a hypothetical model for comparison to experimental data or the effects of manipulation of a control circuit. The mathematical foundation has been known for some time see e.g. [18], but in practice it is often hard to solve the equations. With the exception of a few simple cases it is necessary to use numerical methods to solve the problem. Here, we use approximations to evade some of the problems associated with the Monte Carlo method that is usually applied. The problem is reformulated as a *partial differential equation* (PDE) and questions we would like to answer are:

1. When is the PDE-approximation the better alternative?
2. How is the approximation solved numerically?

Since molecular biology is not a usual application for numerical analysis and numerical analysis is not that common in molecular biology, we begin with a brief introduction to the basics in both fields in order to be able to describe the contributions of the thesis. First a summary of stochastic models

in molecular biology will be made followed by the numerical perspective. Finally the contributions will be described. Also included is a very popular summary in Swedish.

2. Molecular biology

The core of the modern scientific description of life concerns the famous macromolecules DNA, RNA and the proteins. Molecular biology characterizes their structure, chemical and physical properties as well as their interactions with each other. These characteristics are all clues for which parts that are connected and what functions they perform. Ultimately, all aspects of cellular life are a consequence of those functions or corresponding malfunctions. Famous examples are often concerned with malfunctions, such as cancer or the consequences of viral infections. The prerequisites for making such discoveries are of course a comprehension of the principles of a healthy system. When the step is taken from general principles to a detailed quantitative level a revolution is expected in medicine as well as in the biotechnological industry.

2.1 Control of cellular processes

The operational schemes of a cell are not understood just because its genome has been sequenced. That merely identifies the components of the system. For the complete picture the functional connections between the parts must be mapped. The understanding of genetic control circuits is growing quickly, but much remains before the whole picture is elucidated. The field *systems biology* is concerned with quantitative description of complex biological systems [21] and it such ambitions that motivate the methods described in this thesis.

2.1.1 Chemistry

Consider N molecular species named $X_i, i = 1 \ldots N$. The number of X_i-molecules is called *copy number* and is denoted by x_i. The chemical system under study has the volume Ω and is comprised of a solvent (water in biology) and a number of molecules of the different molecular species. The solution is assumed to be *well stirred*, so that the probability to find a certain molecule in a subvolume of the cell depends only on the size of the subvolume in relation to Ω. The state of the chemical system is described by the vector $\mathbf{x} = (x_1, \ldots, x_N)^T$. To follow the chain of events it is necessary to specify the *reactions* that change the composition of the system.

The molecules that are a prerequisite for the reaction are called *reactants* and the ones that are the result are called *products*. The number of X_i-molecules in the reactants and the products in a reaction will be denoted

n_i^- respectively n_i^+, where $n_i^-, n_i^+ \in \mathbb{Z}_+ = \{0,1,2,\ldots\}$. Every reaction is defined by its reactants, its products and a *reaction rate* v which is defined as the number of reactions per time unit. Since it is required that two molecules collide in order for a reaction to take place, the reaction rate depends on the copy numbers of the reactant species and the volume Ω, but we will consider Ω to be constant. The reaction is written as an arrow from reactants to products and the reaction rate above:

$$n_1^- X_1 + n_2^- X_2 + \ldots + n_N^- X_N \xrightarrow{v(\mathbf{x})} n_1^+ X_1 + n_2^+ X_2 + \ldots + n_N^+ X_N, \tag{2.1}$$

or

$$\mathbf{n}^- \mathbf{X} \xrightarrow{v(\mathbf{x})} \mathbf{n}^+ \mathbf{X}, \tag{2.2}$$

where $\mathbf{n}^- = (n_1^-, \ldots, n_N^-)^T$ and $\mathbf{n}^+ = (n_1^+, \ldots, n_N^+)^T$. There is always a reverse reaction which has the products as reactants and vice versa. Normally these two reactions are written together as

$$\mathbf{n}^- \mathbf{X} \underset{v^{-1}(\mathbf{x}-\mathbf{n}^-+\mathbf{n}^+)}{\overset{v(\mathbf{x})}{\rightleftharpoons}} \mathbf{n}^+ \mathbf{X},$$

where v^{-1} is the reaction rate of the reverse reaction. Since it is common to neglect the reverse reaction in a chemical system far from equilibrium the reactions will be treated as separate reactions.

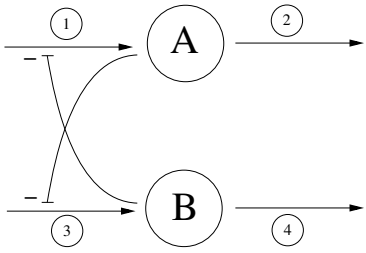
Many reactions are in reality not a single event, but a chain of events, and can be split into several steps. An *elementary reaction* defines the underlying molecular events of a one-step reaction [15]. It is unlikely for three or more molecules to collide simultaneously. Hence, for an elementary reaction $|\mathbf{n}^-| = \sum_i n_i^- \leq 2$. Reaction rates must in general be determined experimentally, but for an elementary reaction it has a simple form:

$$v_r(\mathbf{x}) = k \prod_{i=1}^{N} x_i^{n_i^-}, \tag{2.3}$$

where n_i^- is the i:th component of $\mathbf{n}^-$ and k is a reaction rate constant. This is a good model if the following conditions are met [23]:

1. The cell volume must be well stirred.
2. Reactive collisions must be separated by a large number of non-reactive collisions in order to ensure that the molecule velocities adhere to the Maxwell velocity distribution, otherwise the frequency of reactive collisions will not be proportional to the concentration.
3. The internal states of the molecules are in thermal equilibrium. If not, the reaction rates will be dependent on a varying distribution of excited states.
4. The temperature is constant. Reaction rates are very temperature dependent.

Example 1 *A toggle switch is a control circuit that turns gene expression on or off in a discrete rather than gradual manner. The mode of the switch should naturally be affected by some external signal, but in this model the focus is on the bistable property. The model consists of two molecular species A and B. Assume that a high copy number of A-molecules and few B-molecules corresponds to that the switch is on and the opposite to that it is off. The blunt arrows and the minus sign in the figure represent the metabolites reciprocal inhibition of the synthesis of each other.*



The reactions are

1	$\emptyset \xrightarrow{v_1(b)} A$	$v_1(b) = k/(K+b^2)$	*synthesis*
2	$A \xrightarrow{v_2(a)} \emptyset$	$v_2(a) = \mu a$	*degradation/dilution*
3	$\emptyset \xrightarrow{v_3(a)} B$	$v_3(a) = k/(K+a^2)$	*synthesis*
4	$B \xrightarrow{v_4(b)} \emptyset$	$v_4(b) = \mu b$	*degradation/dilution*

If there are no reactants $\emptyset$ *is used to denote a source. The same symbol is also used when no products are created, but is then called a sink. Notice that the reactions* 1 *and* 3 *are not elementary reactions and do not obey (2.3).*

If the copy number is represented by a continuous function, the time evolution for the copy number of each species can be described by an *ordinary differential equation* (ODE):

$$\frac{d\mathbf{x}}{dt} = \sum_{r=1}^{R} v_r(\mathbf{x})(\mathbf{n}_r^+ - \mathbf{n}_r^-), \tag{2.4}$$

where $\mathbf{n}_r^-$ is the numbers of the reactants and $\mathbf{n}_r^+$ the numbers of the products of reaction number r. The dynamics of the chemical system is then described by the system of coupled ODEs from each species, which are called the *reaction rate equations*. The reaction rate equations are often *stiff* since biochemical systems regularly have many different time scales. This puts certain demands on numerical ODE-solvers, but the problem is well explored, see for example [11] and [12] for numerical methods.

