ACCURATE STOCHASTIC SIMULATION METHODS FOR HOMOGENEOUS BIOCHEMICAL NETWORKS

by

Farida Ansari

Bachelor of Science, University of the Punjab, Pakistan, 1992

Master of Science in Mathematics, QAU, Pakistan, 1995

A thesis presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Science
in the program of
Applied Mathematics

Toronto, Ontario, Canada

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Acknowledgements

I would first like to thank my supervisor Dr. Silvana Ilie for giving me the opportunity to write a fulfilling thesis. Throughout my thesis-writing period, she provided encouragement, sound advice, excellent teaching, and lots of good ideas. I am also grateful from the bottom of my heart for her patience in editing my thesis and bringing it to its good form.

It is also a pleasure to thank Dr. Dejan Delić and Dr. Jean-Paul Pascal, serving as my committee members and for their valuable comments and suggestions. Furthermore, I would also like to acknowledge Dr. Na Yu for being the chairperson of this committee.

Lastly and most importantly, I would like to thank my husband and my children for their support and understanding throughout my graduate studies. I dedicate this thesis to my parents, Amir Hussain and Jamila Akhtar. They raised me, supported me, taught me, and loved me. I could not be successful without their prayers.

Accurate Stochastic Simulation Methods for Homogeneous Biochemical Networks

Master of Science, 2019

Farida Ansari

Applied Mathematics

Ryerson University

Abstract

Stochastic models of intracellular processes are subject of intense research today. For homogeneous systems, these models are based on the Chemical Master Equation, which is a discrete stochastic model. The Chemical Master Equation is often solved numerically using Gillespie's exact stochastic simulation algorithm. This thesis studies the performance of another exact stochastic simulation strategy, which is based on the Random Time Change representation, and is more efficient for sensitivity analysis, compared to Gillespie's algorithm. This method is tested on several models of biological interest, including an epidermal growth factor receptor model.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

One of the most important disciplines of Biology is Genomics, which focuses on the arrangements, performance, development, mapping and rearranging of genomes. We can define genomes as an absolute set of DNA relating to an organism, including all of its genes. Handling the huge data of Genomics after the rapid progress in Genetics and Molecular Biology became a key challenge. For example, in a mammalian cell, more than 10,000 protein coding genes are controlling its physiological activity and cell differentiation [**1**]. The large amount of data can only be processed and investigated using computer simulations. Without computers, it is not possible to study the interactions of genes and proteins, as required for both modelling and data interpretation. To meet this challenge, a significant amount of work is dedicated to the complexity and refinement of simulation algorithms that could possibly compete with the complexity of living cells.

Computational Biology aims to create refined computer simulations with which biological phenomena could be compared. But so far, no such advanced techniques are available that can accomplish these tasks, especially, for the complex biochemical reactions and the gene networks in the cells.

Around 150 years ago, Robert Brown discovered the existence of random fluctuation. He was a botanist and was studying microscopic living phenomena [**2**]. In 1850, Ludwig Wilhelmy used first order ordinary differential equations to describe the conversion of sucrose into glucose and fructose [**3**]. These ODEs laid a foundation for modelling chemical kinetics. But ODEs only model a continuous-deterministic time evolution for the concentration of species.

For many years, the deterministic rate has proved itself successful in modelling chemical reactions, in both chemistry and biochemistry [**4, 5**]. The law of mass action was the base of the deterministic modelling approach, which gives a connection between the reaction rates and the concentration of molecular elements. This law is useful for predicting the species concentration at all future time states, if the initial molecular concentrations are given. Moreover, the law of mass action is based on the assumption that the chemical reactions should be macroscopic, continuous and deterministic [**6**]. Nevertheless, the chemical reactions are the result of discrete random collisions between independent molecules. Hence, the accuracy of a continuous approach in modelling reactions in a small-scale system, with low molecular amounts, is lost. In this case, the stochastic models are more reliable and accurate than deterministic ones.

In 1940, Max Delbruck made an attempt to model the discrete-stochastic behaviour of a chemically reacting system [**7**]. Then in the 1970s, some tools were developed for these systems with the help of large computers. But there were controversies over the correct approaches for modelling and simulations of stochastic chemical kinetics. So, for two decades, there was no certainty regarding the study of molecular discreteness and randomness in biochemical reaction systems. Arkin and McAdams [**8, 9**] showed that discreteness and stochasticity could be important in those living cells where reactant species are in low numbers.

There are three main approaches of implementing stochastic models:

(i) considering the discrete nature of the molecular count of each species and their random character of occurrence,

(ii) following the theories of thermodynamic and stochastic processes,

(iii) describing the small systems and their unpredictability.

As Fedroff and Fontana remarked [**1**], "Stochasticity is evident in all biological processes". For example, there are very low numbers of molecules of certain species, including DNA and regulatory molecules in living cells. Large variations are observed experimentally in the isogenic populations of these living cells, which are given by stochastic effects in gene expression

[**10, 11**]. Also, MacAdams and Arkins analyzed the interactions controlling the expression of a single prokaryotic gene [**8**]. They studied the time interval between the activation of one gene and the regulatory action of its product on a different gene. This time interval is influenced by the stochastic nature of the transcription initiation intervals and the number of protein molecules produced per transcript. Furthermore, they built the repressilator, an oscillating network in Escherichia Coli, which helped study a particular function carried by biomolecules which interact in a living cell. This network behaves as an electrical oscillator system with fixed time period. But this artificial clock shows a noisy behaviour, which may be caused by the stochastic fluctuations of its components.

Hence, we study below three different modelling approaches for analyzing biochemical processes in a single cell:

- stochastic and discrete,
- stochastic and continuous,
- deterministic and continuous.

Chemical Master Equation (CME)[**22**], is a stochastic and discrete model for homogeneous biochemical systems, which gives the time-evolution equations for the probability of the system to be in all possible states. However, this model is very challenging to solve either analytically or numerically, even for the simplest systems. Alternatively, one can generate numerical realizations or sample trajectories in the state space of the stochastic process. These trajectories may be computed using a Monte Carlo method, called the stochastic simulation algorithm (SSA) developed by Gillespie [**12, 13**]. The SSA provides exact realizations of the CME, meaning that their probability distribution is in exact agreement with that obtained from the Chemical Master Equation. Nonetheless, the SSA may be expensive on some models of biochemical systems, the SSA simulates every reaction event. Thus it becomes very slow on systems that involve a large number of such events. Many real biochemical systems involve a significant number of reactions.

A strategy to deal with the large computational cost of the SSA for systems with many reactions is to employ an approximate Monte Carlo method, which trades some numerical accuracy to gain computational efficiency. One of such stochastic acceleration algorithm is the tau-leaping method of Gillespie [**14**]. In this method, the system does not advance on the basis to the time of the next reaction event, but by a pre-selected time step τ. This time interval encloses more than one reaction event. The tau-leaping strategy utilizes Poisson random variables to approximate the number of times each reaction happens during the step of length τ. Under certain conditions, the tau-leaping scheme may be approximated using normal distributions, leading to the so called Langevin tau-leaping technique. This technique may be viewed as a numerical approximation of a stochastic differential equation called the Chemical

Langerin Equation (CLE) [**15**].

In order to use the tau-leaping simulation technique accurately, the τ is selected in accordance with the leap condition. The leap condition requires that no propensity function changes during that τ [**19, 20**]. Sometimes the CLE can be approximated by the reaction rate equation (RRE), which is a deterministic and continuous model. In other words, RRE model is considered as emerging from the CLE, after removing the stochastic term and the reaction rate accordingly. The modelling in terms of concentration and instantaneous rates of change of the RRE is only applicable when very large number of molecules are present in the system. Under the thermodynamic limit, the deterministic term in the CLE expands like the system size, but the stochastic term expands like the square root of the system size. That makes the ODE part dominant. This implies that the RRE may be derived form the CLE model under certain simplifying assumptions [**15**].

While the stochastic simulation algorithm is easy to implement and it was extensively used for computing the numerical solution of the Chemical Master Equation, it is not as useful as a tool for other important studies of stochastic models of biochemical networks, such as sensitivity analysis [**18**]. For estimating parametric sensitivities, another exact Monte Carlo strategy proved to be more accurate and efficient than the SSA. This strategy, the Random Time Change (RTC) algorithm [**18**] is based on the RTC representation of the stochastic system state of the biochemical networks due to Kurtz [**17, 25**].

This thesis studies the numerical properties of the RTC algorithm and compares it with other simulation methods for the Chemical Master Equation.

ln Chapter 2, we discuss the stochastic models of well-stirred biochemical systems in detail along with some of their stochastic simulation methods, including Gillespie's Stocahstic Simulation Algorithm, the tau-leaping method, the Euler-Maruyama scheme for the Chemical Langevin Equation. Chapter 3 introduces the Random Time Change representation of the Markov process modelled by the CME and the RTC algorithm. In Chapter 4, we test these algorithms on three models of well-stirred biochemical systems including a complex model of epidermal growth factor receptor.

Chapter 2

Background

This chapter provides an introduction to stochastic modelling approaches and simulation methods for homogeneous biochemical networks. Standard modelling of chemically reacting systems employs ordinary differential equations to describe the evolution of the system. However, it was observed experimentally that many biochemical systems are inherently noisy [**8, 9, 23**], it was observed experimentally that cells are intrinsically noisy biochemical networks. In a cell low numbers of reactant lead to statistical fluctuations in molecule numbers and reaction rate. Thus stochastic models are necessary to capture the variability observed experimentally. The stochastic models of homogeneous biochemical systems considered in this thesis are the Chemical Master Equation and the Chemical Langevin Equation.

2.1 Homogeneous Biochemical Systems

Let's start with a process that involves N different types of molecules, or chemical species, denoted by $\{S_1, \dots S_N\}$. These molecules are subject to M types of chemical reactions, $\{R_1, \dots R_M\}$. In general, we consider a system in which the velocity and position of every molecules are known and let the system evolve while keeping track of the future positions and velocities of each molecule. But keeping the record of this molecular dynamics is very expensive. Whenever possible, it is preferred to only consider the evolution of the number of molecules of each species as a function of time. This applies when the system is well-stirred, where molecules of each type are laid out evenly throughout the domain. We also assume that this system is in thermal equilibrium and the volume is constant.

Suppose that we know the number of molecules of each species present at time $t = 0$, and we wish to track the number of these molecules with respect to time. Let us consider a state vector, denoted by $X(t) = (X_1(t), \dots X_N(t))$, where $X_p(t)$ is the number of molecules of species S_p in the system at the time t. The state vector $X(t)$ is a Markov process. A Markov process is a stochastic (random) process in which future behaviour is independent of the past, if the current state of the system is given. In other words, the information about the past behaviour of the system is of no help, if the current system state is provided that will be helpful in predicting the evolution of the process which is time dependent. Moreover, this process is accommodating both a theoretical and a computational analysis. Also, it can adequately model the dynamic behaviour of well-stirred biochemical network.

Whenever, one of the M reactions happens, the state vector will change. For any reaction R_r ($1 \leqslant r \leqslant M$), there is a corresponding state-change vector, $v_r \in R^N$. The *p-th* component of v_r is showing the change in the number of molecules of S_p after the R_r reaction happens. Hence the state vector will change from $X(t)$ to $X(t) + v_r$ after reaction R_r occurs.

Every reaction R_r is associated with a propensity function $\alpha_r(X(t))$, which depends on the molecular amounts of the reactant species. The propensity of reaction R_r is defined as $\alpha_r(X(t))dt$ is the probability of this reaction to occur in an infinitesimal time interval $[t, t + dt)$.

The propensity functions are computed as follows:

First Order: $S_p \xrightarrow{c_r} products$ has a propensity of the form

$$\alpha_r(X(t)) = c_r X_p(t).$$

Second Order: $S_p + S_q \xrightarrow{c_r} products$, with $p \neq q$, has a propensity expressed as

$$\alpha_r(X(t)) = c_r X_p(t) X_q(t).$$

Dimerization: $S_p + S_p \xrightarrow{c_r} products$, has a propensity given by

$$\alpha_r(X(t)) = c_r \tfrac{1}{2} X_p(t)(X_p(t) - 1).$$

The existence of constant c_r is implied by the kinetic theory. For example, assuming that the system is in state x, then

$$S_1 \xrightarrow{c_1} S_2$$

has propensity $\alpha_1(x) = c_1 x_1$ and state change vector $v_1 = (-1, 1, 0, \ldots, 0)^T$,

while

$$S_1 + S_2 \xrightarrow{c_2} 2S_1$$

has propensity $\alpha_2(x) = c_2 x_1 x_2$ and state change vector $v_2 = (+1, -1, 0, \ldots, 0)^T$.

Also,

$$2S_1 \xrightarrow{c_3} S_1 + S_2$$

has propensity $\alpha_3(x) = c_3 x_1(x_1 - 1)/2$ and state change vector $v_3 = (-1, +1, 0, \ldots, 0)^T$. Here $x_1(x_1 - 1)/2$ represent the numbers of times one can choose two out of x_1 existing molecules the species S_1.

2.2 Stochastic Models of Biochemical Systems

To study the behaviour of a biochemical system, it is required to build a mathematical model which considers the components of the system, their state and interactions with other elements. These models are expected to include all the essential features of the system, that can be analyzed using computer simulation or theoretical tools. When modelling the dynamics of a biochemical system, the first issue we come across is to decide which features should be included in the model that are related to the requirements of that model.

The purpose of modelling plays an important role in establishing the features of the model. In many cases, the primary purpose of the modelling is to specify the model's components and the interaction between them. This will help us understand the present state of the specific system. To test that our representation of a system is accurate, we verify whether the theoretical or numerical solution of our model is consistent with the observed experimental data. Another purpose of modelling may be to integrate several models or mechanisms into a bigger model. That will help us investigate how the components of the model interact with each other and what their effects are on the behaviour of the system. Lastly, models may be very useful for designing complex biological experiments and analyzing the results of their experiments.

At first, deterministic approaches were considered for the modelling of biological systems at the level of single cell. Nonetheless, many cellular processes involve species with low molecular counts, which have a non negligible level of randomness. Often, we are not able to model and simulate biological systems of realistic size, in spite of the rapid development in the field of computing technology. And we are still not capturing the complexity of the systems to the full extent from the perspective of molecular dynamics. Sometimes we can exclude many features of the system state, such as position, alignment and momentum of every single molecule of the biochemical system under examination. So, at this level, the dynamics of the biochemical systems are naturally considered stochastic rather than deterministic. Subsequent subsections introduce some stochastic models of well-stirred biochemically reacting systems.

2.2.1 Chemical Master Equation

The state vector $X(t)$ changes whenever a reaction R_r $(1 \leqslant r \leqslant M)$ takes place. Since we are not taking into account the spatial information, we only consider the probability of a reaction taking place. This probability is based on the molecular amounts of the reactant species, for the current state of the system. Thus, we are interested in the probability of the system being in a specific state at time t. The evolution of these probabilities will direct us to the Chemical Master Equation (CME). This is a set of ordinary differential equations (ODEs), where each ODE models the evolution of the probability of the system to be in a given state at time t. Most importantly, the dimension of the ODE is based on the number of all possible states of the system and not on the number of species N, as in the case for deterministic models of chemical reactions.

Let's start with the quantity $P(x, t)$, which is defined as the probability that the state vector $X(t)$ is in state x at time t, that is $X(t) = x$, given that

$X(0)$ is known. Suppose that we know the probability of the system to be in any possible state at time t and we wish to determine the probability of the system to be in any state x at time $t+dt$. One more assumption is made here, that dt is so small that at most one reaction can take place over $[t, t+dt)$.

If the system is in state x at time $t+dt$, then there are only two basic scenarios for time t.

Table 2.1: Two basics scenarios for time t.

. . .

Scenario	State at t	Reaction taken place over $[t, t+dt)$	State at $t+dt$
1	x	0	x
2	$x - v_r$	R_r $(1 \leq r \leq M)$	x

Table 2.1 shows that in the first scenario, the system was in state x at time t and no reaction happend during $[t, t+dt)$. Consequently, the system state at $t+dt$ remains x. In the second scenario, the system was in state $x - v_r$ and one reaction R_r occurred in $[t, t+dt)$, therefore bringing the system to state $x - v_r + v_r = x$ at time $t+dt$.