Example 2 *The reaction rate equations for Example 1 are:*

$$\begin{aligned} \tfrac{da}{dt} &= k/(K+b^2) - \mu a \\ \tfrac{db}{dt} &= k/(K+a^2) - \mu b. \end{aligned}$$

An ODE-solver will find one of the two modes of the switch depending on the initial conditions. The switch property will show itself if the system is perturbed far enough from the locally stable attractor. The perturbation can be thought of as an external signal.

The description of a chemical system here differs a little from the one normally used in chemistry. The ulterior motive is to simplify the notation in the current context. These are the discrepancies:

1. The concept *concentration* is avoided in benefit of *copy number*. In biochemical systems, the discrete character of the molecules is tangible since the copy numbers are very low.
2. Reaction rates are written out explicitly in the reaction. In chemical notation reaction rate constants or equilibrium constants are used. Reaction rates are implied by the constant and reactant concentrations, since reactions are assumed to be elementary.
3. Every reaction has a reverse reaction. As previously mentioned these reactions will be regarded as two separate reactions and often one of them is neglected.
4. To make a tractable model of cellular control circuits it is important to limit the number of components. Reactions will seem to violate principles of mass conservation. Molecules will emerge from nowhere and disappear in the void. The models are all open systems where mass can be exchanged with the environment. The $\emptyset$-symbol (source or sink) is used for such exchange when there are no other reactants or products, see, Example 1.

2.1.2 Stochastic models in molecular biology

Biochemical processes in the cell occur at a very small scale. The volume of a bacterial cell is in the order of $10^{-15}l$. Furthermore, many of the important molecules are present in very low copy numbers, DNA in one or a few copies, RNA in tenths and proteins perhaps in hundreds. Simultaneously there are a large number of molecular species that selectively interact and whose dynamics are dependent to a high degree.

Because of the discrete nature of molecules, there is always a certain noise in chemical systems that is not the result of *external noise*, such as for instance variation in temperature, but comes from the randomness of collisions in the reactor volume. This inherent property is called *internal noise* or *intrinsic noise* [23].

The system can be described at different levels of details. A description where the position, velocities, rotation and inner state of the molecules are accounted for and physical principles act on them would be a *microscopic* description. This approach would be very computationally demanding. Usually a *macroscopic* description is used instead. In the macroscopic description, the dynamics of the mean copy numbers are determined by a system of ordi-

nary differential equations (2.4). If the system tends towards equilibrium and the copy numbers are large, the noise is insignificant and the system is well modelled by this description. Both these conditions are frequently violated in biological systems. A rule of thumb is that the fluctuation in copy number for a species with an average of M molecules is of order $M^{1/2}$ [23]. There are examples, however, where this is not a sufficient level of detail. In such cases a *mesoscopic* description might be necessary. At the mesocopic level, the copy numbers are described in a probabilistic sense which accurately captures the fluctuations in copy numbers. The macroscopic description can be viewed as an average of the mesocsopic variables.

The fluctuations are important in the following cases:

1. If copy numbers are small then the relative variation is large and highly relevant for determining the macroscopic properties of the system [18].
2. Near-critical systems can exhibit very large fluctuations for comparatively large copy numbers [4].
3. The dynamics of systems with several stable fixed points or almost stable fixed points cannot be studied without characterizing the fluctuations [24].

Let $p(\mathbf{x},t)$ be the probability for the system to be in state $\mathbf{x}$ at time t and $p(\mathbf{x}_1,t_1;\mathbf{x}_2,t_2;\ldots)$ be the joint probability for the state $\mathbf{x}_1$ at time t_1 and the state $\mathbf{x}_2$ at time t_2 and so on. It is practical to define the conditional probabilities

$$p(\mathbf{x}_1,t_1;\mathbf{x}_2,t_2;\ldots|\mathbf{y}_1,\tau_1;\mathbf{y}_2,\tau_2,\ldots)=\frac{p(\mathbf{x}_1,t_1;\mathbf{x}_2,t_2;\ldots;\mathbf{y}_1,\tau_1;\mathbf{y}_2,\tau_2,\ldots)}{p(\mathbf{y}_1,\tau_1;\mathbf{y}_2,\tau_2,\ldots)},$$

where $\mathbf{x}_1$, $\mathbf{x}_2,\ldots$ and $\mathbf{y}_1$, $\mathbf{y}_2,\ldots$ are the states of the system at the times t_1, t_2, ... and τ_1, τ_2, ... respectively. The conditional probability is simply the probability for the states $\mathbf{x}_i$, $i=1,2,\ldots$ if the states $\mathbf{y}_j$, $j=1,2,\ldots$ are known to have occurred. If the times are ordered $t_1\geq t_2\geq\ldots\geq\tau_1\geq\tau_2\geq\ldots$ the states $\mathbf{y}_j$ are the memory of the system.

A *stochastic process* defines the time evolution of an individual in the population that is described by the probability distribution. It is defined by state transitions in contrast to reactions which deal with the molecular circumstances for such a transition. A general state transition that corresponds to the reaction (2.1) is written

$$\mathbf{x}\xrightarrow{w(\mathbf{x})}\mathbf{x}+\mathbf{n}^+-\mathbf{n}^-.$$

where w denotes the transition rate which is closely related to the reaction rate v. The stochastic transition rate is often referred to as *propensity* and is defined by

$$w(\mathbf{x}'|\mathbf{x})=\lim_{\Delta t\to 0}\frac{p(\mathbf{x}',t+\Delta t|\mathbf{x},t)}{\Delta t}. \tag{2.5}$$

Assume that the limit exists for every reaction and all $\mathbf{x},\mathbf{x}'\in\mathbf{Z}_+^N$. $w\Delta t$ is simply the probability for a reaction event in the infinitesimal time interval Δt. The propensity is essentially the same thing as the reaction rate, but the former is

defined in a stochastic context. w can replace v in (2.3). For a reaction, $\mathbf{n}^+$ and $\mathbf{n}^-$ determines the end state $\mathbf{x}'$ for a given state $\mathbf{x}$ and the propensity is written $w(\mathbf{x})$.

For systems with reactions with time-independent propensities, the previous states of the system are clearly unimportant. The propensities can be computed from the current state of the system. Such a process, which has no memory is called a *Markov process*. As (2.3) shows, there is no time dependence in the propensity of elementary reactions. That means that systems of approximately elementary reactions are well described as Markov processes.

Example 3 *The reactions in Example 1 written as state transitions:*

1	$(a,b) \xrightarrow{w_1(b)} (a+1,b)$	$w_1(b) = k/(K+b^2)$	*synthesis*
2	$(a,b) \xrightarrow{w_2(a)} (a-1,b)$	$w_2(a) = \mu a$	*degradation/dilution*
3	$(a,b) \xrightarrow{w_3(a)} (a,b+1)$	$w_3(a) = k/(K+a^2)$	*synthesis*
4	$(a,b) \xrightarrow{w_4(b)} (a,b-1)$	$w_4(b) = \mu b$	*degradation/dilution*

For low copy numbers it is reasonable to correct rates for elementary reactions for actual number of possible reaction pairs. The difference is that the propensity of a reaction like

$$2A \xrightarrow{k\cdot a\cdot(a-1)} B$$

is proportional to $a\cdot(a-1)$ not just a^2 since an A-molecule must collide with one of the other $a-1$ A-molecules. The importance of this correction vanishes quickly with increasing copy numbers, but is very important for very small copy numbers, e.g. if $a = 1$ in the example above. The modified propensity for an elementary reactions is

$$w(\mathbf{x}) = k\prod_{i=1}^{N}\frac{x_i!}{(x_i-n_i^-)!}.$$

2.2 The chemical master equation

The *master equation* [6] [23] for a general Markov process describes the time evolution of the probability distribution of the state of the system. The underlying process is defined by an initial distribution $p(\mathbf{x},0)$ and a transition rate per unit time, here called the propensity $w(\mathbf{x}'|\mathbf{x})$ defined in (2.5). The master equation is a difference-differential equation and is written

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = \sum_{\mathbf{x}'\in\mathbb{Z}_+^N}\left(w(\mathbf{x}|\mathbf{x}')p(\mathbf{x}',t) - w(\mathbf{x}'|\mathbf{x})p(\mathbf{x},t)\right). \tag{2.6}$$

The equation is defined in every point in the *state space*, that is $\mathbf{x}\in\mathbb{Z}_+^N$ for an N-dimensional problem. The total probability in the state space is conserved since copy numbers cannot become negative.