Let us apply now the *Law of total probability.*

Law of total probability : Suppose that $\{A_1, A_2, ..., A_n\}$ is a partition of a sample space, S. If E is an event, then

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E \mid A_i)\mathbb{P}(A_i)$$

where $\mathbb{P}(E \mid A_i)$ is the conditional probabilities that the event E happens if

A_i occurred.

Now, suppose that I is the event of interest and $J_0, J_1, J_2, ..., J_M, J_{M+1}$ are disjoint and exhaustive events. Then, according to the law of total probability,

$$\mathbb{P}(I) = \sum_{r=0}^{M+1} \mathbb{P}(I \mid J_r)\mathbb{P}(J_r). \tag{2.1}$$

Consider the following notation:

I : the event that the system is in state x at time $t + dt$,

J_0 : the event that the system is in state x at time t,

J_r : the event that the system is in state $x - v_r$ at time t, for $1 \leq r \leq M$,

J_{M+1} : the event that the system is in any other state at time t,

$\mathbb{P}(I \mid J_r)$: the probability of the reaction R_r firing over $[t, t + dt)$.

According to the definition of propensity functions,

$$\mathbb{P}(I \mid J_r) = \alpha_r(x - v_r)dt, \tag{2.2}$$

where α_r is the propensity of the reaction R_r. Also,

$$\mathbb{P}(I \mid J_0) = 1 - \sum_{r=1}^{M} \alpha_r(x)dt, \tag{2.3}$$

where $\mathbb{P}(I \mid J_0)$ is the probability, that no reaction took place over the interval $[t, t + dt)$, Then $\mathbb{P}(I \mid J_0)$ must be equal to 1 minus the probability of any reaction firing in $[t, t + dt)$.

Lastly,

$$\mathbb{P}(I \mid J_{M+1}) = 0, \tag{2.4}$$

where J_{M+1} contains all the states that are more than one reaction away from x, since during $[t, t + dt)$ at most one reaction may occur. Substituting (2.2), (2.3) and (2.4) in (2.1) as per the definition of $P(x, t)$, we get

$$P(x, t + dt) = \left(1 - \sum_{r=1}^{M} \alpha_r(x)dt\right) P(x, t) + \sum_{r=1}^{M} \alpha_r(x - v_r)dtP(x - v_r, t)$$

The above equation may be written as

$$P(x, t + dt) - P(x, t) = \sum_{r=1}^{M} \Big[\alpha_r(x - v_r)P(x - v_r, t) - \alpha_r(x)P(x, t)\Big] dt$$

Dividing by dt, we derive

$$\frac{P(x, t + dt) - P(x, t)}{dt} = \sum_{r=1}^{M} \Big[\alpha_r(x - v_r)P(x - v_r, t) - \alpha_r(x)P(x, t)\Big]$$

Letting dt⟶ 0, we obtain the following system of ordinary differential equations

$$\frac{dP(x,t)}{dt} = \sum_{r=1}^{M} \Big[\alpha_r(x - v_r)P(x - v_r, t) - \alpha_r(x)P(x,t)\Big] \qquad (2.5)$$

Equation (2.5) represents the Chemical Master Equation model for well-stirred biochemical reactions. This is an ODE system with one ODE for each possible system state x. The entries of the state x may only take discrete values.

Generally, the CME is a model of very high dimension, so it can not be handled analytically or computationally. One way of computing the solution of the CME indirectly is by using the Stochastic Simulation Algorithm (SSA) also known as Gillespie's algorithm [**12, 13**]. The SSA generates numerical realizations of the stochastic process $X(t)$ governed by the CME. These realizations are sample trajectories in the state space.

While the SSA is an algorithm that produces exact realizations of the CME, it is often very expensive to simulate numerically on realistic models of biochemically reacting systems. Indeed, the SSA becomes very slow when it simulates every reaction event of a system that has a large numbers of such events, as many reacting systems do. One can attempt to improve its computational time by combining reactions and only updating the state vector after many reactions have taken place. In this case, we are searching for an algorithm that gives up the exactness of the SSA, for the sake of higher simulation speed. Tau-leaping method is one of approximate accelerated algorithms

[**15**], for the Chemical Master Equation.

2.2.2 Chemical Langevin Equation

Gillespie [**15**], introduced a scheme called tau-leaping scheme to improve the computational cost of SSA. As we shall see later, in the SSA, at each iteration, we have to draw a random variable to compute the reaction time and one to evaluate the reaction index and then, accordingly, we update the state vector and the propensity functions. In case there are large numbers of some molecules or some very fast reactions, then a large amount of random number generations are required as well as extra effort is needed to keep their records. It is desirable to design an approximate algorithm which trades some accuracy for a significant gain in efficiency. The basic idea of the tau-leaping method is to know that how many times each reaction channel fires in an interval of predefined length, τ. Let us define $\omega_r(\tau; x, t)$ to be the number of times the reaction channel R_r fires in the time interval $[t, t+\tau)$, for each $1 \leqslant r \leqslant M$, given that $X(t) = x$. Assume that the following *Leap Condition* is satisfied.

Leaping Condition: *Require τ to be small enough such that no propensity function will suffer an appreciable change in its value during* $[t, t+\tau)$.

Then, $\omega_r(\tau; x, t)$ is may be approximated by a Poisson random variable, $\mathcal{P}_r(\alpha_r(X(t))\tau)$, with mean and variance $\alpha_r(X(t))\tau$. Assume that we find

$\tau > 0$ such that for $t \leqslant s \leqslant t+\tau$,

$$\alpha_r(X(s)) \backsimeq \alpha_r(X(t)),$$

then,

$$\begin{aligned}\int_t^{t+\tau} \alpha_r(X(t))ds &= \alpha_r(X(t)) \int_t^{t+\tau} ds \\ &= \alpha_r(X(t))\tau,\end{aligned}$$

where $\alpha_r(X(t))$ is assumed constant with respect to s. Hence,

$$\mathcal{P}_r\Big(\int_t^{t+\tau} \alpha_r(X(s))ds\Big) \backsimeq \mathcal{P}_r(\alpha_r(X(t))\tau).$$

We approximated the number of reactions R_r during $[t, t+\tau)$ by

$$\omega_r(\tau; x, t) \backsimeq \mathcal{P}_r(\alpha_r(X(t))\tau). \tag{2.6}$$

The system state may be updated as

$$X(t+\tau) = X(t) + \sum_{r=1}^{M} \omega_r v_r. \tag{2.7}$$

Substituting (2.6) into (2.7), we derive the tau-leaping method:

$$X(t+\tau) = X(t) + \sum_{r=1}^{M} v_r \mathcal{P}_r(\alpha_r(X(t))\tau), \tag{2.8}$$

which holds when τ satisfies the leap condition. In addition to (2.8), where we choose τ small enough to satisfy the leap condition so that the approximation (2.6) is accurate, we assume that τ is large enough that

$$\alpha_r(X(t))\tau \gg 1, \tag{2.9}$$

for all $r = 1, ..., M$. Since $\alpha_r(X(t))\tau$ is the mean of the Poisson random variable $\mathcal{P}_r(\alpha_r(X(t))\tau)$, the condition (2.9) requires that in the next τ, each reaction channel will be fired, on average, many times. The conditions that τ satisfies the leap condition as well as (2.9), are simultaneously satisfied if the population of each reactant species is large.

When these two conditions hold, the tau-leaping formula (2.8) can be approximated using a standard result from probability theory, that a Poisson random variable with large mean may be approximated by a normal random variable with the same mean and variance. Therefore, if every reaction fires many times over $[t, t+\tau)$, then we may replace the Poisson distribution by the normal distributions in the tau-leaping method. Hence, estimating each Poisson variable with mean and variance $\alpha_r(X(t))\tau$, by a normal distribution having the same mean and variance $\alpha_r X(t))\tau$ gives

$$\mathcal{P}_r(\alpha_r(X(t)\tau) \simeq \mathcal{N}_r(\alpha_r(X(t))\tau, \alpha_r(X(t))\tau)$$

$$= \alpha_r(X(t))\tau + \sqrt{\alpha_r(X(t))\tau}\mathcal{N}_r(0,1), \qquad (2.10)$$

where $\mathcal{N}_r(0,1)$ are statistically independent, normal random variables, with mean 0 and variance 1, for $1 \leq r \leq M$. The above step follows from the fact that

$$N(\mu, \sigma^2) \sim \mu + \sigma N(0,1).$$

Substituting (2.10) into (2.8) gives

$$X(t+\tau) = X(t) + \sum_{r=1}^{M} v_r \left[\alpha_r(X(t))\tau + \sqrt{\alpha_r(X(t))\tau}\mathcal{N}_r(0,1)\right]$$

and thus

$$X(t+\tau) = X(t) + \sum_{r=1}^{M} v_r \alpha_r(X(t))\tau + \sum_{r=1}^{M} v_r \sqrt{\alpha_r(X(t))}\sqrt{\tau}\mathcal{N}_r(0,1) \qquad (2.11)$$

Equation (2.11) is called the *Langevin Leaping formula.* It clearly indicates the increment in the state, that is $X(t+\tau) - X(t)$, as the sum of two terms:

$$\sum_{r=1}^{M} v_r \alpha_r(X(t))\tau : \textit{a deterministic drift term proportional to } \tau$$

$\sum_{r=1}^{M} v_r \sqrt{\alpha_r(X(t))} \mathcal{N}_r(0,1)\sqrt{\tau}$: *a stochastic diffusion term proportional to* $\sqrt{\tau}$.

Equation (2.11) may be viewed as an approximation, which is based on two conditions:

• τ is small enough such that it satisfies the leap condition (no propensity function changes its value remarkably during τ).

• τ is large enough that every reaction occurs much more than once during that interval.

The approximate nature of equation (2.11) shows that $X(t)$ has changed from a discrete (integer-value) random variable to a continuous (real-value) random variable. The discreteness has been lost when the integer-valued Poisson random variable was estimated by a real-valued normal random variable. The Langevin leaping formula (2.11) gives faster simulations than the tau-leaping formula (2.8), for the following reasons:

(a) condition (2.9), that suggests that many reactions are fired over each step,

(b) normal random numbers required by (2.11) are computationally faster to generate than the Poisson random numbers in (2.8).

By subtracting $X(t)$ from both sides of (2.11), we get

$$X(t+\tau)-X(t)=\sum_{r=1}^{M}v_r\alpha_r(X(t))dt+\sum_{r=1}^{M}v_r\sqrt{\alpha_r(X(t))}\sqrt{dt}\mathcal{N}_r(0,1)$$

Taking $\tau \longrightarrow dt$ and $dt \longrightarrow 0$, we derive

$$dX(t)=\sum_{r=1}^{M}v_r\alpha_r(X(t))dt+\sum_{r=1}^{M}v_r\sqrt{\alpha_r(X(t))}dW_r(t). \tag{2.12}$$

Here $W_r(t)$ is an independent scalar *Brownian Motion* for each $1 \leq r \leq M$, and

$$dW_r(t)=W_r(t+dt)-W_r(t)\simeq\sqrt{dt}\,\mathcal{N}_r(0,1)$$

Definition :*A scalar standard Brownian motion, or standard Wiener process over* $[0,T]$ *is a random variable* $W(t)$ *that depends continuously on* $t \in [0,T]$ *and satisfies the following three conditions.*

1. $W(0)=0$ *with probability 1,*

2. *for* $0 \leq s < t \leq T$*, the random variable* $W(t)-W(s)$ *is normally distributed with mean zero and variance* $(t-s)$*; equivalently,* $W(t)-W(s) \sim \sqrt{t-s}N(0,1)$*, where* $N(0,1)$ *denotes a normally distributed random variable with zero mean and unit variance,*

3. *for* $0 \leq s < t < u < v \leq T$ *the increments* $W(t)-W(s)$ *and* $W(u)-W(v)$ *are independent.*

Equation (2.12) is called the *Chemical Langevin Equation* (CLE). The CLE is a system of stochastic differential equations in the system state $X(t)$. The dimension of the CLE is N, the number of reacting species. The CLE model applies when condition (1) and (2) above apply, that is, when all species have large molecular numbers. Further, we use the formula (2.11) to solve (2.12) numerically.

Next, we will derive the reaction rate equations (RRE) as a series of limiting approximations.

2.2.3 Reaction Rate Equation

Stochastic differential equations emerge in many fields of physics, but normally we obtain them when we start with a drift term and some form for the diffusion term. The CLE has the drift term ,

$$\sum_{r=1}^{M} v_r \alpha_r(X(t)) d\tau,$$

and diffusion term,

$$\sum_{r=1}^{M} v_r \sqrt{\alpha_r(X(t))} dW_r(t).$$

Below, we assume that the biochemical system is in the thermodynamic limit, when the model becomes macroscopic.

Thermodynamic limit: The system volume Ω *and the molecular counts* X_i, *all approach ∞, in such a way that the species concentrations X_i/Ω remain constant.*

Now, we need to determine the behaviour of the propensity functions in the thermodynamic limit. As the system approaches the thermodynamic limit, all propensity functions grow linearly with the system size. The behaviour of a unimolecular propensity function of the form $c_r X_p$ is easy to see, as c_r is independent of the system size. For a bimolecular propensity function of the form of $c_r X_p X_q$, for which c_r's are inversely proportional to Ω, it can also be shown that it is proportional to the system size.

Hence, near the thermodynamic limit, the deterministic drift term in (2.11) grows like the size of the system, while the fluctuating diffusion term grows like the square root of the size of the system.

In the full thermodynamic limit, the size of the diffusion term of (2.12) will normally become insignificantly small compared to the size of drift term. In that case the Chemical Langevin Equation (2.12) reduces to the reaction rate equations(RRE):

$$\frac{dX(t)}{dt} = \sum_{r=1}^{M} v_r \alpha_r (X(t)). \tag{2.13}$$

Thus, the RRE was derived as a series of limiting approximations of the Chemical Master Equation. The RRE was obtained under the assumption that the system is in thermodynamic limit, therefore this model is valid when all molecular amounts are very large. The tau-leaping method (2.8) and Langevin leaping formula (2.11) clearly yield a connection between the stochastic chemical kinetics and the traditional deterministic chemical kinetics, that is between the CME/SSA and the RRE.

Consider the state vector $X(t)$ in the stochastic approach of the chemical kinetics and $x_p(t)$, the non negative real number representing the concentration of species S_p at time t. Usually, concentrations are measured in M (moles per litre) and the number of molecules in a mole is given by Avogardo's constant, $n_A \approx 6.023 \times 10^{23}$. So $x_p(t) n_A vol$ moles of a species S_p is the concentration $x_p(t)M$ of that species in a fixed volume Ω. The *law of mass action*, is used to derive the RRE.

Law of mass action: *The rate of any chemical reaction is proportional to the product of the concentrations of the reacting substances, with each concentration raised to power equal to the coefficient that occurs in the chemical equation.*

In a more accurate way, we can say that the instantaneous rate of change of a reaction is proportional to the product of the concentrations of the reacting species.

The RRE can be formulated in terms of the concentration vector

$x(t) = (x_1(t), \ldots, x_N(t))^T$ as follows:

$$\frac{dx(t)}{dt} = \sum_{r=1}^{M} v_r \alpha_r(x(t)). \tag{2.14}$$

Consequently, the elementary reactions have the following propensities:

First Order: $S_p \xrightarrow{k_r} products$, has a propensity of the form, $k_r x_p(t)$

Second Order: $S_p + S_q \xrightarrow{k_r} products$, has a reaction propensity given by, $k_r x_p(t) x_q(t)$, where with $p \neq q$.

Dimerization: $S_p + S_p \xrightarrow{k_r} products$, has the following propensity function $k_r x_p(t)^2$.