The difference in molecule numbers due to reaction r is denoted $\mathbf{n}_r = \mathbf{x} - \mathbf{x}'$ and is used to rewrite the master equation (2.6) for a system with R reactions

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = \sum_{\substack{r=1 \\ (\mathbf{x}+\mathbf{n}_r)\in\mathbb{Z}_+^N}}^{R} q_r(\mathbf{x}+\mathbf{n}_r,t) - \sum_{\substack{r=1 \\ (\mathbf{x}-\mathbf{n}_r)\in\mathbb{Z}_+^N}}^{R} q_r(\mathbf{x},t), \tag{2.7}$$

where $q_r(\mathbf{x},t) = w_r(\mathbf{x})p(\mathbf{x},t)$ and w_r is the propensity for the r:th reaction. We will call this form the *chemical master equation* (CME).

2.3 The Fokker-Planck equation

If the CME is approximated with a continuous approximation, it is possible to use a substantially coarser computational grid than if the CME is solved numerically as a system of ODEs on $\mathbb{Z}_+^N$. Define a CME on a continuous state space $\mathbb{R}_+^N = \{\mathbf{x} \in \mathbb{R}^N | x_i \geq 0\}$, so that the continuous equation coincides with the discrete CME in the points in the discrete state space. By Taylor-expansion at $\mathbf{x}$ and truncation at the third order term, the *Fokker-Planck equation* (*FPE*) is obtained:

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = \sum_{r=1}^{R} \left\{ \sum_{i=1}^{N} n_{ri} \frac{\partial \left(q_r(\mathbf{x},t)\right)}{\partial x_i} + \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{n_{ri}n_{rj}}{2} \frac{\partial^2 \left(q_r(\mathbf{x},t)\right)}{\partial x_i \partial x_j} \right\}, \tag{2.8}$$

where n_{ri} is the i:th element of $\mathbf{n}_r$. In order to find a conservation form of the FPE, the probability current [6] is defined

$$F_{ri} = n_{ri}(q_r + \frac{1}{2}\mathbf{n}_r \cdot \nabla q_r)$$

and $\mathbf{F}_r = (F_{r1}, \ldots, F_{rN})^T$. The FPE on conservation form is now written

$$\frac{\partial p(\mathbf{x},t)}{\partial t} = \sum_{r=1}^{R} \nabla \cdot \mathbf{F}_r. \tag{2.9}$$

The boundary condition is that the probability current over the boundary is zero

$$F_{ri} = 0 \quad \text{on} \quad \Gamma_i = \{x | x_i = 0\},$$

since the total probability of the state space is conserved. The FPE is a parabolic PDE.

3. Numerical methods

The usefulness of numerical methods depends on two properties. How accurate they are and how large problems that can be solved. These are both aspects of the efficiency of the method. To evaluate the efficiency it necessary to measure the error in the numerical solution and relate it to the time for making the computation. Methods also differ in the memory demands which also may limit the size of problems that can be solved.

3.1 Numerical solution of PDEs

In order to solve a continuous problem, such as a PDE, it needs to be discretized, that is a corresponding discrete problem must be derived. The discrete equation must be constructed so that its solutions approximates the solutions of the original equation. It is common to first discretize the PDE in the spatial domain to obtain a system of ODEs that can be subsequently discretized in time by some method for solving ODEs. This approach is called *the method of lines*.

There are several established discretization methods to choose from. In this thesis *finite difference methods* and *finite volume methods* are used for the discretization of the state space, but there are other alternatives such as *finite element methods*.

3.1.1 Finite difference methods

Finite difference methods require a *structured grid*, that is a computational grid where every grid point is identifiable by an ordinal number in each dimension. If the distance between points is constant in each dimension, the grid is a *uniform grid*. If the distance between points vary it is a non-uniform grid or a *stretched grid*.

A finite difference method approximates the derivatives of the equation in the grid points using the solution in adjacent points. A simple example is the approximation the derivative of the function $f(x)$ in the points x_i, $i = 1, \ldots, N$,

$$\frac{df(x_i)}{dx} \approx \frac{f(x_{i+1}) - f(x_{i-1})}{2h},$$

where h is the distance between points. Near the boundaries of the grid this approximation cannot be computed since one of the points in the approximation

lies outside the grid. There, a boundary condition must be applied. Mathematical boundary conditions of the original equation can be used, but often there are not enough such conditions and numerical boundary conditions must be devised.

3.1.2 Finite volume methods

A finite volume method is based on a *conservation form* of the equation to be solved. The computational domain is divided into a number of subvolumes which are called cells. The equation must be written on the form

$$\frac{\partial u}{\partial t} = \nabla \cdot \mathbf{F}(u).$$

Integrate over the cell ω with boundary $\partial\omega$ and apply Gauss' formula to form the following equation for the mean value of u in ω

$$|\omega| \frac{\partial}{\partial t} \bar{u} = \frac{\partial}{\partial t} \int_{\omega} u \, d\omega = \int_{\partial\omega} \mathbf{F}(u) \cdot \hat{\mathbf{n}}_{\omega} ds,$$

where $\hat{\mathbf{n}}_{\omega}$ is the normal of the boundary $\partial\omega$ and $|\omega|$ is the volume of ω. The subvolumes in the discretization do not have to have any particular structure, but on a structured grid a standard second order discretization can be interpreted as a second order finite difference method with grid points in the center of the cells. All finite volume discretizations in this thesis can be interpreted in this way.

3.1.3 Properties of the discretization

To be able to talk about accuracy, the numerical solution has to converge. That is, approach the analytical solution as some discretization parameter L goes to infinity, where an increasing L implies an increase in the computational work. For structured grids, L is the product of the number of points in each dimension. In practice L is replaced by the distance between points as the measure of the resolution. For an equidistant grid in N dimensions and $h_i = h$, $L \propto 1/h^N$.

To describe the qualities of a discretiztion the concepts *stability* and *order of accuracy* is used. Stability means that the discrete solution does not grow without control due to amplification of numerical errors. An unstable method is useless because the solution is drowned in an explosion of numerical noise. Many methods must satisfy certain conditions on the resolution in order to be stable. The order of accuracy is the convergence speed of the approximate solution. Let h denote the distance between grid points on a uniform grid and k the time step. If the solution error τ is bounded, so that $\tau \leq C(h^p + k^q)$, where C is a constant depending on the solution, then (p,q) is the order of accuracy. Let us assume that $p = q$ for simplicity. A first order method implies that a

twofold increase in resolution in each dimension will cause a twofold decrease in the error. The same increase in resolution for a second order method would decrease the error fourfold. Assuming the error due to the space and time discretizations are of the same order of magnitude, it is necessary to refine the grid in both time and space to decrease the error. In the same way the resolution for every spatial dimension in a high-dimensional problem, must be increased to achieve error reduction. A large increase in the total number of points in the grid is necessary since it is a product of the number of points in each dimension. A twofold reduction in the error requires an increase of 2^N points for a first order method.

Many biochemical systems have reactions with reactions rates that differ by orders of magnitude. That implies that the dynamics of the system has several time scales. Some species vary rapidly due to fast reactions, and may reach a balance point, while slow reactions change the conditions in the systems which will probably move the balance point in the fast scale. As mentioned before, such systems are called stiff. Stiff systems put special demands on the time discretization. Time discretizations are divided into *explicit methods* and *implicit methods*. The difference between them is that explicit schemes compute the solution at the time t_{k+1} using only the solution at previously computed time steps, $t_0, \ldots, t_k$ while implicit methods use the yet not computed solution at t_{k+1} to generate a system of equations. Implicit schemes make it necessary to solve a system of equations in each time step. That is expensive, but usually make it possible to take longer time steps. Explicit methods have a much stricter stability condition. Implicit methods are especially suitable for stiff problems. The large sparse systems of equations is solved efficiently in each time step by *iterative methods* [10].

Apart from stability and order of accuracy a numerical method may be constructed to preserve some other property of the original problem, such as conservation of mass in a closed system or that probabilities are non-negative. Finite volume methods are constructed from a conservation law and ensure that the discrete equation is conservative as well.

3.2 The stochastic simulation algorithm

The method that dominates the study of stochastic models in molecular biology is a Monte Carlo-method called the *stochastic simulation algorithm* (*SSA*) [8]. The method generates realizations of the stochastic process that models the reaction network. A realization corresponds to a chain of events for an individual in the population that is described by the stochastic process. It is a *trajectory* through the state space representing the evolution of the individual system. The realizations can be used for determining different statistical properties of the process. To approximate the solution of the master equation, the probability for a state $\mathbf{x}$ at the time t is estimated by making a large number

of independent simulations and computing the number of trajectories passing the point $(\mathbf{x}, t)$ divided by the total number of trajectories.

4. Summary of papers

My contribution to these papers are:

Paper I I implemented the method, made the tests and wrote parts of the report.

Paper II I implemented the SSA code, conducted the SSA experiments and wrote small parts of the report.

Paper III I implemented and conducted all experiments and wrote parts of the report.

Paper IV I am the sole author.

Paper V I am the sole author.

4.1 Paper I

Paper I introduces CME in numerical analysis as a high-dimensional problem and presents the FPE-approximation as an alternative for solving it using finite differences for a four-dimensional example. The FPE approximation is more flexible than CME for computational purposes since the spatial resolution can be varied.

4.1.1 Numerical solution of the FPE-approximation

The state space is truncated at a sufficiently high copy number and the truncated computational domain is discretized using second order accurate centered finite differences. *Dirichlet boundary conditions* are used, that is the probability is assumed to be zero at the boundary. The discretized equation is written

$$\frac{d\mathbf{p}(t)}{dt} = A\mathbf{p}(t),$$

where A is the discrete space operator and $\mathbf{p}(t)$ is the numerical solution. For time discretization the second order *backward difference formula* (BDF-2) was used [11]. If the solution in time step n is represented by the vector $\mathbf{p}^n$ the scheme can be written

$$\left(\frac{3}{2}I - \Delta t A\right)\mathbf{p}^n = 2\mathbf{p}^{n-1} - \frac{1}{2}\mathbf{p}^{n-2}, \tag{4.1}$$

where I is the identity matrix and Δt is the time step.

BiCGSTAB [22] is used to solve the system of equations (4.1) generated by the implicit method. BiCGSTAB is an iterative method for solving large sparse non-symmetric systems of equations. It only has to store seven vectors of size N besides the sparse matrix $\mathbf{A}$ during the iterations. The modest memory demands are important since it is the memory that limits how large problems that can be solved. In order to accelerate the convergence of the iterative method, incomplete LU-factorization (ILU) with zero fill-in is used as a preconditioner [10]. ILU with a threshold value might use more memory and is not an alternative. In addition it is more efficient to compute the factorization knowing the structure of non-zero elements in the incomplete factors.

4.1.2 A model of coupled flows

The method is tested on a model from [4] that has been extended by two enzymes. An overview of the model is shown in Figure 4.1 In the model, two

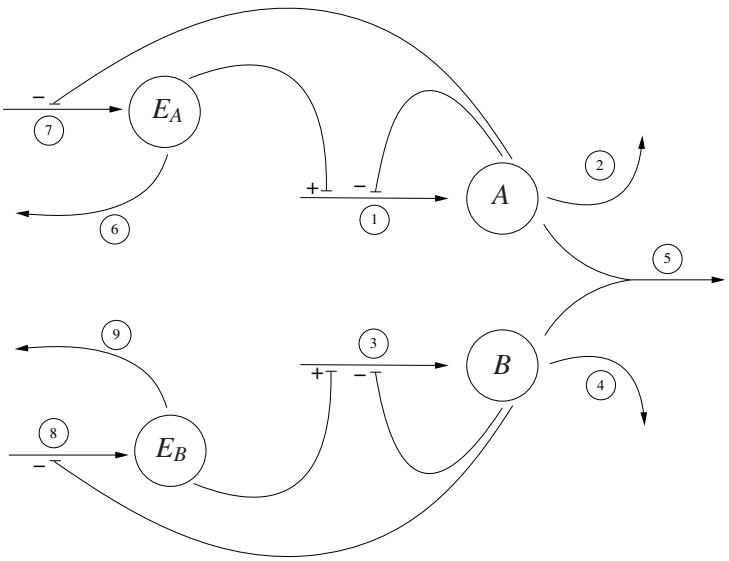


Figure 4.1: The model of a coupled flows. 1. synthesis of A. 2. degradation of A. 3. synthesis of B. 4. degradation of B. 5. high consumption of metabolites. 6. degradation of enzyme E_A. 7. synthesis of E_A. 8. synthesis of the enzyme E_B. 9 degradation of E_B.

enzymes E_A and E_B are regulated by their respective products, the metabolites A and B. Enzymes catalyze the reaction which means that they take part in the reaction and lower the activation energy threshold and thereby speed up the reaction, but they are not consumed in the reaction. That is the definition of *catalysis*. In the figure it is indicated by blunt arrows with a plus sign at the catalyzed reaction. Correspondingly a blunt arrow with a minus sign means that the reaction is slowed down, which is called *inhibition*. When the product is

abundant the synthesis rate of the corresponding enzyme goes down. Furthermore the activity of the enzyme is regulated by the amount of product. There are two time scales in the control system, the slow regulation of enzyme levels and the fast feedback on metabolite synthesis rate. The metabolites are consumed together by a reaction whose products are removed at a high enough rate for the reaction to be regarded as irreversible. In [4] the coupled consumption of A and B is catalysed by an enzyme, which here is assumed to be unsaturated so that the reaction rate can be slightly simplified.

4.2 Paper II

The conservation form of the FPE is used to solve the steady state and the time-dependent problem problem efficiently. The steady state solution is the eigenvector corresponding to the zero eigenvalue of the space operator. A model [24] of a Barkai-Leibler oscillator [1] is approximated by an FPE discretized by a finite volume method.

4.2.1 The Barkai-Leibler oscillator

Molecular clocks are present in many organisms and cell types. It is important to control periodic processes without having to depend on external signals such as daylight. It is common with molecular clocks that oscillate in 24 hour cycles. They are usually referred to as circadian clocks.