Remark that the RRE representation (2.13) may be obtained from the CLE, after omitting the diffusion term and applying the transformation from the number of molecules $X_p(t)$ to the concentration $x_p(t)$ of each species S_p. Indeed, let us compare the term of each reaction, while keeping in mind the conversion between concentration and molecule counts as follows:

First Order: $S_p \longrightarrow products$, in this case the deterministic rate $k_r x_p(t)$ gives a molecular rate of $k_r X_p(t)$. So

$$c_r = k_r. \tag{2.15}$$

Second Order: $S_p + S_q \longrightarrow products$, with $p \neq q$, Here the concentration-based rate of change is $k_r x_p x_q M s^{-1}$, with $X_p = x_p n_A vol$ and $X_q = x_q n_A vol$ indicating the number of molecules of S_p and S_q. Equating this rate with the propensity function value, we get

$$c_r X_p X_q(t) = \frac{k_r X_p X_q(t)}{n_A vol}$$

$$\Rightarrow c_r = \frac{k_r}{n_A vol} \tag{2.16}$$

Dimerization: $S_p + S_p \longrightarrow products$, The concentration-based rate $k_r x_p(t)^2$ corresponds to a molecular rate of $2 \times k_r X_p(t)^2/(n_A vol)$. Equating this with the propensity function, we derive

$$\frac{2 \times c_r X_p(X_p - 1)}{2} \approx \frac{2 \times k_r X_p(t)^2}{n_A vol}$$

$$\Rightarrow c_r \approx \frac{2k_r}{n_A vol} \tag{2.17}$$

Consequently, the traditional RRE model of chemical kinetics is considered as a simplification of the CLE , after removing the stochastic terms and the reaction constants are transformed according to (2.15) - (2.17). Also, in the case of dimerization-type reaction, $X_p(t)(X_p(t)-1)$ is approximated to $X_p(t)^2$ because of large molecular number $X_p(t)$.

2.3 Stochastic Simulation Methods for Homogeneous Biochemical Systems

Below, we present several stochastic simulation algorithms for well-stirred biochemical systems, ranging from exact simulation methods for the Chemical Master Equation, to approximate strategies for the solution of the Chemical Langevin Equation and the simplified model of the reaction rate equation. Stochastic models of biochemically reacting systems are solved numerically using Monte Carlo simulation strategies. These strategies generate stochastic trajectories, in accordance with the probability given by the CME or CLE, respectively.

2.3.1 Stochastic Simulation Algorithm

Since the CME model is generally very high dimensional, it is hard to solve it directly, either numerically or analytically, so we need another approach to solve it, and that is to construct numerical realizations of $X(t)$, which are simulated trajectories of the Markov process $X(t)$ with respect to t. The algorithm below computes a single realization of the state vector, instead of an entire probability distribution.

Let us begin by introducing the quantity $g(x, \tau)$, where the system state at time t is $X(t) = x$, defined as the probability that no reaction takes place in the time interval $[t, t + \tau)$. Let consider the time interval $[t, t + \tau + d\tau)$ and suppose that what happens over $[t, t+\tau)$ and over $[t+\tau, t+\tau+d\tau)$ are not dependent on each other. Then, the probability that no reaction happens in the interval $[t, t+\tau+d\tau)$ can be computed as the product of the probability that no reaction happens in the interval $[t, t+\tau)$ and the probability that no reaction happens in the interval $[t+\tau, t+\tau+d\tau)$. Hence,

$$\begin{aligned} g(x, \tau + d\tau) &= g(x, \tau)\times \\ &\Big(1 - \textit{sum of prob. of each } R_r \textit{ firing in } [t+\tau, t+\tau+d\tau)\Big) \end{aligned}$$

We can write this as after using the definition of the propensity function,

$$g(x, \tau + d\tau) = g(x, \tau)\Big(1 - \sum_{k=1}^{M} \alpha_k(x) d\tau\Big),$$

that is,

$$\frac{g(x, \tau + d\tau) - g(x, \tau)}{d\tau} = -\sum_{k=1}^{M} \alpha_k(x) g(x, \tau).$$

Denote

$$\alpha_{sum}(x) = \sum_{k=1}^{M} \alpha_k(x).$$

Consider now the limit $d\tau \longrightarrow 0$ in the above equation. It will give a linear scalar ODE, with the initial condition of $g(x, 0) = 1$. Solving this ODE leads to

$$g(x, \tau) = e^{-\alpha_{sum}(x)\tau}. \tag{2.18}$$

Now, we define an important quantity for the SSA, $p(r|x, \tau)$, as follows.

Given that, $X(t) = x$, $p(r|x,\tau)d\tau$ is the probability that the next reaction:

(A) will be the reaction R_r, and

(B) it will fire in the time interval $[t+\tau, t+\tau+d\tau)$

Then, we have that

$$\begin{aligned}\textit{Prob. of (A) and (B)} &= \textit{Prob. no reaction took place over } [t, t+\tau) \\ &\quad \times \textit{Prob. } R_r \textit{ reaction took place over } [t+\tau, t+\tau+d\tau)\end{aligned}$$

We assume here that $d\tau$ is so small that no more than one reaction can take place over that length of time. Using the definitions of g and α_r, we derive that

$$p(r|x,\tau)d\tau = g(x,\tau)\alpha_r(x)d\tau$$

Substituting (2.18) in the equation above, we obtain

$$p(r|x,\tau) = e^{-\alpha_{sum}(x)\tau}\alpha_r(x),$$

which can be written as

$$p(r|x,\tau) = \Big[\frac{\alpha_r(x)}{\alpha_{sum}(x)}\Big]\Big[\alpha_{sum}(x)e^{-\alpha_{sum}(x)\tau}\Big] \tag{2.19}$$

Here, $p(r|x,\tau)$ may be viewed as a joint density function of two random variables. The two random variables involved are:

- r**-density function**: it is a random variable that provides the index r of next reaction. The discrete random variable $\left[\frac{\alpha_r(x)}{\alpha_{sum}(x)}\right]$, finds the index r of the next reaction according to the rule that the chance of choosing the reaction R_r is proportional to the propensity $\alpha_r(x)$.

- τ**-density function**: this continuous random variable gives the time τ until next reaction. Note that $a_{sum}(x)e^{-a_{sum}(x)\tau}$ is the density function of an exponential distribution with parameter $a_{sum}(x)$.

Often, exponential random variables are used in representing the time elapsed between uncertain events.

To justify the above statements, we give below some propositions and a lemma [**16**].

Proposition 1 [16]: *Consider that X_i are independent exponentially distributed random variables with parameters α_i, for all $i = 1, 2, \ldots, M$, then*

$$X_0 = \min_{1}\{X_i\} \backsim Exp(\alpha_0)$$

where $\alpha_0 = \sum_{i=1}^{M} \alpha_i$.

Proof. To prove this proposition, we recall that for an exponential distribution X with parameter α, $X \sim Exp(\alpha)$, the probability $P(X > x) = e^{-\alpha x}$.

Therefore,

$$\begin{aligned} P(X_0 > x) &= P(\min_i X_i > x) \\ &= P([X_1 > x] \cap [X_2 > x] \cap \cdots \cap [X_M > x]). \end{aligned}$$

Since the distributions X_i are independent, we get

$$\begin{aligned} P(X_0 > x) &= \Pi_{i=1}^{M} P(X_i > x) \\ &= \Pi_{i=1}^{M} e^{-\alpha_i x} \\ &= e^{-x\Sigma_{i=1}^{M}\alpha_i} \\ &= e^{-\alpha_0 x}. \end{aligned}$$

In conclusion, $P(x_0 \leq x) = 1 - e^{-\alpha_0 x}$, which shows that X_0 is exponentially distributed with parameter α_0, $X_0 \sim Exp(\alpha_0)$.

We shall need the next lemma for the Preposition 2.

Lemma [16]: *Assume that $X \sim Exp(\alpha)$ and $Y \sim Exp(\beta)$ are independent random variables which are exponentially distributed, then*

$$P(X < Y) = \frac{\alpha}{\alpha + \beta}.$$

.

Proof. Consider the probability

$$\begin{aligned} P(X < Y) &= \int_0^\infty P(X < Y|Y = y)f(y)dy \\ &= \int_0^\infty P(X < y)f(y)dy. \end{aligned}$$

Since X and Y are exponentially distributions, then

$$\begin{aligned} P(X < Y) &= \int_0^\infty (1 - e^{-\alpha y})\beta e^{-\beta y} dy \\ &= \frac{\alpha}{\alpha + \beta}. \end{aligned}$$

This lemma will be used in the proof of following Proposition.

Proposition 2 [16]: *If $X_i \backsim Exp(\alpha_i)$, $i = 1, 2, \ldots, M$ are independent exponentially distributed random variables and j represents the index of the smallest value of the X_i, then the probability mass function of the discrete random variable j is*

$$\pi_i = \frac{\alpha_i}{\alpha_0}$$

with $i = 1, 2, \ldots, M$, where $\alpha_0 = \sum_{i=1}^{M} \alpha_i$.

Proof: Consider the following probability

$$\begin{aligned}\pi_j &= P(X_j < \min_{i \neq j} \{X_i\}) \\ &= P(X_j < Y)\end{aligned}$$

where $Y = \min_{\ i \neq j} \{X_i\}$. Then, according to Proposition 1, $Y \sim Exp(\alpha_{j^*})$, where $\alpha_{j^*} = \Sigma_{i \neq j} \alpha_i$.

Now, from the lemma above,

$$\begin{aligned}&= \frac{\alpha_j}{\alpha_j + \alpha_{j^*}} \\ &= \frac{\alpha_{j^*}}{\alpha_0}.\end{aligned}$$

Proposition 2 helps us determine the index j of the reaction that fires next, given that $X(t) = x$.

Numerically, to compute the time τ to the next reaction and the index r of the next reaction, we proceed as follows. We draw two random numbers d_1 and d_2 from the uniform distribution in the unit-interval $(0, 1)$. we select τ and r according to the following rules:

$$\tau = \frac{1}{\alpha_{sum}(x)} \ln\left(\frac{1}{d_1}\right) \tag{2.20}$$

$$r = \textit{the smallest integer satisfying} \sum_{k=1}^{r} \alpha_r(x) > d_2 \alpha_{sum}(x) \qquad (2.21)$$

This constitutes the basis for the *Stochastic Simulation Algorithm*.

Stochastic Simulation Algorithm

1. Initialize the time $t = t_0$ and the state of the system , $X(0) = x_0$.

2. Evaluate all the propensities, $\alpha_r(X(t))$, and their sum , $\alpha_{sum}(X(t))$.

3. Generate values of τ and r according to equations (2.20) and (2.21).

4. Update $X(t+\tau) = X(t) + v_r$ and t to $t + dt$.

5. Go back to step 2 or stop.

Practically, a termination condition is also included in step 5, for example,

(i) the simulation will be stopped when t passes a given value,

(ii) when some molecular population number reaches a given upper or lower bound, or

(iii) when the number of iterations reaches to a given number.

2.3.2 Tau-leaping Method

The SSA is rather simple to implement and also logically equivalent to CME. Even when the CME is intractable, the SSA is easy to apply. However, on many models of biochemical reactions arising in applications, the SSA is prohibitively slow and the origin of this high computational cost is related to the factor $\frac{1}{\alpha_{sum}(X(t))}$ in the equation (2.20), as $\alpha_{sum}(X(t))$ could be very large if the population of one or more reactant species is significant.

To implement the SSA , some modifications may be made that will increase the efficiency of its computation [**30**]. Still, any strategy that simulates every reaction event, one event at a time, will be slow on such models. To accelerate the stochastic simulation of well-stirred biochemical systems which involve some fast reactions, Gillespie [**14**], proposed the tau-leaping method. This method advances the system by a pre-selected time τ which step overs several reactions. The tau-leaping scheme requires that the τ be chosen in such a way that it satisfies the "leap condition", that is during that time-interval, the propensity functions will not change their value by a notable amount.

The number of events that will happen in the time τ may be represented by a Poisson random variable $\mathcal{P}(\alpha(X(t)\tau)$ where $\alpha(X(t))dt$ is the probability that an event will happen in a very small interval of time of length dt. Here $\alpha(X(t))$ is considered as positive scalar. If τ is chosen small enough to satisfy the leap condition, then the number of times R_r fires during $[t, t+\tau)$ is approximately $\mathcal{P}_r(\alpha_r(X(t))\tau)$.

Thus, we can leap by a time τ simply by taking

$$X(t+\tau) = X(t) + \sum_{r=1}^{M} v_r \mathcal{P}_r(a_r(X(t))\tau) \tag{2.22}$$

which requires generating M Poisson random numbers for each leap. Equation (2.22) is the tau-leaping method. It is faster than the SSA if the total number of reactions fired over $[t, t + \tau)$, $\sum_{r=1}^{M} v_r \mathcal{P}_r(\alpha_r(X(t))\tau)$, is larger than M.

Clearly, we need a way to estimate in an effective way the largest value of τ that obeys the leap condition. One possible choice for such a τ is to estimate the largest value of τ for which the increment in each propensity over $[t, t + \tau)$ is bounded above by $\varepsilon\alpha_{sum}(X(t))$ for a small ε. During that τ, no propensity function will change its value significantly. Here ε $(0 < \varepsilon << 1)$ is an accuracy control parameter. The explicit tau-leaping simulation procedure will be executed in the following way, regardless of the τ-selection strategy. [**19, 20, 21**].

Tau-leaping Algorithm

1. Initiate $t = t_0$, and $X(t_0) = x_0$.

2. Select a τ that fulfils the leap condition.

3. For each *r=1,2,...,M*, generate the number ω_r of times the R_r fires during $[t, t + \tau)$ as a sample of a Poisson random variable $\omega_r = \mathcal{P}(\alpha_r(X(t))\tau)$.

4. Update $t + \tau \longleftarrow t$ and $X(t + \tau) \longleftarrow X(t) + \sum_{r=1}^{M} \omega_r v_r$.

5. Return to step 2.

Practically, the leap time τ is chosen adaptively, according to current state vector and the values of the propensity functions, and the accuracy control parameter ε.

2.3.3 Euler-Maruyama Method for the Chemical Langevin Equation

The tau-leaping method is inefficient in the limit $\tau \longrightarrow 0$. In that case, it is mathematically equivalent to the SSA because all the ω_r's will also approach zero. In such a small time step, no reaction is fired. Moreover, the tau-leaping scheme becomes ineffective when the largest value of τ, that satisfies the leap condition is less than a small multiple of $\frac{1}{\alpha_{sumX(t)}}$, the expected time to the next reaction. The tau-leaping strategy may be faster than the SSA, but being an approximate scheme, it may lead to errors. For example, a large step may result in negative population numbers of some species. And stiffness is also common in biochemical systems due to the different scale reactions. Stiffness arises when the dynamical modes in the system have different time scales, and the fastest of these time-scales are stable. It may cause instability in simulations for large values of τ.

Since the Poisson random variable $\mathcal{P}(\alpha_r(X(t))\tau)$ has mean $\alpha_r(X(t))\tau$, this is also the expected number of times for the reaction R_r fires over $[t, t+\tau)$. The variance of $\mathcal{P}(\alpha_r(X(t))\tau)$ is also $\alpha_r(X(t))\tau$. Now assume that τ is chosen, such that the value of the mean of the Poisson random variable $\mathcal{P}(\alpha_r(X(t))\tau)$ is large for $1 \leq j \leq M$. In this case, each Poisson random variable approaches a normal random variables with the same mean and variance.

After substituting $\mathcal{P}(\alpha_r(X(t))\tau)$ in (2.8) by $\alpha_r(X(t))\tau + \sqrt{_r(X(t))\tau}\mathcal{N}_r$, where $\mathcal{N}_r$ are independent normal (0,1) random variables, we get

$$X(t+\tau) = X(t) + \sum_{r=1}^{M} v_r \alpha_r(X(t))\tau + \sum_{r=1}^{M} v_r \sqrt{\alpha_r(X(t))}\sqrt{\tau}\mathcal{N}_r \tag{2.23}$$

Equation (2.23) is known as the *Euler-Maruyama* discretization of the Chemical Langevin equation. For the Chemical Langevin Equation the discrete time recurrence (2.9) converges to a continuous time process (2.13), by taking the limit $\tau \to 0$.

Euler-Maruyama Algorithm for the CLE

1. Initiate $t = t_0$, $X(t_0) = x_0$
2. Draw independent samples $\{n_r\}$ from a normal (0,1) distribution where $1 \leq r \leq M$.
3. Update $t \longleftarrow t + \tau$ and

   $X(t+\tau) = X(t) + \sum_{r=1}^{M} v_r \alpha_r(X(t))\tau + \sum_{r=1}^{M} v_r \sqrt{\alpha_r(X(t))\tau} n_r$
4. Return to step 2.