In [1] a model is proposed for a class of biological oscillators that will work in a noisy environment. It is necessary for a reliable molecular oscillator to be robust with respect to variations in copy numbers. The principle of the oscillator is shown in Figure 4.2. It is comprised of an activator A and a repressor R. The activator stimulates the synthesis of itself and of the repressor. The repressor binds to the activator forming inactive complexes I. Because R is degraded at a slower rate than A, all A-molecules will eventually be bound into I-complexes and the synthesis rates of A and R decrease to a base level. As the repressor molecules are degraded as well, the activator molecules will get a new chance to rise in numbers. Since activator molecules are synthesised at a higher rate than the repressors they will have time to grow before the repressors catch up and the copy number peaks.

A specific example of a model in this class of oscillators is published in [24]. It has nine components but is also reduced to a simpler two-dimensional model that still captures the properties of the full model. An SSA simulation of the model is shown in Figure 4.3. An interesting feature of the model is that there are parameter choices that generate oscillations in a mesoscopic simulation, but not in a deterministic one. Another important difference to a macroscopic simulation is that a deterministic description cannot predict the lag of the molecular clock. Because of the internal noise,

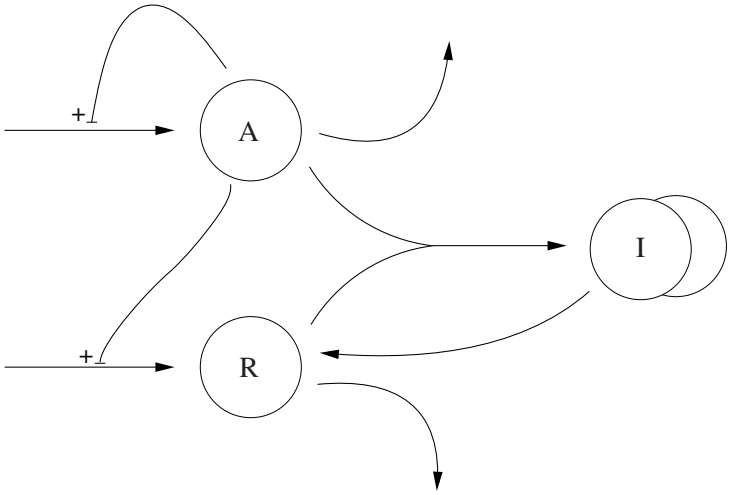


Figure 4.2: The principle of a Barkai-Leibler oscillator.

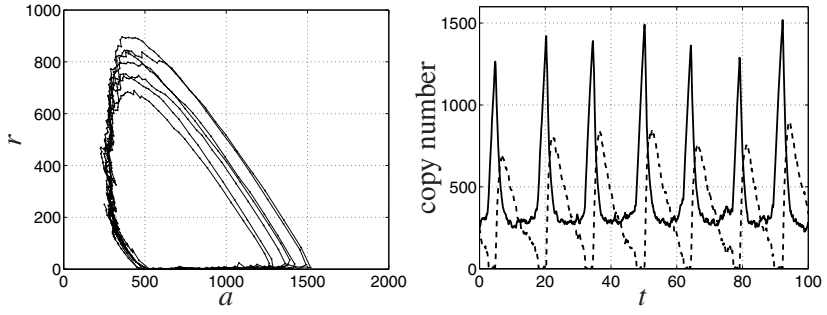


Figure 4.3: A realization of the two-dimensional Vilar model of a circadian clock. Left: a trajectory is plotted in the state space. Right: the copy numbers of A (solid line) and R (dashed line) are plotted as a function of time.

individual cells that are synchronized will lose their phase until they are no longer correlated. The loss of phase is a measurable property that can be compared between simulations and experiments.

4.2.2 The steady state solution as an eigenvalue problem

The master equation has a unique steady state solution if some mild conditions are fulfilled, i.e. as $t \to \infty$ the solution $\mathbf{p}(t) \to \mathbf{p}^{\infty}_{ME}$ [23]. That implies that the space operator of the CME has an eigenvalue zero and that the corresponding eigenvector is $\mathbf{p}^{\infty}_{ME}$. If a conservation form of the FPE, (2.9) is used the approximation also has an eigenvalue zero and an eigenvalue solver can be used

for the steady state problem

$$A_{FPE}\mathbf{p}^{\infty} = 0.$$

The conservative finite volume discretization guarantees that there is an eigenvalue to the space operator that is zero. The corresponding eigenvector converges to the steady state solution of the continuous approximation of the CME. The eigenvalue problem can be solved by efficient methods for finding a few eigenvalues of large sparse systems such as the Arnoldi [16] or Jacobi-Davidson methods [20]. The error in the discretization is controlled by dynamic adaptation of cell sizes and time steps.

4.2.3 Reflecting boundary conditions

Besides the boundary condition from the FPE approximation a boundary condition is needed at the artificial boundary that is introduced by the truncation of the state space. In order to be a conservative discretization the boundary condition must be chosen so that the probability mass is conserved. No probability currents may cross the boundary, which is the same boundary condition that is prescribed at the boundary of the mathematical problem. It is called a reflecting boundary condition [6]. If the solution is zero close to the artificial boundary it is a good approximation, otherwise the computational domain is too small. Since the number of molecules always is limited, it is certain that there is a truncated computational domain that will satisfy the approximation condition.

4.3 Paper III

In order to compare the FPE approximation with SSA the error must be estimated for the two methods. A limit for the error in FPE can be derived from theory for numerical PDEs, while the error in SSA is estimated by statistical methods. The FPE solver and the SSA simulations are evaluated for two models: the toggle switch and coupled flows.

4.3.1 The toggle switch

An implementation of an artificial toggle switch has been made in living bacteria [7] using the circuit shown in Example 1.

A toggle switch has two locally stable attractors in the state space. The function of a toggle switch is to represent an operational pathway junction in an organism. A well-known example is the switch in the λ*-phage* which is a very well-studied virus that attacks *E. coli* bacteria (see for instance [19] for an overview). The virus has two developmental pathways once it enters a bacterial cell. One is to make as many copies of itself as possible and burst the

cell, so called *lysis*, to spread the copies. The other is to incorporate the viral DNA into the bacterial genome and quietly be copied along with the bacterium in a *lysogenic state*. The virus can switch into the lytic state if business goes bad for the bacterium. A toggle switch may direct less dramatic decisions, but it is typical that the switch maintains its position in absence of whatever signal that set it to its current one.

The principle of the artificial toggle switch is to use two genes, E_A and E_B which reciprocally inhibits the other gene through their respective products, A and B. If the protein A is in excess the amount of B will be suppressed and A will remain abundant while B will remain scarce. If some signal enables the number of B-molecules to increase, A synthesis will be inhibited and the system will switch to the other locally stable attractor where B is in excess. When the signal is removed the switch will remain in the same mode. In [7] chemical and temperature signals are used to turn the switch.

In a deterministic description the system will find a stable fixed point (the locally stable attractor) and stay there unless the system is perturbed. In a mesoscopic description there will always be a small probability for the switch to make a random switch between attractors. The probability for such a random event is an important property of the circuit.

Toggle switches may also be a design for introducing variation in a population [14]. Random switches or random initiations of the switch let the circuit become a molecular random number generator. Each individual has a certain probability for the switch to be on or off, and the population may show heterogeneity in proportions determined by the toggle switch.

4.3.2 A reduced model of coupled flows

The model mentioned in Section 4.1.2 and was derived from [4] has interesting properties in two dimensions as well. The enzyme dimensions are simply removed and the rate of A and B syntheses are modified by substituting the amount of enzymes with constants.