Hence we replaced integer-valued Poisson random variables to real-valued normal random variables. Then, real numbers will be used for counting the amount of molecules of each species present in the system. Even though continuous in the state variable, the recurrence (2.23) runs over a discrete sequence of times. A sequence of random variables $\{X(0), X(\tau), X(2\tau), \dots\}$ is generated that corresponds to the state vector at discrete times $\{0, \tau, 2\tau, \dots\}$.

2.3.4 Approximate Methods for the Reaction Rate Equation

As explained before, in the thermodynamic limit, the Chemical Langevin Equation may be reduced to the reaction rate equations, which is a continuous deterministic model. Hence, the RRE model is a limiting approximation of the CLE.

Stiffness is an important challenge in the numerical solution of ODEs and SDEs. A problem becomes stiff when its solution varies slowly, but there are some solutions that are close to this solution that vary rapidly, so small time steps should be taken by the numerical solution to maintain stability. Efficiency of the numerical solution will be affected by small steps, as the computational time of the algorithm depends on the number of time-steps, and will increase when the stepsize is reduced. We give below a summary of the MATLAB ODE solvers, which may be used for simulating the solution of the RRE.

The MATLAB solvers ode45, ode23 and ode113 are used for nonstiff problems, while the solvers ode15s, ode23s, ode23t and ode23tb may be applied to stiff models of biochemical systems represented using the reaction rate equations. Since for biochemical systems, the number of molecules for each species are positive or zero, the MATLAB solvers should be utilized with the option ‘NonNegative’, or whenever applicable.

Chapter 3

RTC for Discrete Stochastic Processes

In biological networks at the cellular level, intrinsic noise may play an important role. Stochastic effects due to low molecular numbers of some reacting species may be significant in biochemical networks arising in applications.

The most popular stochastic model of well-stirred biochemical systems is the Chemical Master Equation [**22**], while the SSA is an easy to implement simulation method for the Chemical Master Equation, it is not always the exact Monte Carlo strategy of choice when analyzing the CME model. One important tool for analyzing biochemical systems is sensitivity analysis. Although sensitivity analysis is easy to apply to the deterministic continuous model of the reaction rate equations, for the stochastic discrete model of the Chemical Master Equation such an analysis is much more challenging to perform.

Sensitivity analysis of the CME is not the focus of this thesis. Nonetheless, it is one of the key applications of the exact Monte Carlo simulation techniques discussed in this thesis. Biochemical reaction models have kinetic parameters which are often difficult to measure experimentally or are unknown. Small changes in these parameters may have a significant impact on the system's dynamics. It is therefore important to know the influence of the model's parameters on the behaviour of the system, that is, it is crucial to study the parametric sensitivity of biochemical kinetic models. Sensitivity analysis, or parametric sensitivity, measures how the evolution of a given biochemical network depends on the system's parameters. If a small perturbation of a parameter leads to a significant variation in the behaviour of the system, we say that the model is sensitive with respect to that parameter. Otherwise, the system is robust with respect to the given parameter. Sensitivity analysis is a powerful tool for studying properties of the system, in model design and model reduction. For example, in large biological networks, sensitivity analysis can guide us in drug targeting [**8, 23**].

For stochastic discrete models of biochemical systems, one cannot directly apply the sensitivity analysis methods developed for ODEs or SDEs. However, the parametric sensitivities for the Chemical Master Equation, for example, may be estimated using the exact or approximation Monte Carlo algorithms for numerically solving the model. Indeed, one may employ finite difference approximations of the sensitivity of $E(X(t,p))$ evaluated for the given parameter of interest p, by evaluating $E(X(t,p))$ and $E(X(t,p+\varepsilon))$ for a small perturbation $p + \varepsilon$. If the streams of random numbers used to generate samples of $E(X(t,p))$ and $E(X(t,p+\varepsilon))$ are independent, then the statistical estimator of the sensitivity will have a large variance, giving thus large errors for a low number of sample paths [**31**]. To overcome this problem, a large number of trajectories are required to improve the accuracy of the estimator, which results in increased computational efforts. The efficiency of computing are estimation of the parametric sensitivities, can be increased by using common random numbers, which will give an estimator with low variance, that requires fewer sample path for a good accuracy.

One of the most accurate sensitivity estimation methods for the CME is the common reaction path (CRP) developed by Rathinam, *et al* [**18**]. This method employs an exact stochastic simulation algorithm for the Chemical Master Equation, based on the Random Time Change representation (RTC) of the Markov process governed by the CME model [**18**]. This exact stochastic simulation algorithm known as the RTC is studied in detail in this thesis.

3.1 Stochastic Chemical Kinetics

Consider a biochemical reaction system with N chemical species. The sample space Ψ is a set of sample trajectories and ψ, representing the randomness, is an element of Ψ. As before, there are M reaction channels in the system. The propensity function $\alpha_r(X(t), c)$, for $r = 1, \ldots, M$, corresponding to the reaction R_r depends on the system state $X(t)$ and c, which represents one or more kinetic parameters. The probability of the reaction R_r to occur in a very small time interval $[t, t + \delta t)$ is given by $\alpha_r(X(t), c)\delta t$.

The evolution of the Markov process $X(t)$ is governed by the Chemical Master Equation. Exact Monte Carlo simulation methods for the CME include the direct and first reaction methods proposed by Gillespie [**12, 13**], or the next reaction method introduced by Gibson and Bruck [**24**], and the Random Time Change strategy. The Random Time Change representation and algorithm are presented in this chapter.

3.2 Random Time Change Representation

We give below the Random Time Change description of the Markov process $X(t)$ of a biochemical system modelled by the CME. This representation was proposed by Either and Kurtz in [**25**]. This description shows that each reaction channel is carrying its own internal clock. These internal clocks have the rate given by the propensity function of the corresponding reaction channel. These internal times, denoted by, $S_r(t, \psi, c)$, are defined by

$$S_r(t, \psi, c) = \int_0^t \alpha_r(X(s, \psi), c)ds \tag{3.1}$$

for the reaction channel S_r, with $r = 1, \ldots, M$. The S_r are dimensionless quantities. The reaction R_r's clock is modelled as a unit rate Poisson process. The system state is updated according to the following equation:

$$X(t,\psi,c) = X(0,\psi,c) + \sum_{r=1}^{M} v_r Y_r(S_r(t,\psi,c),\psi) \tag{3.2}$$

known as the Random Time Change representation **[18, 25]**. Here $Y_1,\ldots,Y_M$ represent unit rate Poisson processes for the reactions $R_1,\ldots,R_M$, respectively. Equation (3.2) holds pathwise, that is for each realization ψ, here c is a system parameter, it could be an initial condition or reaction rate constant.

The state change vectors v_r are independent of the parameters but the propensity functions α_r are dependent on these parameters. Let us mention again that Y_r are Poisson random variables with unit rate in their frames of internal time and are independent of c explicitly. We assume that the initial condition is independent of c and also deterministic. For instance, c may represent some of the reaction rate parameters. Equation(3.2) allows to find two processes, $X(.,.,c_1)$ and $X(.,.,c_2)$, corresponding to different values of parameters c_1 and c_2, respectively. Hence, they represent functions of the same sample space and thus we can compare them directly.

3.3 Pathwise Computations Using on the Random Time Change Representation

Assume that $Y_1(.,\psi),\ldots,Y_M(.,\psi)$ are realizations of the noise. Let us solve $X(.,\psi,c)$, using the RTC representation (3.2). The approach described below follows that from Rrathinam, *et al* [**18**]. The random internal jump times of the Poisson processes are denoted by I_i^r, where $r = 1, \ldots, M$ and $i = 1, 2, \ldots,$ such that

$$I_1^r < I_2^r < I_3^r \ldots$$

for each r.

Let us take $S_r(t, \psi, c) = I_i^r(\psi)$, where the value of t is the physical time. Then I_i^r means the i-th firing of the reaction R_r happens at time t. Denote this physical time by $T_i^r(\psi, c)$. According to the definition :

$$S_r(T_i^r(\psi, c), \psi, c) = I_i^r(\psi).$$

In what follows, we denote by $T_i(\psi, c)$, the random time of occurrence of the i-th reacting event of any type and by $J_i(\psi, c)$ the (random) type of the i-th reaction event, for $i = 1, 2, \ldots$. Therefore J_i is an integer $1 \leqslant i \leqslant M$. For $i = 1, 2, \ldots$ and $r = 1, 2, \ldots, M$, from the point of view of the information stored, keeping track of the collection (T_i, J_i) is equivalent to recording the collection T_i^r. Either the sequence (T_i, J_i) or T_i^r for $i = 1, 2, \ldots,$ and $1 \leqslant r \leqslant M$, will determine a unique path $X(., \psi, c)$. Moreover, $S_r(t, \psi, c)$ is a piecewise linear function in time t, for $T_i \leq t < T_{i+1}$, given by

$$S_r(t,\psi,c) \longleftarrow S_r(t,\psi,c) + \alpha_r(X(s,\psi,c),c)(t - T_i),$$

for $1 \leqslant r \leqslant M$ and $T_i \leqslant t \leqslant T_{i+1}$.

For $1 \leqslant r \leqslant M$, we denote by $I_+^r(t,\psi,c)$, the following minimum

$$I_+^r(t,\psi,c) = \min\{I_l^r(\psi) \mid S_r(t,\psi,c) < I_l^r, l = 0,1,2,\dots,\},$$

for $r = 1,\dots,M$. Here $I_+^r(t)$ represents the internal time of the next occurrence of reaction channel R_r at real (physical) time t. If $T_1,\dots,T_i$ and $J_1,\dots,J_i$ are computed for a given i, then we can also compute I_+^r for $1 \leqslant r \leqslant M$ and $X(T_i)$. We evaluate T_{i+1} as

$$T_{i+1} = T_i + min\left\{\frac{I_+^r(T_i) - S_r(T_i)}{\alpha_r(X(T_i))} \mid r = 1.\dots,M\right\}.$$

Remark that:

(i) the internal times of the stochastic processes are $S_r(T_i)$, when the physical time is equal to T_i.

(ii) for $T_i \leqslant t \leqslant T_{i+1}$, $S_r(t)$ and $1 \leqslant r \leqslant M$ grows at the constant rate, $\alpha_r(X(T_i))$,

(iii) $I_+^r(T_i)$ are the next internal times of firing of the reactions.

Consequently, the minimum over all $r = 1, 2, \ldots, M$ of $\frac{I_+^r(T_i) - S_r(T_i)}{\alpha_r(X(T_i))}$ is the increment in physical time ($T_{i+1} - T_i$), before the next firing of a reaction. Moreover, J_{i+1} is the index for which the minimum is obtained. Hence,

$$T_{i+1}(\psi, c) = I_{i+1}^{J_{i+1}(\psi,c)}$$

. We can now find the first jump time T_1 as follows:

$$T_1(\psi, c) = \min \left\{ \frac{I_1^r}{\alpha_r(x_0, c)} \mid r = 1, \ldots, M \right\}$$

and $J_1(\psi, c)$ as the index of the minimum. Consequently, $T_1(\psi, c) = I_1^{J_1(\psi,c)}$.

3.4 Random Time Change Algorithm for Stochastic Biochemical Systems

The Random Time Change formulation **[17]**, is a mathematical representation of the stochastic process $X(t)$ which has a direct connection to the sample paths, so it is may be more useful than the Chemical Master Equation. In addition, it is useful for the analysis and derivation of the Monte Carlo algorithm for simulating realizations of the stochastic process, that could be exact or approximate. It is also helpful for the justification of Gillespie's direct method and in understanding some asymptotic properties, scaling limit and approximations. Moreover, this representation may be employed as a component of hybrid stochastic simulation algorithms.

The RTC representation of $X(t)$, due to Kurtz [**17**], constitutes the theoretical basis of the RTC algorithm for the CME, which is an exact simulation method developed bt Rathinam *et al* [**18**] . The following variables are used, S_r, I_+^r and k_r, denoting

k_r: index showing the r-th stream of exponential numbers

S_r: current internal time of reaction R_r

I_+^r: internal time of reaction channel R_r occurring in the next firing

Consider M arrays of entries E_i^r, representing unit exponential random variable where $r = 1, \dots\ M$ and $i = 1, 2, \dots$. These exponential random numbers will be used to select the internal times between successive firing of the unit rate Poisson processes and are also related to the internal firing times, I_i^r. Indeed, $I_{i+1}^r - I_i^r = E_i^r$, independent, unit rate, exponential random variables may be computed using independent uniform random numbers in $[0, 1)$.

Below, we shall describe the random time change algorithm which can be derive from the RTC representation [**18**]. This algorithm computes a single trajectory $X(.,\psi,c)$. It is an exact Monte Carlo simulation method for the CME.

RTC Algorithm

1. Initialize $i = 0$, $T = 0$, $X(T) = x_0$, $S_r = 0$, $k_r = 1$ and $I_+^r = E_i^r$ for $r = 1, \dots\ M$
2. Exit if terminal condition is reached, otherwise continue.
3. Calculate propensity functions $\alpha_r(X(T_i))$, $1 \leqslant r \leqslant M$
4. Compute $T_{i+1} = T_i + min\{\frac{I_+^r - S_r}{\alpha_r(X(T_i))}\}$,
5. Set $X(T_{i+1}) = X(T_i) + v_{r^*}$, where r^* is the index of the min in the above equation.
6. Set $S_r \longleftarrow S_r + \alpha_r(X(T_i))(T_{i+1} - T_i)$
7. Increment k_{r^*}.
8. Set $I_+^{r*} \longleftarrow I_+^{r^*} + E_{k_{r*}}^{r^*}$.
9. Increment i and return to step 2.

Chapter 4

Numerical Results

In this chapter, we study the performance of the stochastic simulation algorithms presented in the previous chapters, by testing them on a rich set of models of well-stirred biochemical systems of practical interest. The focus is on the accuracy and efficiency of the (exact) Random Time Change algorithm, [**18**], by comparing it with Gillespie's stochastic simulation algorithm or SSA, which is an exact Monte Carlo method for the Chemical Master Equation. In addition, we study the accuracy of an approximate strategy of the CME, the explicit tau-leaping method as well as that of a numerical technique for the Chemical langevin Equation model. Each of the RTC, SSA, tau-leaping method and Euler-Maruyama Scheme for the CLE is applied to generate 10,000 trajectories for the biochemical model under consideration. The histograms obtained for the RTC, tau-leaping and Euler- Maruyama algorithms are compared to those generated by the exact SSA, to establish the accuracy of the methods above[**26**].

The biochemical models used for testing have some interesting features, much as a degree of stiffness. Moreover, the third model represents a complex model, that of the epidermal growth factor system which involves 23 species subject to 47 reactions.

4.1 Simple Stiff Model

Let us consider the following system of well-stirred biochemical reaction [**27**]:

$$S_1 \xrightarrow{c_1} S_2$$

$$S_2 \xrightarrow{c_2} S_1$$

$$S_2 \xrightarrow{c_3} S_3$$

The first two, reversible reactions are fast and the third reaction channel is slow. In this model, three species are subject to three reactions. Its stochiometric matrix is

$$V = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

The properties of the reactions are $\alpha_1(X) = c_1X_1$, $\alpha_2(X) = c_2X_2$, and $\alpha_3(X) = c_3X_2$. The reaction rate parameters are $c(1) = 1$, $c(2) = 1$, and

$c(3) = 50$, while the initial conditions are:

$$X(0) = \begin{bmatrix} 1000 \\ 100 \end{bmatrix}$$

The system is studied on the time interval $[0, 1]$. This biochemical system is stiff as it has both fast and slow reactions. A sample path computed with the SSA is plotted in Figure 4.1, representing the time-evolution of the numbers of S_1 molecules and in Figure 4.2, showing the time-dependence of the number of S_2 molecules.

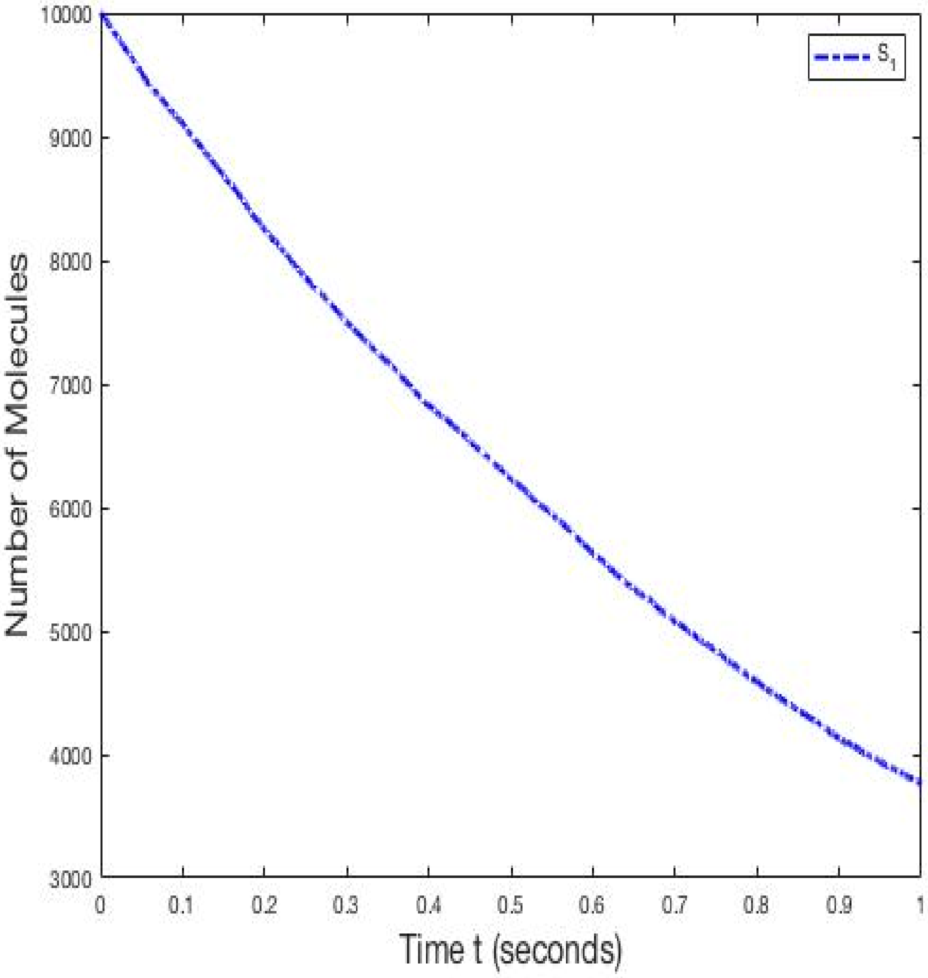

Figure 4.1: Simple stiff model: Evolution in time of the amount of S_1 molecules, on the time interval [0,1], using the SSA

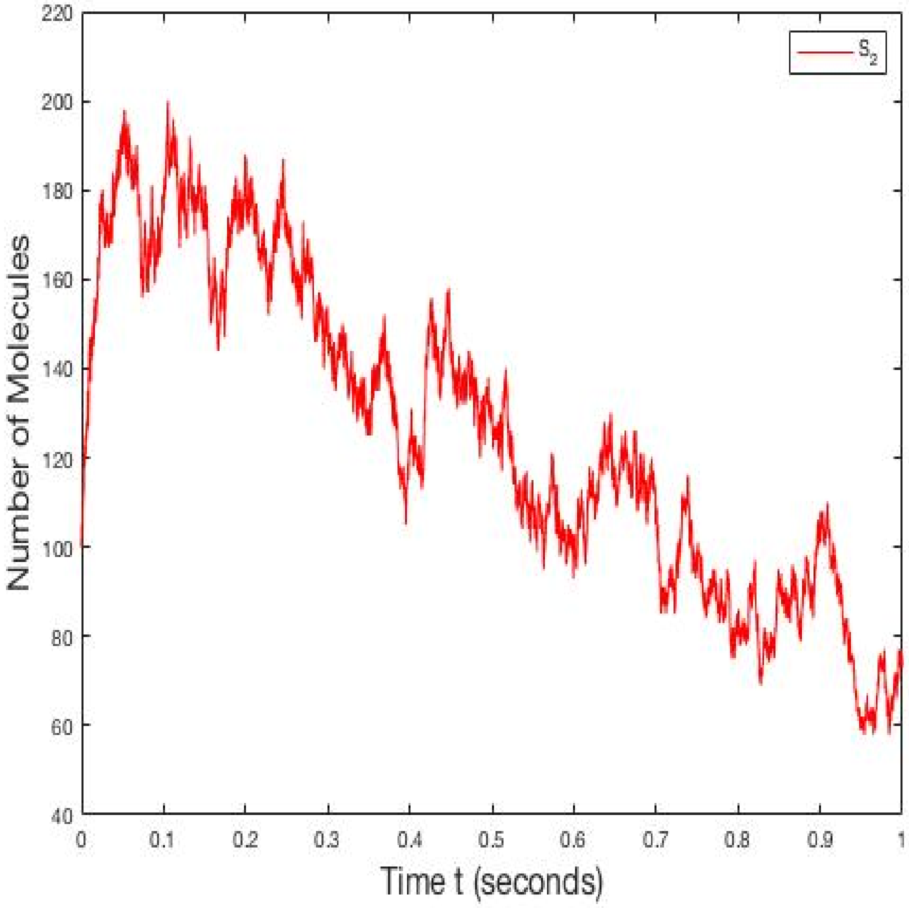

Figure 4.2: Simple stiff model: Evolution in time of the amount of S_2 molecules, on the time interval [0,1], using the SSA.

The histograms for 10,000 trajectories simulated using the SSA, RTC, the tau-leaping method and the Euler-Maruyama (EM) method for the CLE are shown in Figure 4.3, for the molecular amounts of S_1 molecules, and in Figure 4.4, the molecular amounts of S_2 molecules.

We observe the excellent accuracy of the RTC , the tau-leaping and the EM method compared to the exact SSA. In particular, the RTC strategy provides a high accuracy. The tau-leaping and the Eulear-Maruyama techniques were applied for a stepsize $\tau = 2 \times 10^{-4}$.

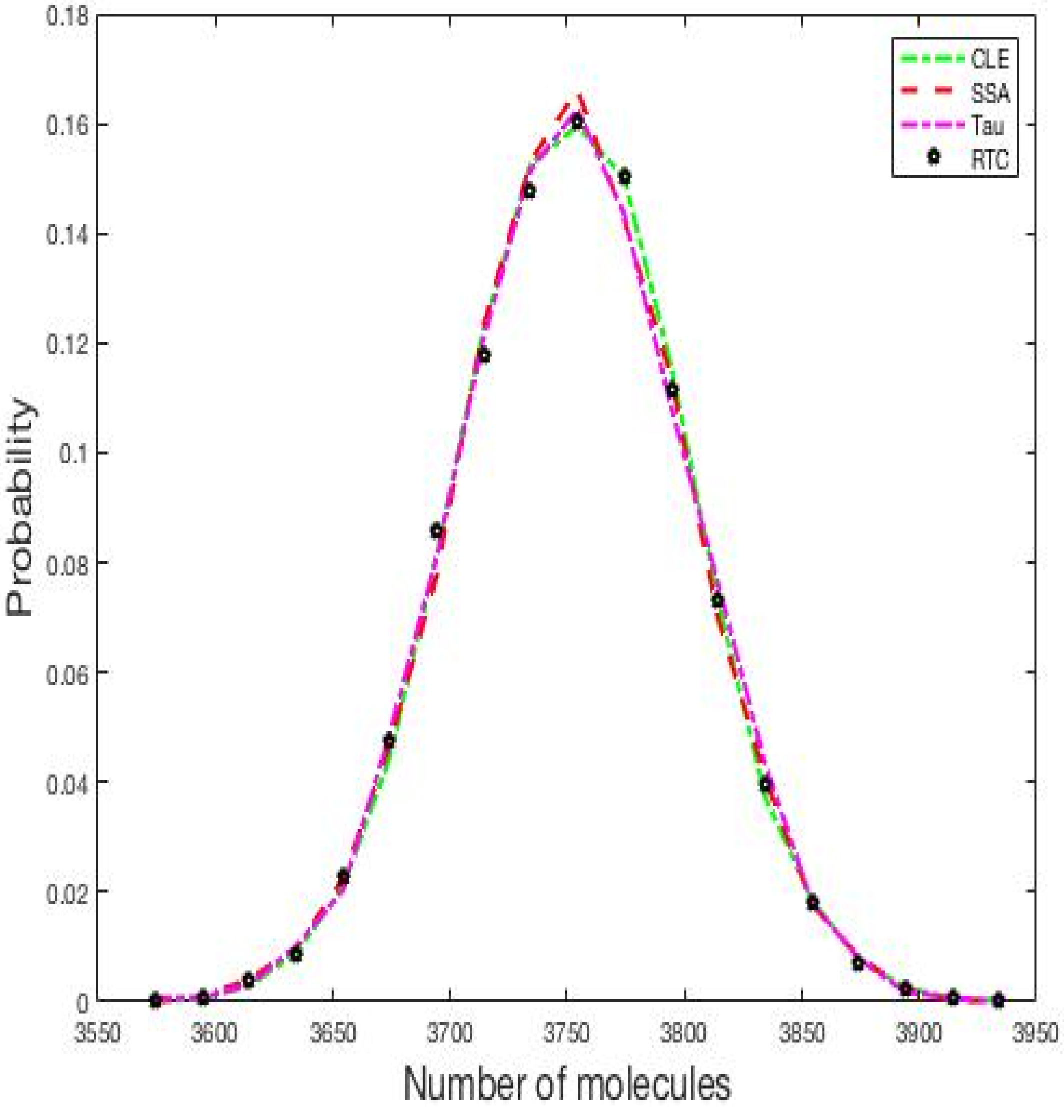

Figure 4.3: Simple stiff model: The histograms of the number of S_1 molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM schemes

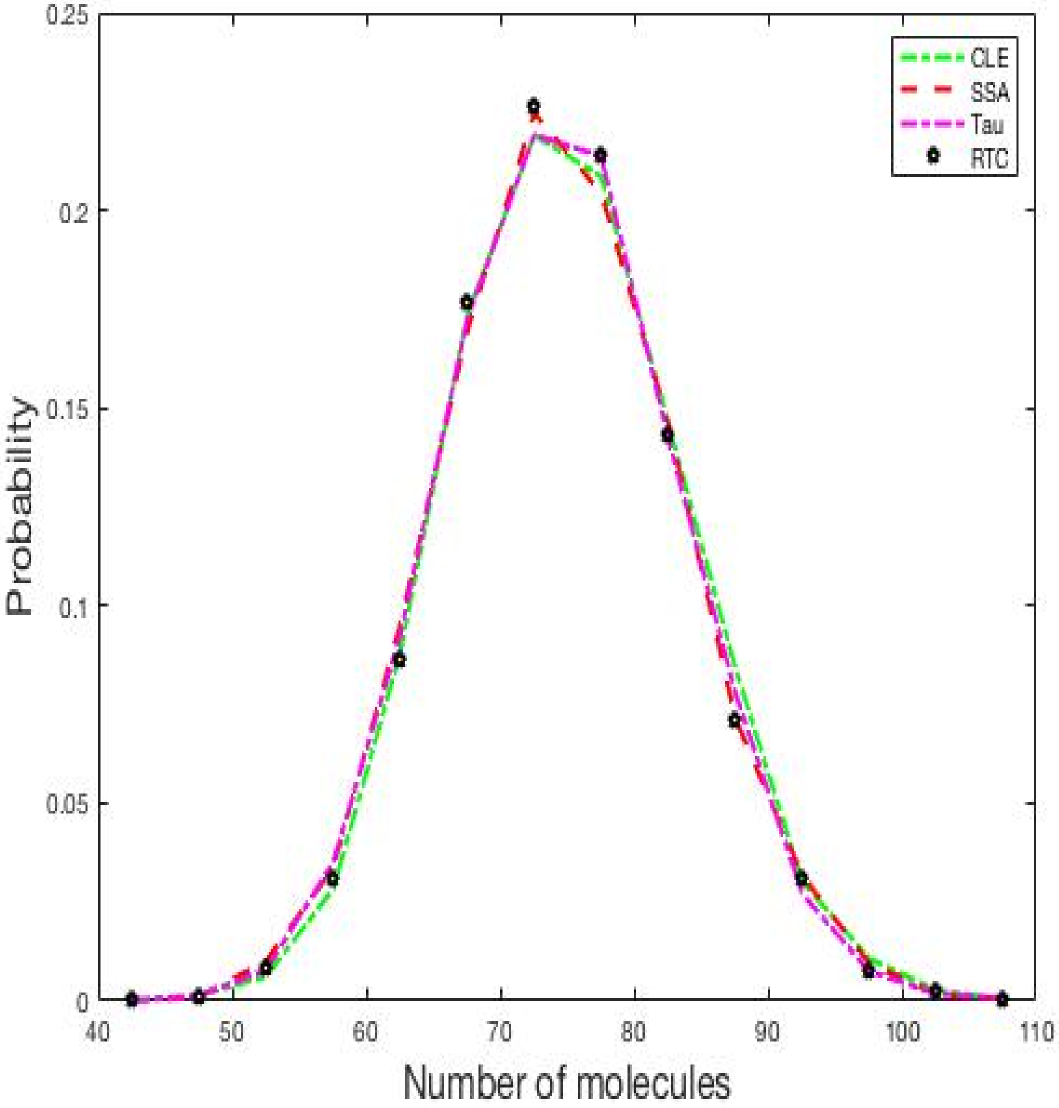

Figure 4.4: Simple stiff model: The histograms of the number of S_2 molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM schemes

4.2 Stiff Decay-dimerization Model

Consider the decay dimerization model, which was proposed by Gillespie[**12, 13, 14**]. The biochemical system consists of three species which are interacting through four reactions:

$$S_1 \xrightarrow{c_1} 0$$

$$S_1 + S_1 \xrightarrow{c_2} S_2$$

$$S_2 \xrightarrow{c_3} S_1 + S_1$$

$$S_2 \xrightarrow{c_4} S_3$$

The reactions rate constants are $c(1) = 0.01$, $c(2) = 10$, $c(3) = 30$ and $c(4) = 0.1$ where as the initial conditions are given by $X(0) = [100, 10000, 0]^T$. The stochiometric matrix for this reaction system is:

$$V = \begin{bmatrix} -1 & -2 & 2 & 0 \\ 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and the propensities are given by $\alpha_1(X) = c_1X_1$, $\alpha_2(X) = c_1X_1(X_1 - 1)/2$, $\alpha_3(X) = c_3X_2$ and $\alpha_4(X) = c_4X_2$. The model is integrated on the time interval $[0, 1]$. As with the previous example, this model exhibits stiffness, as both fast and slow time-scales are present in this system. Stiffness is a challenge for solving numerically this problem.

The decay-dimerization model with the parameters given above is simulated numerically using an ensemble of 10,000 paths generated with SSA, RTC, tau-leaping method and the EM scheme for the CLE, respectively. The approximate methods utilized a fixed stepsize $\tau = \frac{1}{3} \times 10^{-4}$. A sample trajectory of the evolution in time of the number of S_1 molecules, generated using the SSA is given in Figure 4.5. Figure 4.6 depicts the time-dependence of the number of S_2, computed using a single SSA run.