The model is interesting since it is an example of when a mesoscopic description is important for rather large copy numbers. When the synthesis rates of A and B are of equal size, the variations in A and B are strongly correlated and the system is poorly described by a macroscopic description. If the system in steady state is perturbed the flow rate is rapidly recuperated by a new balance between A and B. The relaxation back to the steady state occurs along the trajectory $k \cdot a \cdot b$, where k is a constant, at slow scale. Depending on the regulatory parameters, i.e. the strength of the inhibition, the probability density function will deviate strongly from a normal distribution and have a very large relative variance even for rather high copy numbers of A and B.

4.3.3 Error bounds for the FPE approximation

The error in the FPE approximation consists of an error from the truncation of the Taylor-expanded CME and a discretization error. Both parts depends on higher derivatives of $q(\mathbf{x},t) = \sum_r w_r(\mathbf{x})p(\mathbf{x},t)$. The approximation error depends on the third derivatives of $q(\mathbf{x},t)$. The numerical method determines the discretization error. These are both an error in the equation, but using a maximum principle a limit for the error in the solution can be derived.

By solving the problem on grids using different resolutions, the numerical error can be estimated. An adaptive time stepping scheme is used to control the error in the time discretization. It controls the local error of a time step, but the global error can be estimated in the same way as was used for the space operator, solving the problem with different resolutions.

4.3.4 Error estimates for the SSA

The original version of SSA has no discretization errors. The computational error in the probability density function obtained from SSA trajectories is a statistical error that depends on the number of sample trajectories. The convergence is a result of the decrease in the uncertainty as the sampling increases.

The *central limit theorem* implies that the error will decrease as $1/\sqrt{n}$ where n is the number of samples. The absolute error depends on the standard deviation, which can be viewed as a measure of the expected deviation from the exact solution. The standard deviation shows a great variation between different biological models.

In practice it is often not possible to use the exact resolution in $\mathbb{Z}_+^N$ that SSA produces in its computation. It would require too much memory to store the solution in every point in the state space. It is therefore necessary to collect states close to each other in the state space into batches. Comparing the SSA solution to FPE solutions, it is reasonable to use the discretization of the PDE method and let the states inside a computational cell be collected together. The grouping of states will also affect the standard deviation and the statistical error will depend on the space resolution.

4.3.5 Computational work

Using the error estimates, the expected computational work for a certain accuracy ε can be estimated. The work $W_{FPE}(\varepsilon)$ for solving the steady state problem using the FPE method with the error ε is

$$W_{FPE}(\varepsilon) = C_{FPE}(N)\varepsilon^{-\frac{N}{r}}, \tag{4.2}$$

where C_{FPE} does not depend on ε, N is the number of dimensions and r is the order of accuracy of the discretization. For the same problem using SSA, the work $W_{SSA}(\varepsilon)$ is

$$W_{SSA}(\varepsilon) = C_{SSA}\varepsilon^{-2}, \tag{4.3}$$

where C_{SSA} is independent of ε. Hence, for a second order accurate method ($r = 2$) the work grows slower for the FPE method for $N < 4$ when ε decreases. The formulas (4.2) and (4.3) is a general measure for how well the methods will do for problems in a certain dimension, but the actual computational work also depends on C_{FPE} and C_{SSA}. If the error in the intersection point ε_0 of W_{FPE} and W_{SSA} is much smaller than what is used in practical computations, then the method with the faster asymptotic convergence is still slower in practice.

For the steady state problem the ergodicity property of the underlying Markov process can be exploited. Instead of computing many trajectories, one single trajectory is computed for a long time. After an initial transient phase the trajectory is a realization of an individual in the steady state population. The time that is spent in each state is accumulated and is proportional to the probability for being in that state. Compared to using the SSA for time-dependent problems this method is very efficient, even though the convergence properties are the same, simply because C_{SSA} is much smaller for the steady state problem.

4.3.6 Comparison of FPE and SSA approaches

The FPE method is much more efficient for the two-dimensional test problems, while SSA is faster for the problems in three and four dimensions. However, the biological model has a very large impact on the actual computational times. The main purpose in this paper was to examine the convergence rates and to compare the error of the methods in an appropriate manner.

SSA is sensitive to scales in the state space and time and there are methods that attempts to improve the algorithm in this respect, see e.g. [3] and [13]. Any such remedy must introduce some approximation and a fair evaluation requires some carefulness. It is important to establish which accuracy that is needed and probably also a characterization of biological models in order to assign the right method to the problem.

The FPE has less problems with scale. The downside is that it is hard to determine how suitable it is *a priori* since it depends on the solution. For two-dimensional problems that have multi-nodal solutions such as the toggle switch, or near-critical fixed points such as the coupled flows example, FPE seems well-suited, especially for accurate solutions.

4.4 Paper IV

The FPE approximation may not be a good approximation for an entire model. However, the state space can be divided in order to apply the approximation on a subspace of the state space. The subspace that is not approximated is treated as a CME without approximations.

4.4.1 Splitting the state space

By dividing the variables in two subsets X and Y the state space is split into two subspaces that are spanned by the variables in the two subsets. Correspondingly, the state vector is split into $\mathbf{x} \in X$ and $\mathbf{y} \in Y$, and $\mathbf{n}_r$ into μ_r and η_r so that a reaction r corresponds to a state transition from state $(\mathbf{x}+\mu_r, \mathbf{y}+\eta_r)$ to $(\mathbf{x}, \mathbf{y})$. Now, the CME (2.7) can be rewritten

$$\frac{\partial p}{\partial t}(\mathbf{x},\mathbf{y},t) = \sum_r q_r(\mathbf{x}+\mu_r, \mathbf{y}+\eta_r, t) - \sum_r q_r(\mathbf{x},\mathbf{y},t),$$

and the equation can be divided into two parts

$$\frac{\partial p}{\partial t}(\mathbf{x},\mathbf{y},t) = \sum_r q_r(\mathbf{x}+\mu_r, \mathbf{y}+\eta_r, t) - q_r(\mathbf{x}+\mu_r, \mathbf{y}, t) + \sum_r q_r(\mathbf{x}+\mu_r, \mathbf{y}, t) - q_r(\mathbf{x},\mathbf{y},t). \quad (4.4)$$

If the division of variables is such that the subspace Y is suitable for FPE approximation, the first part of (4.4), where the state is constant in the subspace X, can be approximated. The second part is treated directly as a CME.

4.4.2 A model of gene regulation

Proteins are functional units in cells and are also part of the regulation of their own synthesis by taking part in the *transcription* (synthesis of RNA from the DNA template) as well as the *translation* (synthesis of proteins from an mRNA template). One example in [9] captures the main aspects of what is called the *central dogma of molecular biology*, i.e. how the information in the genome is translated to functionality. Here, a somewhat simplified model is derived from the model in [9] by letting the gene product regulate the transcription rate by binding to the gene regulatory region on DNA. In the paper it is the dimer of the gene product that binds. The full model will be treated in Section 4.5. The simplified model is shown in Figure 4.4. In addition to transcription and translation, the figure shows the reactions for degradation of mRNA and the gene product M (curly arrows), and transitions between binding configurations of the gene regulatory region on the DNA. There are two binding sites, S_1 and S_2, in the regulatory region. The binding is *cooperative*. That means that M-molecules bind stronger to the binding sites if both sites are bound to an M-molecule. As a simplification the binding to the S_1 is assumed to be so much stronger that it is always occupied before the S_2 and never unoccupied while S_2 has a bound M-molecule. An M-molecule bound at S_1 enables transcription of the gene unless S_2 is occupied as well, which blocks mRNA production. The number of genes is not changing in the model.