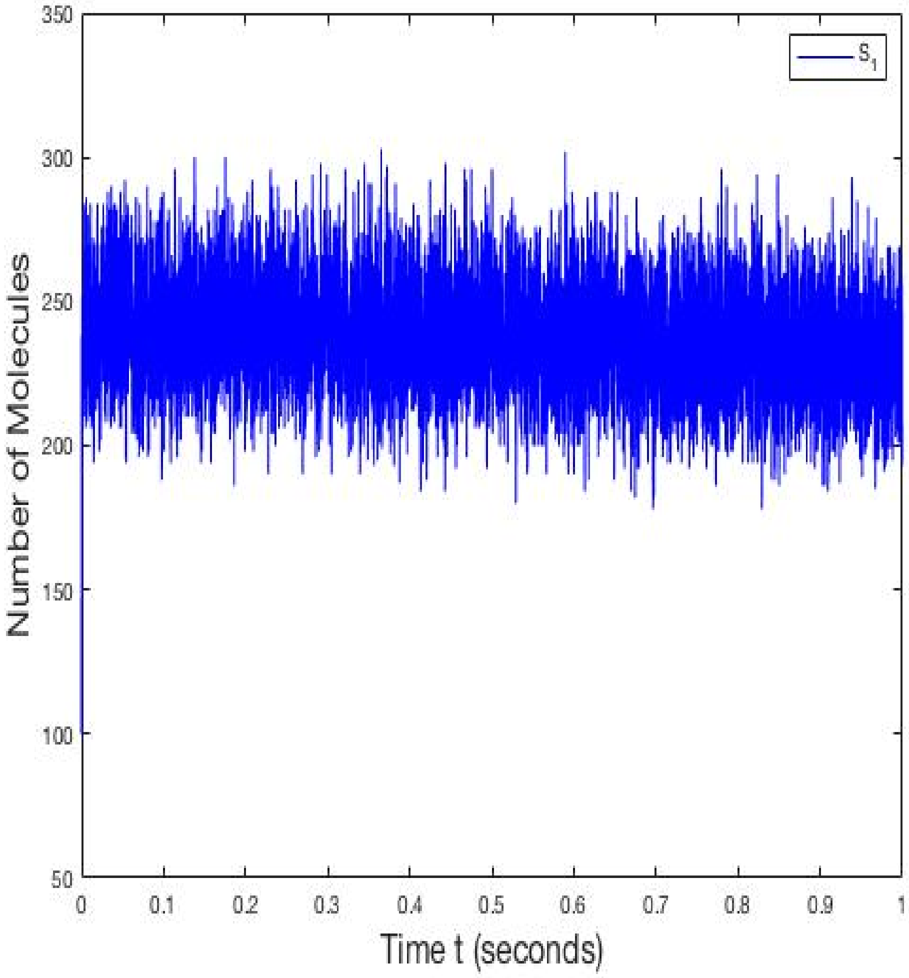

Figure 4.5: Decay-dimerization model: Evolution in time of the amount of S_1 molecules on the time interval [0,1], using the SSA

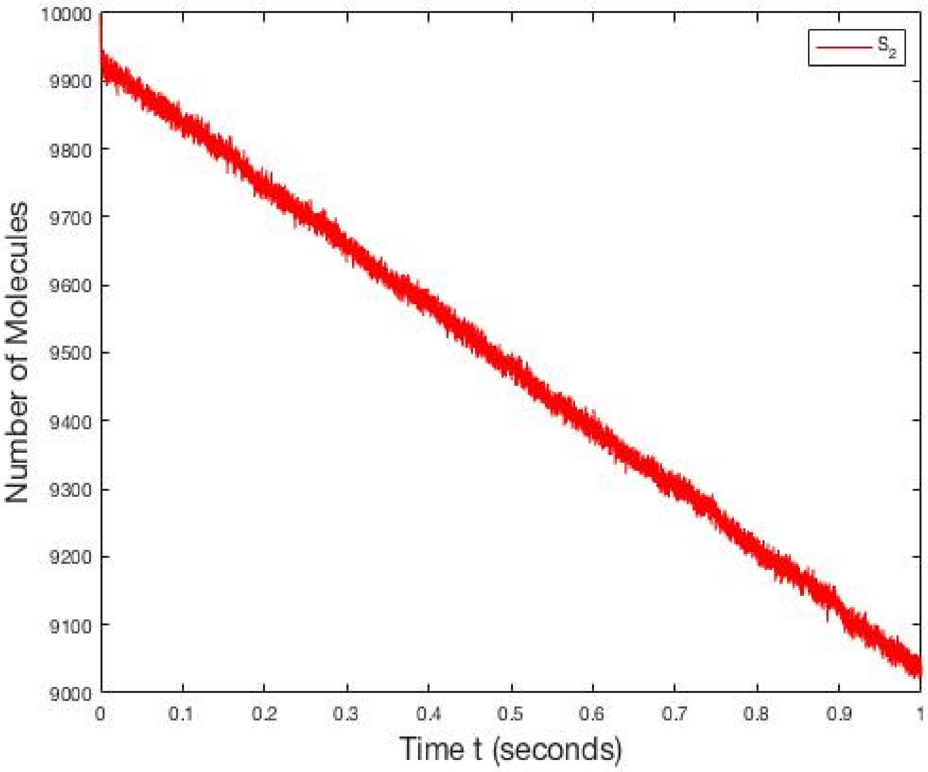

Figure 4.6: Decay-dimerization model: Evolution in time of the amount of S_2 molecules on the time interval [0,1], using the SSA

The histograms of the same three species at time $t = 1$, which are computed using 10,000 trajectories obtained with the SSA, the RTC, the tau-leaping method and the EM scheme for the CLE, are shown in Figures 4.7, 4.8 and 4.9 for the numbers of S_1, S_2 and S_3 molecules, respectively. These histograms show the very good accuracy of the RTC, tau-leaping and EM schemes, the highest accuracy being given by the RTC algorithm and the least accuracy being obtained using EM scheme . The step $\tau = \frac{1}{3} \times 10^{-4}$ was chosen very small to improve the accuracy of the tau-leaping and EM technique and maintained numerical stability of these explicit strategies applied to a stiff model.

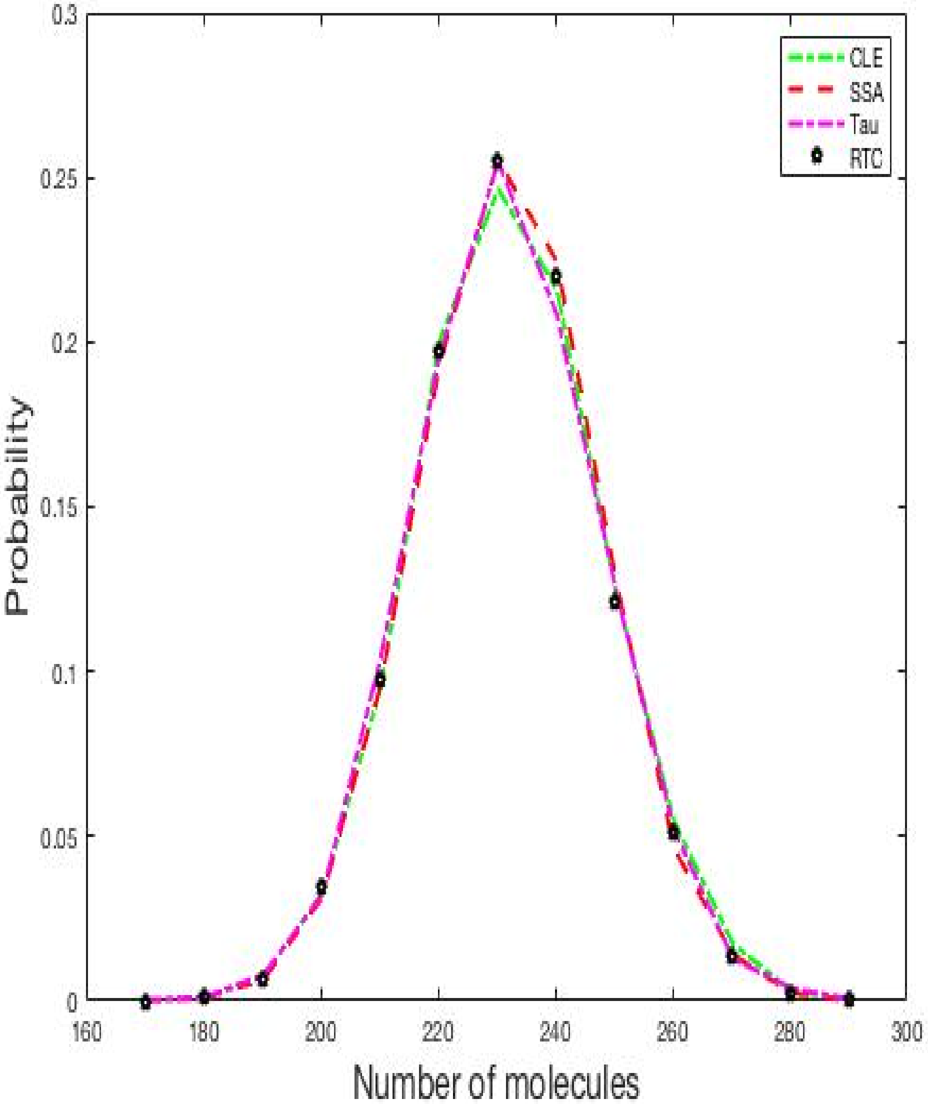

Figure 4.7: Decay-dimerization model: The histograms of the number of S_1 molecules at time $t = 1$ using the SSA, RTC, tau-leaping and EM schemes.

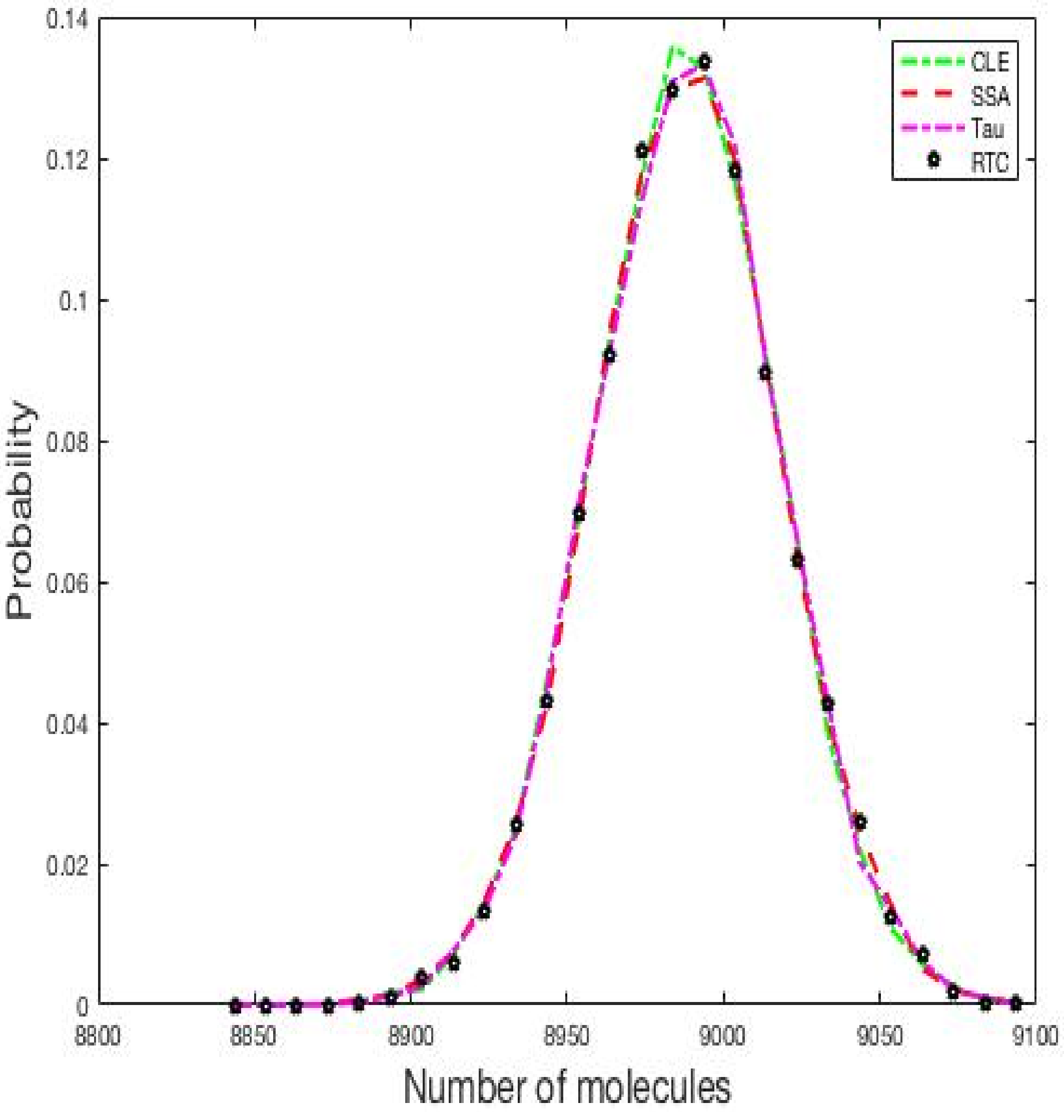

Figure 4.8: Decay-dimerization model: The histograms of the number of S_2 molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM schemes

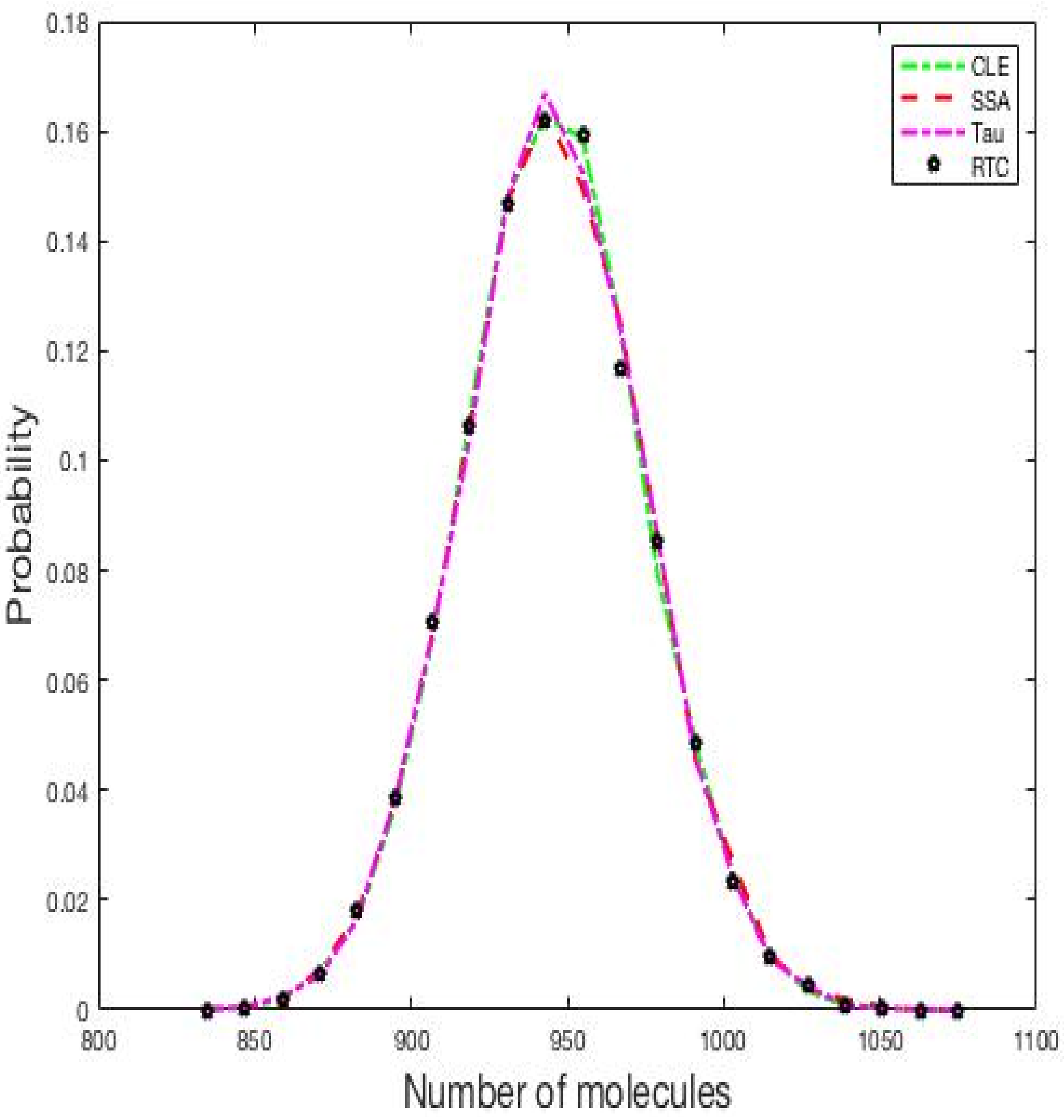

Figure 4.9: Decay-dimerization model: The histograms of the number of S_3 molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM schemes

4.3 Epidermal Growth Factor Receptor Model

A transmembrane protein called epidermal growth factor receptor (EGFR), is a receptor for members of the epidermal growth factor family. It is involved in cell proliferation and differentiation [**28, 29**]. A shortage in signalling of EGFR in humans may lead to diseases such as Alzheimer and over-expression to the development of a wide variety of tumors. The EGFR signalling pathway model involves 23 species and 47 reactions. The reactions and their rate constants are listed in Table 4.1 with rate constants:

Table 4.1: EGFR signalling pathway reactions and rate constants

Rid	**Reaction**	**Rate Constant**
1	$EGF + R \longrightarrow Ra$	3×10^{-3}
2	$Ra \longrightarrow EGF + R$	6×10^{-2}
3	$Ra + Ra \longrightarrow R2$	2×10^{-2}
4	$R2 \longrightarrow Ra + Ra$	1×10^{-1}
5	$R2 \longrightarrow RP$	1
6	$RP \longrightarrow R2$	1×10^{-2}
7	$RP \longrightarrow R2[MM]$	$4.5 \times 10^{2}, 5 \times 10^{1}$
8	$RP + PLCg \longrightarrow R\text{-}PL$	6×10^{-2}
9	$R\text{-}PL \longrightarrow RP + PLCg$	2×10^{-2}
10	$R\text{-}PL \longrightarrow R\text{-}PLP$	1
11	$R\text{-}PLP \longrightarrow R\text{-}PL$	5×10^{-2}
12	$R\text{-}PLP \longrightarrow RP + PLCgP$	3×10^{-1}
13	$RP + PLCgP \longrightarrow R\text{-}PLP$	6×10^{-2}
14	$PLCgP \longrightarrow PLCg[MM]$	$1, 1 \times 10^{2}$
15	$RP + Grb \longrightarrow R\text{-}G$	3×10^{-3}

Rid	**Reaction**	**Rate Constant**
16	$R\text{-}G \longrightarrow RP + Grb$	5×10^{-2}
17	$R\text{-}G + SOS \longrightarrow R\text{-}G\text{-}S$	10^{-2}
18	$R\text{-}G\text{-}S \longrightarrow R\text{-}G + SOS$	6×10^{-2}
19	$R\text{-}G\text{-}S \longrightarrow RP + G\text{-}S$	3×10^{-2}
20	$RP + G\text{-}S \longrightarrow R\text{-}G\text{-}S$	4.5×10^{-3}
21	$G\text{-}S \longrightarrow Grb + SOS$	1.5×10^{-3}
22	$Grb + SOS \longrightarrow G\text{-}S$	1×10^{-4}
23	$RP + Shc \longrightarrow R\text{-}Sh$	9×10^{-2}
24	$R\text{-}Sh \longrightarrow RP + Shc$	6×10^{-1}
25	$R\text{-}Sh \longrightarrow R\text{-}ShP$	6
26	$R\text{-}ShP \longrightarrow R\text{-}Sh$	6×10^{-2}
27	$R\text{-}ShP \longrightarrow RP + ShP$	3×10^{-1}
28	$RP + ShP \longrightarrow R\text{-}ShP$	9×10^{-4}
29	$ShP \longrightarrow Shc[MM]$	$1.7, 3.4 \times 10^{2}$
30	$R\text{-}ShP + Grb \longrightarrow R\text{-}Sh\text{-}G$	3×10^{-3}
31	$R\text{-}Sh\text{-}G \longrightarrow R\text{-}ShP + Grb$	1×10^{-1}
32	$R\text{-}Sh\text{-}G \longrightarrow RP + Sh\text{-}G$	3×10^{-1}
33	$RP + Sh\text{-}G \longrightarrow R\text{-}Sh\text{-}G$	9×10^{-4}
34	$R\text{-}Sh\text{-}G + SOS \longrightarrow R\text{-}Sh\text{-}G\text{-}S$	10^{-2}
35	$R\text{-}Sh\text{-}G\text{-}S \longrightarrow R\text{-}Sh\text{-}G + SOS$	2.14×10^{-2}
36	$R\text{-}Sh\text{-}G\text{-}S \longrightarrow RP + Sh\text{-}G\text{-}S$	1.2×10^{-1}

Rid	**Reaction**	**Rate Constant**
37	$RP + Sh\text{-}G\text{-}S \longrightarrow R\text{-}Sh\text{-}G\text{-}S$	2.4×10^{-4}
38	$ShP + Grb \longrightarrow Sh\text{-}G$	3×10^{-3}
39	$Sh\text{-}G \longrightarrow ShP + Grb$	1×10^{-1}
40	$Sh\text{-}G + SOS \longrightarrow Sh\text{-}G\text{-}S$	3×10^{-2}
41	$Sh\text{-}G\text{-}S \longrightarrow Sh\text{-}G + SOS$	6×10^{-2}
42	$Sh\text{-}G\text{-}S \longrightarrow G\text{-}S + ShP$	1×10^{-1}
43	$G\text{-}S + ShP \longrightarrow Sh\text{-}G\text{-}S$	2.1×10^{-2}
44	$R\text{-}ShP + G\text{-}S \longrightarrow R\text{-}Sh\text{-}G\text{-}S$	9×10^{-3}
45	$R\text{-}Sh\text{-}G\text{-}S \longrightarrow R\text{-}ShP + SG\text{-}S$	4.29×10^{-2}
46	$PLCgP \longrightarrow PLCg\text{-}I$	1
47	$PLCg\text{-}I \longrightarrow PLCgP$	3×10^{-2}

Note that first and second order reaction rate constants have units in s^{-1} and $nM^{-1}s^{-1}$, respectively. The cell volume is 3×10^{-12} liters, such that $1nM$ concentration is equal to 1800 molecules per cell. For Michaelis-Menten kinetics, denoted in Table 4.1 as $[MM]$, the rates are, $\dfrac{V_m[S_i]}{K_m + [S_i]}$ where $[S_i]$ is the concentrations of reactant species. The Michaelis-Menten type reaction constants are expressed as $V_m[1/s]$ and $K_m[nM]$. The Michaelis-Menten type reactions are R_7, R_{14} and R_{29}.

The species together with their initial values of are given in Table 4.2:

Table 4.2: EGFR signalling pathway species and their initial values

Sid	**Species**	**N**
1	EGF	23 040 183
2	R	335
3	Ra	11 774
4	$R2$	9 514
5	RP	1 360
6	R-PL	59
7	R-PLP	91
8	R-G	947
9	R-G-S	300
10	R-Sh	23
11	R-ShP	618
12	R-Sh-G	195
13	R-Sh-G-S	124
14	G-S	1 776
15	ShP	152 296
16	Sh-G	56 545
17	$PLCg$	1 195
18	$PLCgP$	2 160
19	$PLCg$-I	185 357

Sid	**Species**	**N**
20	Grb	32 547
21	Shc	2 634
22	SOS	4 689
23	Sh-G-S	52 301

The model is simulated over the time-interval $[0, 100]$, using 10,000 runs of the SSA, RTC, the *tau-leaping* scheme and the Euler-Maruyama for he CLE, respectively. Individual trajectories representing the graphs of the molecular amounts of some key species as functions as time are plotted in Figure 4.10 for species EGR, Figure 4.11 for species R, Figure 4.12 for R-G, Figure 4.13 for $PLCg$-I and Figure 4.14 for Sh-G-S.

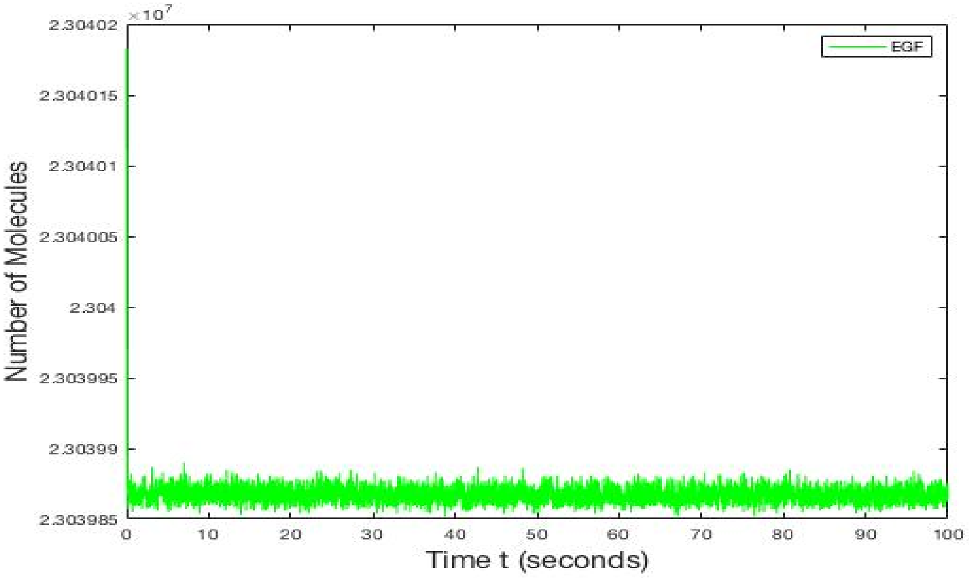

Figure 4.10: EGFR model: Evolution of the number of *EGF* molecules on the time interval [0,100] using SSA.

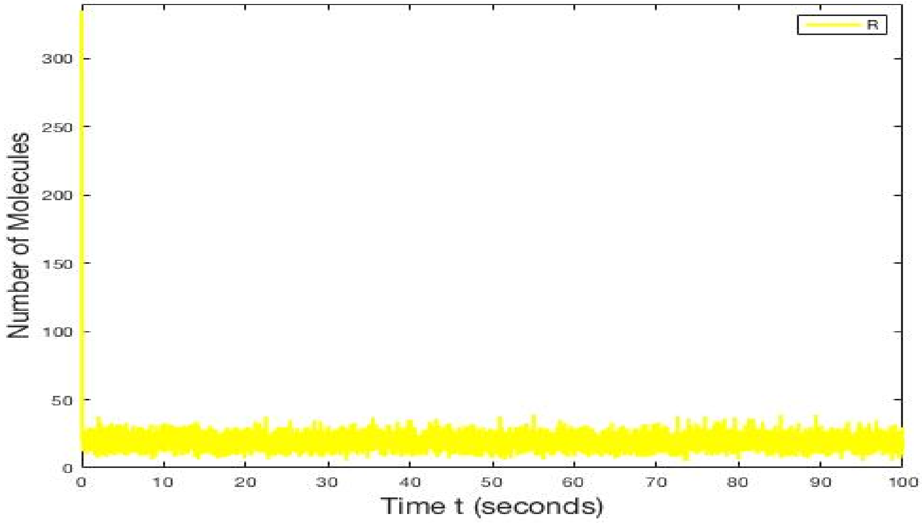

Figure 4.11: EGFR model: Evolution of the number of R molecules on the time interval [0,100] using SSA.

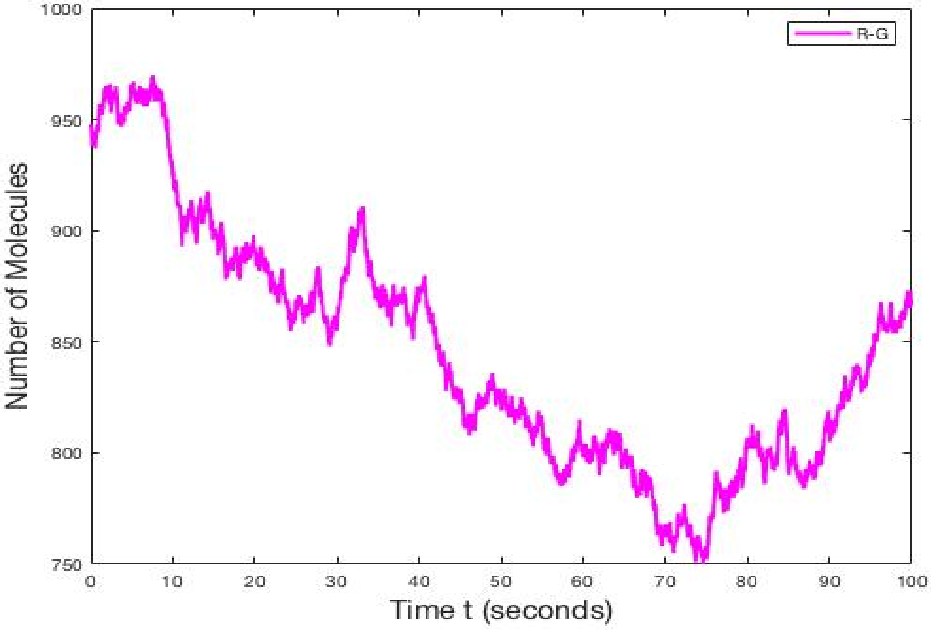

Figure 4.12: EGFR model: Evolution of the number of R-G molecules on the time interval [0,100] using SSA.

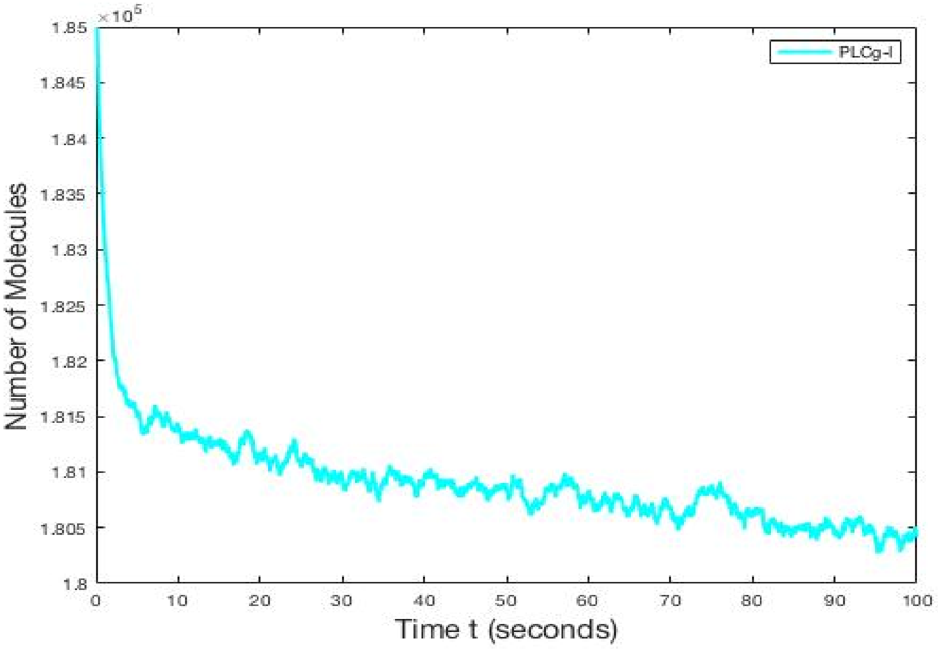

Figure 4.13: EGFR model: Evolution of the number of *PLCg-I* molecules on the time interval [0,100] using SSA.