For this model it is obvious that the FPE approximation is not suitable for the representation of the gene binding configurations, because there are few

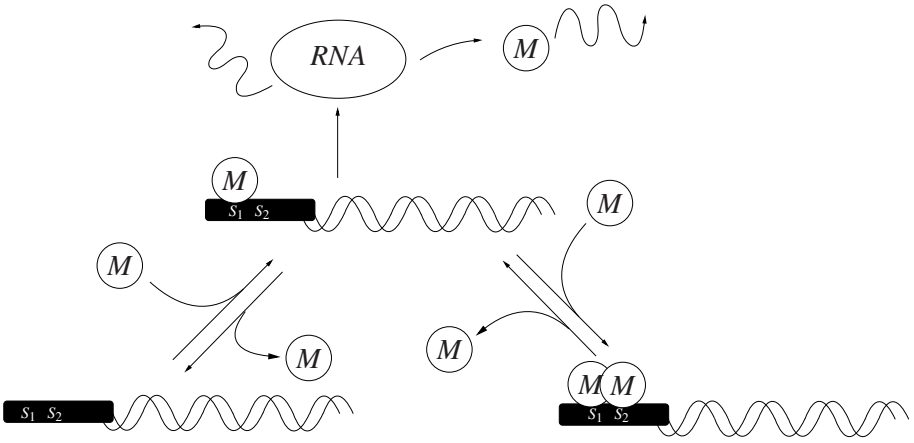


Figure 4.4: The reaction scheme of the gene regulation example.

gene copies and hence very few states in these dimensions of the state space. FPE approximation is used only for the $RNA \times M$ subspace.

4.5 Paper V

The operator splitting introduced in Section 4.4 is applied on a problem in five dimensions. It proves to be necessary to use a fourth order accurate discretization of the state space.

4.5.1 An extended model of gene regulation

The example published in [9] differs from the model in Section 4.4 by the addition of D, the dimer of M, which binds to the regulatory region instead of M. The full model is shown in Figure 4.5. In this model D not degraded but break down into M monomers. With respect to the gene regulation, D works as M did before. The FPE approximation is used in the $mRNA \times M \times D$ subspace.

When the system is simulated it turns out the mean values and the variances for binding states with zero or one bound D-molecule vanish. That means that the probability for being in a binding state where both binding sites are occupied becomes close to one. The solution obviously shrinks to a nearly three-dimensional distribution in the five-dimensional space.

The space discretization is a conservative fourth order finite difference method on a uniform grid. When higher derivatives become large it is not only the discretization error that is affected but also the error in the FPE approximation of the CME. Since $h > 1$ if the FPE approximation shall be an increase in efficiency, a higher order of accuracy may still be justified if the discretization error dominates the error.

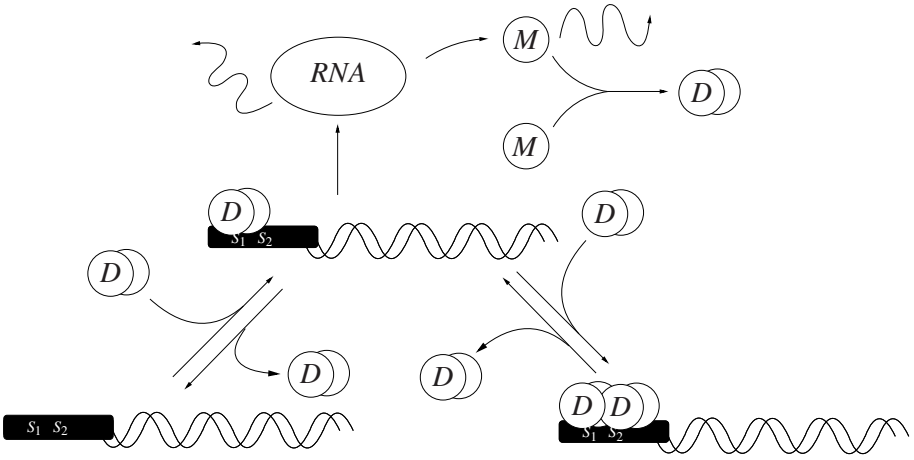


Figure 4.5: The reactions of the extended gene regulation example.

4.5.2 Comparing the PDE and SSA approaches

The resolution must be sufficiently high for the PDE error estimation to work. The estimate relies on the convergence rate which is an asymptotic property, and not valid for too low resolution. For high dimensions it might be hard to resolve the problem well enough for a correct estimate, especially if the grid is not adaptive, since the number of grid points increase rapidly if the dimension is high.

Here, a reference solution is computed by SSA for an estimate of the error in the FPE solution. Using the error estimate for SSA from Section 4.3 the error in the reference solution can be estimated, and it is small enough to be used for computing the error in the FPE solution. The computation is made in five dimensions, but are projected into different low-dimensional subspaces for comparison. The computation time is shorter for SSA when the solutions are compared at $t = 200s$, but the FPE time discretization is adaptive. The time step is modified so that the local error is bounded by an error tolerance [17]. So, comparison for longer simulations allowing larger time steps may be different. An interesting result is that SSA has a certain dependence on the dimensionality of the problem.

4.5.3 The fourth order discretization

In a comparison of solutions using the FPE approach and SSA it is clear that FPE is not able to approximate the solution well enough using the same second order discretization as in Section 4.4 for a reasonable resolution. The large error in the second order method is explained by the large higher derivatives of the solution. By implementation of a fourth order accurate finite difference method, the error in the FPE approach can be controlled by a limited num-

ber of discretization point and the solutions become comparable to the SSA results.

5. Conclusion

This thesis describes how the solution of the CME can be approximated numerically using a FPE and finite differences or finite volume discretizations. It is an advantage to use a conservative discretization since the steady state problem can be solved more efficiently and the total probability mass is preserved. The computational space can be divided into subspaces so that approximations can be applied to a part of the problem which results in an operator splitting. The splitting is important since the method is well suited for low-dimensional problems and the prerequisites for approximation is often not met for the entire computational space of high-dimensional models.

A recurring theme has been the comparison to SSA which is natural since it is the method that is used for this kind of problems in molecular biology. The result of the comparison is very problem dependent. FPE approximation can be much more efficient for very stiff problems with large variation and low dimension. The advantage of PDE methods is that adaptive methods developed for such problems can be applied.

High-dimensional problems have been a trend in numerical analysis for the last years and it will continue since it opens up for important applications such as quantum chemistry, financial mathematics and not least molecular biology. A common denominator of these application fields is the stochastic perspective. Molecular biology is an application with many interesting problems. Biochemical reactions in non-homogeneous volumes and time-delays in reactions are two examples. We expect that numerical analysis will provide useful and powerful methods for problems in molecular biology in the years to come.

6. Sammanfattning på svenska

Den här avhandlingen handlar om att utveckla tekniker för att göra simuleringar av modeller för hur gener och proteiner samspelar. Modellerna, som tas fram av molekylärbiologer, är ganska enkla eftersom de inte innehåller så många komponenter, men det är ändå svårt att förstå hur de fungerar och vad som händer om man ändrar några detaljer. För att kunna göra experiment på modellerna behöver man simulera dem i en dator. Simuleringarna kan bli mycket krävande så det gäller att utveckla metoder som är snabba och inte kräver för mycket minne.

6.1 Molekylärbiologi

Generna lagar all information om hur alla levande organsimer fungerar. Det gör de genom att innehåller en mall för hur en cell ska kunna bygga olika proteiner. Proteinerna är sedan de molekyler som utför alla funktioner i cellen. Molekylärbiologi handlar om att beskriva hur dessa processer fungerar och hur cellen reglerar produktionen av proteiner och därmed hela cellens liv.

Modellerna beskrivs i termer av kemiska reaktioner. Ett fåtal gener och proteiner väljs ut för att de är intressanta i visst sammanhang. De kemiska reaktionerna beskriver hur proteinerna och generna påverkar varandra. En stor skillnad jämfört med den kemi som man utför i fabriker eller provrör är att det handlar om mycket få molekyler i en cell. En gen finns bara i ett fåtal kopior och många proteiner har färre än hundra kopior. Eftersom det finns så få molekyler kan slumpmässiga variationer i cellen få stor betydelse och det är viktigt att studera hur modellen fungerar under dessa störningar. Därför använder man en statistisk beskrivning som inte bara anger hur många kopior det finns i medeltal av ett visst protein utan också hur stor sannolikheten är för olika antal kopior. Problemet är att om modellen innehåller flera olika molekylslag blir beskrivningen omfattande. Vi behöver nämligen inte bara veta sannolikheten för varje molekylantal utan sannolikheten för varje kombination av molekylantal (x molekyler av den ena *och* y molekyler av den andra). Varje kombination av molekylantal kallas för ett tillstånd. Vilka reaktioner som kan inträffa i en cell beror på vilket tillstånd den befinner sig i, d.v.s. hur många molekyler det finns av de olika molekylslagen.

Även om det matematiska ramverket för att formulera detta matematiskt är känt sedan länge är det i praktiken ofta svårt att lösa ekvationerna. Förutom för några enkla fall måste man använda datorberäkningar för att lösa problemet.

Ekvationen som beskriver sannolikheten för att cellen ska vara i ett visst tillstånd kallas för *den kemiska masterekvationen*. Det är denna ekvation vi vill lösa.

6.2 Numeriska metoder

Även om en matematisk ekvation beskriver precis vad man vill veta är det inte alltid möjligt att lösa den exakt. För att en dator ska kunna behandla problemet konstruerar man en ekvation som approximerar den ekvation man egentligen vill lösa, men som passar en dator sätt att arbeta. Den approximativa ekvationen konstrueras så att den blir noggrannare om man använder mer minne och längre beräkningstid.

Det finns ofta många sätt att hitta en approximativ ekvation men även om de används för att uppskatta samma lösning kan de ha väldigt olika egenskaper. Det är viktigt att metoden är noggrann, effektiv och stabil. Noggrannheten innebär att felet i lösningen beror på hur mycket minne man tilldelar problemet. Det viktiga är att lösningen blir bättre och bättre ju mer minne som används. Metoder skiljer sig åt i hur stor förbättring i felet man får för en viss ökning av minnet. Effektiviteten innebär att det inte får ta för lång tid att lösa problemet och har ett samband med hur mycket minne som används. Slutligen behöver metoden vara stabil, d.v.s. inte förstärka små fel i datorberäkningarna. Om metoden inte är stabil kommer bruset att förstärkas så att lösningen försvinner ungefär som när man får rundgång i ett högtalarsystem.

Den här avhandlingen handlar om en metod att approximera den kemiska masterekvationen och jämföra den med en annan numerisk metod. Det stora problemet med den kemiska masterekvationen är att den är ett högdimensionellt problem. Varje molekylslag som beskrivs i modellen bidrar med en dimension till problemet och i en cell finns det en stor mängd olika molekylslag. Även om modellerna är förenklade och bara innehåller fyra eller fem molekylslag är det tillräckligt många dimensioner för problemet ska bli väldigt stort, vilket ställer speciella krav på beräkningsmetoderna.

6.3 Slutsatser

Avhandlingen beskriver en approximation av den kemiska masterekvationen med en *partiell differentialekvation* (PDE) och undersöker metodens fördelar gentemot den metod som dominerar fältet idag den s.k. *stochastic simulation algorithm* (SSA) och jämför vilken av metoderna som är effektivast. Det visar sig att metoderna passar olika bra för olika problem och det gäller att utnyttja de respektive metodernas starka sidor. En metod för att dela upp problemet i olika delar så att olika approximationer kan användas på de olika delarna utvecklas också.

7. Acknowledgements

I am very grateful for my supervisors Per and Måns. Per, you are always available and very patient, which I have appreciated very much. I suspect that you early made a psychological profile of me and made a secret plan for keeping me on track. You will correct me if I'm wrong - you always do. Måns, you have unknowingly been an inspiration for me ever since your lectures during my undergraduate years and it is always a great pleasure to have a meeting with you.

There are so many colleges I would like to thank for making my work easier. You make other things easier as well, but I will keep this professional. Johan, our discussions stand out in this respect. They have have meant much. All the rest of you is not left with just the good karma. Thank you everybody! I hope you will collect the favours you have done to me so I get the opportunity to see you often in the future. Take care!

This work was funded by the Swedish Research Council, the Swedish National Graduate School in Scientific Computing and the Swedish Foundation for Strategic Research.

References

[1] N. Barkai and S. Leibler. Circadian clocks limited by noise. *Nature*, 403:267–268, 2001.

[2] O. G. Berg. A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.*, 71:587–603, 1978.

[3] Y. Cao, D. T. Gillespie, and L. R. Petzold. The slow-scale stochastic simulation algorithm. *J. Chem. Phys.*, 122:014116, 2005.

[4] J. Elf, J. Paulsson, O. G. Berg, and M. Ehrenberg. Near-critical phenomena in intracellular metabolite pools. *Biophys. J.*, 84:154–170, 2003.

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

[6] C. W. Gardiner. *Handbook of Stochastic Methods*. Springer-Verlag, Berlin, 2nd edition, 1985.

[7] T. S. Gardner, C. R. Cantor, and J. J. Collins. Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, 403:339–342, 2000.

[8] D. T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[9] J. Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *J. Chem. Phys.*, 122:184102, 2005.

[10] A. Greenbaum. *Iterative Methods for Solving Linear Systems*. SIAM, Philadelphia, 1997.

[11] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations, Nonstiff Problems*. Springer-Verlag, Berlin, 2nd edition, 1993.

[12] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II, Stiff and Differential-Algebaric Problems*. Springer, Berlin, 2nd edition, 1996.

[13] E. L. Haseltine and J. B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic kinetics. *J. Chem. Phys.*, 117(15):6959–6969, 2002.

[14] M. Kærn, T. C. Elston, W. J. Blake, and J. J. Collins. Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, 6:451–464, 2005.

[15] J. Keizer. *Statistical Thermodynamics of Nonequilibrium Processes*. Springer-Verlag, New York, 1987.

[16] R. B. Lehoucq, D. C. Sorensen, and C. Yang. *ARPACK users' guide. Solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1998.

[17] P. Lötstedt, S. Söderberg, A. Ramage, and L. Hemmingsson-Frändén. Implicit solution of hyperbolic equations with space-time adaptivity. *BIT*, 42:134–158, 2002.

[18] D. A. McQuarrie. Stochastic approach to chemical kinetics. *J. Appl. Prob.*, 4:413–478, 1967.

[19] M. Ptashne. *A genetic switch: gene control and phage λ*. Cell Press & Blackwell Scientific Publications, cop., Cambridge, Mass., 1986.

[20] L. G. Sleijpen and H. A. Van der Vorst. A Jacobi-Davidson iteration method for linear eigenvalue problems. *SIAM Rev.*, 42:267–293, 2000.

[21] Z. Szallasi, J. Stelling, and V. Periwal, editors. *Systems Modeling in Cell Biology : From Concepts to Nuts and Bolts*. MIT Press, Cambridge, Mass, 2006.

[22] H.A. van der Vorst. BiCGSTAB: A fast and smoothly converging variant of the Bi-CG for the solution of nonsymmetric linear systems. *SIAM J. Sci and Stat. Comp*, 13:631–644, 1992.

[23] N. G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam, 2nd edition, 1992.

[24] J. M. G. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proc Natl Acad Sci USA*, 99:5988–5992, 2002.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 358

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-8293

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2007