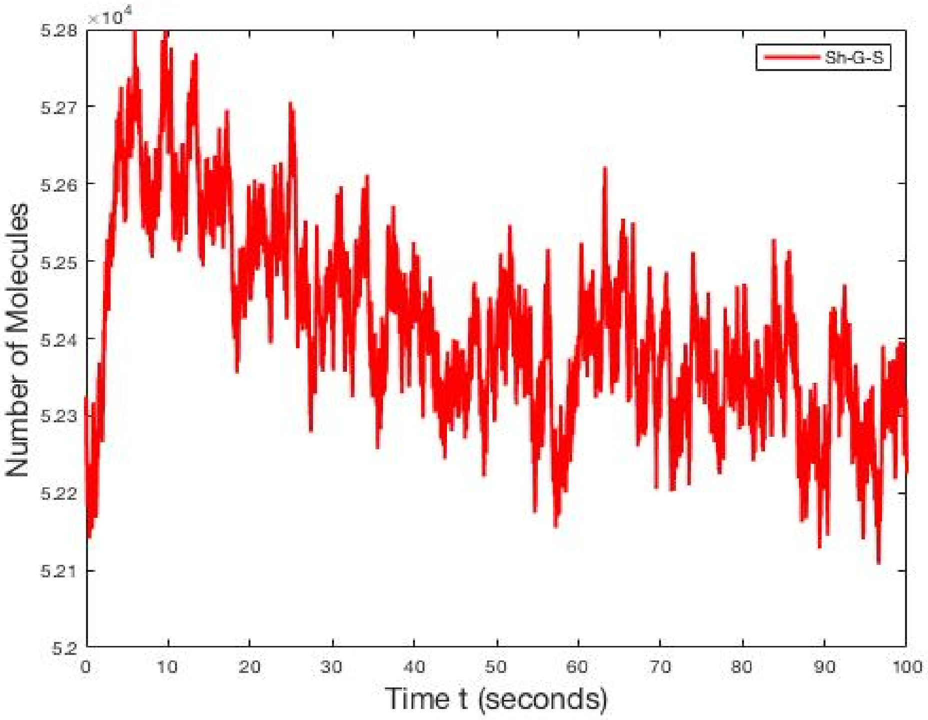

Figure 4.14: EGFR model: Evolution of the number of *Sh-G-S* molecules on the time interval [0,100] using SSA.

Finally, we compare the histograms computed over 10,000 trajectories at time $t = 1$ for the number of molecules of EGF in Figure 4.15, of R in Figure 4.16, of *R-G* in Figure 4.17, of *PLCg-I* in Figure 4.18 and of *Sh-G-S* in figure **??** . The accuracy of RTC, the tau-leaping algorithm and the EM scheme for the CLE are computed, considering the exact SSA as reference. All methods have excellent accuracy. The approximate tau-leaping method and the Euler-Maruyama scheme employed a very small stepsize, $\tau = 10^{-4}$, thus the high accuracy of these methods.The RTC produces the histogram closest to that generated utiliting the SSA.

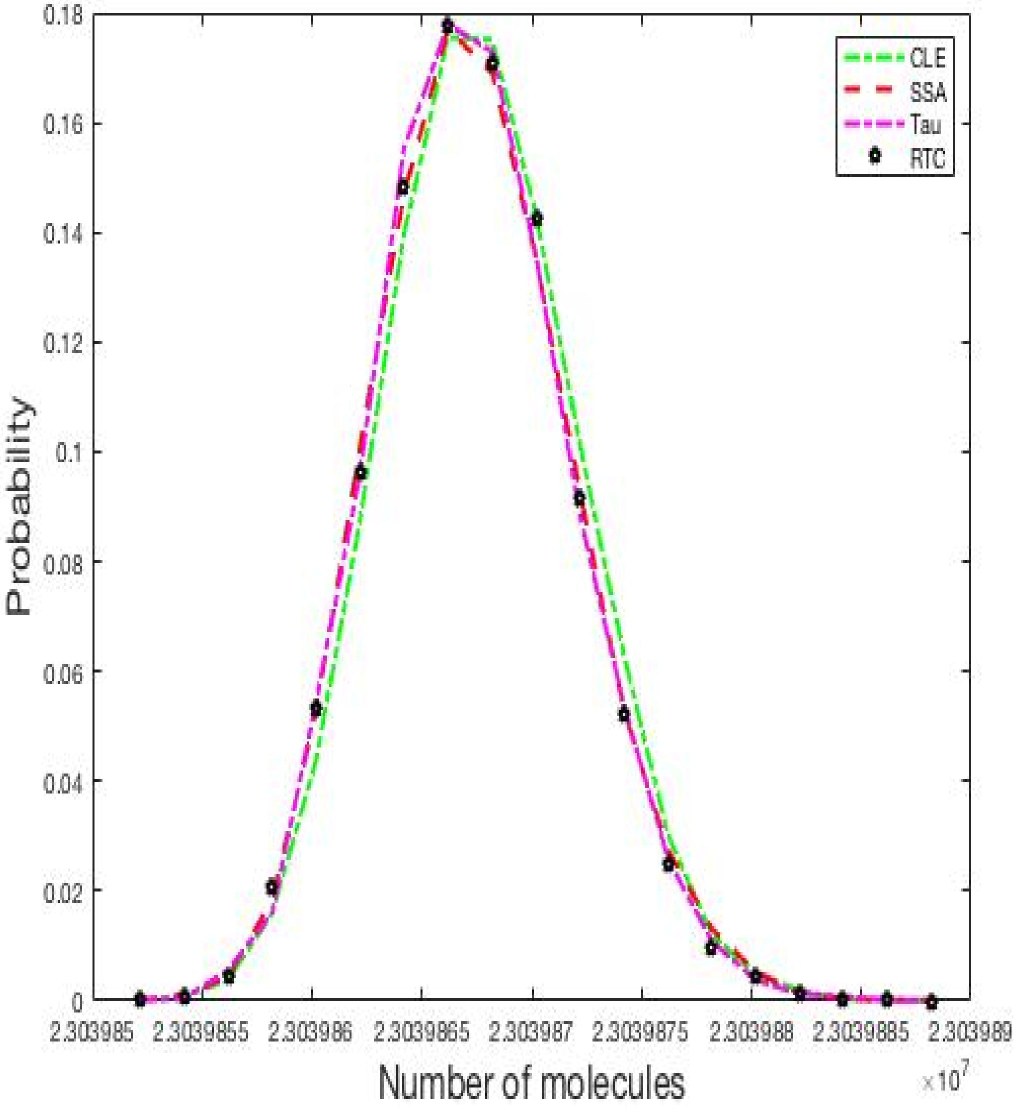

Figure 4.15: EGFR model: The histograms of the number of EGF molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM scheme

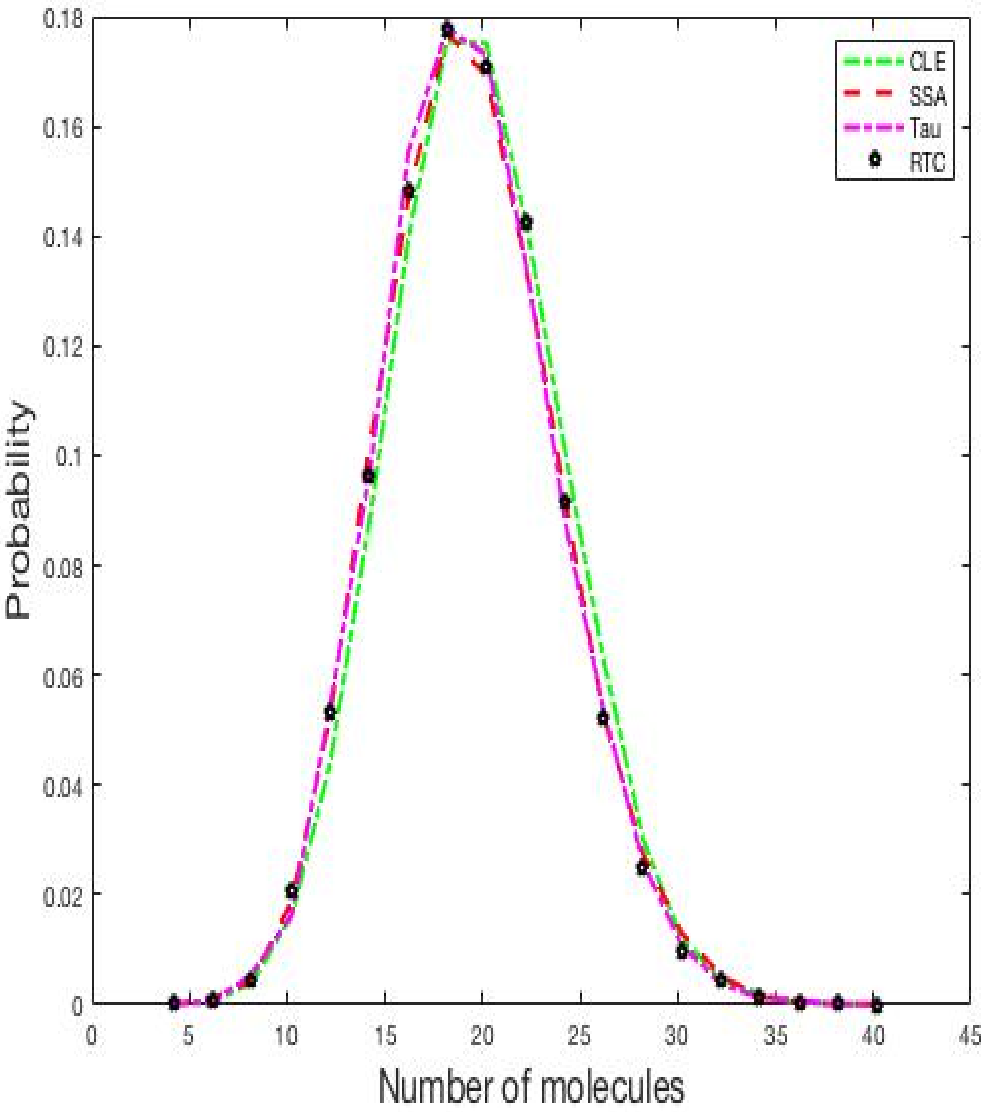

Figure 4.16: EGFR model: The histograms of the number of R molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM scheme

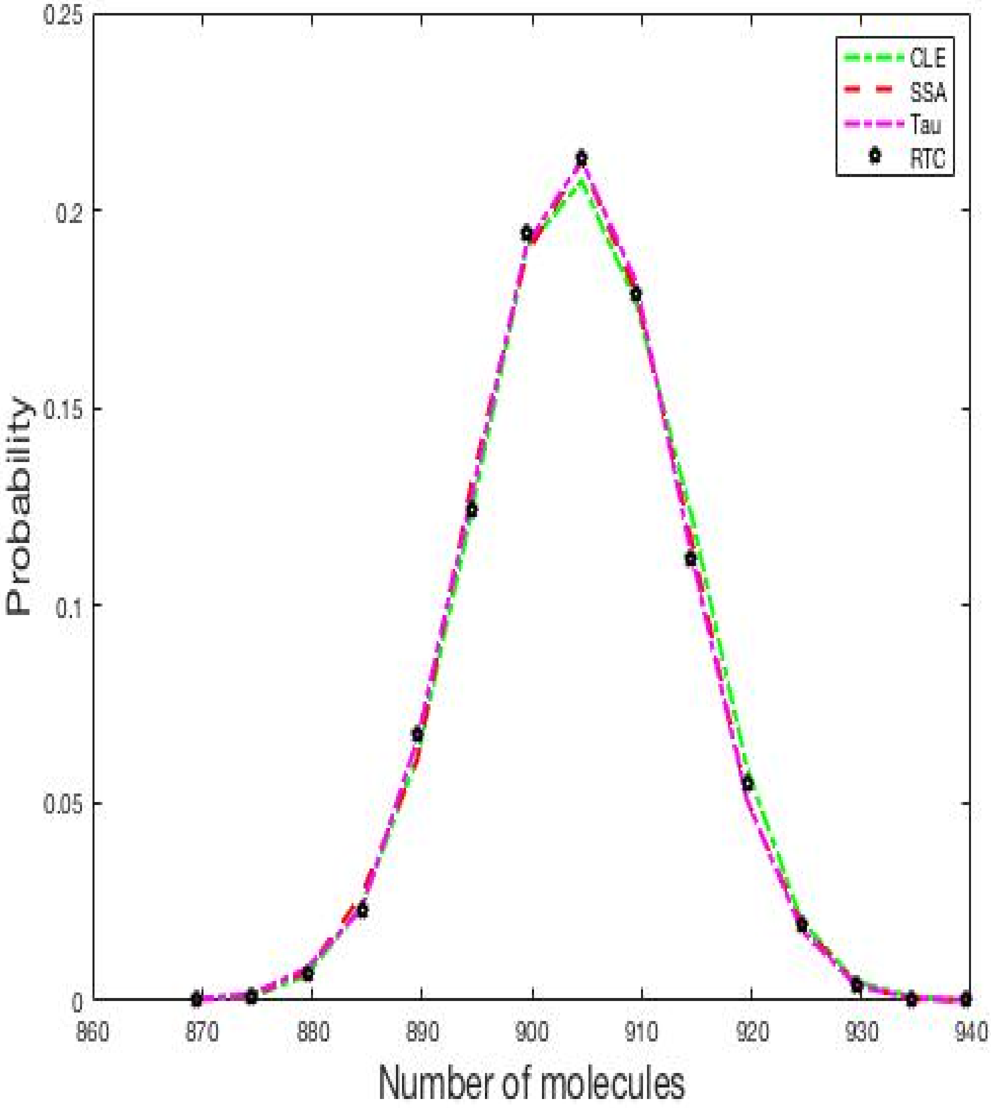

Figure 4.17: EGFR model: The histograms of the number of R-G molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM scheme

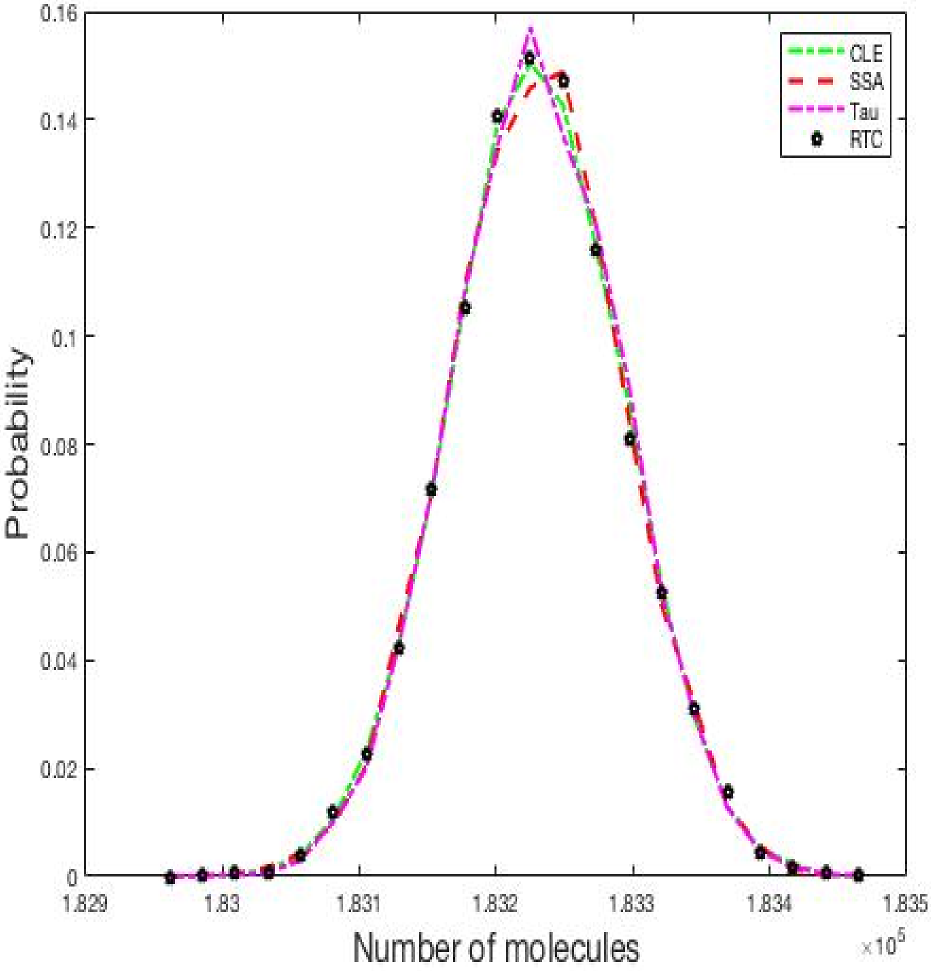

Figure 4.18: EGFR model: The histograms of the number of $PLCg\text{-}I$ molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM scheme

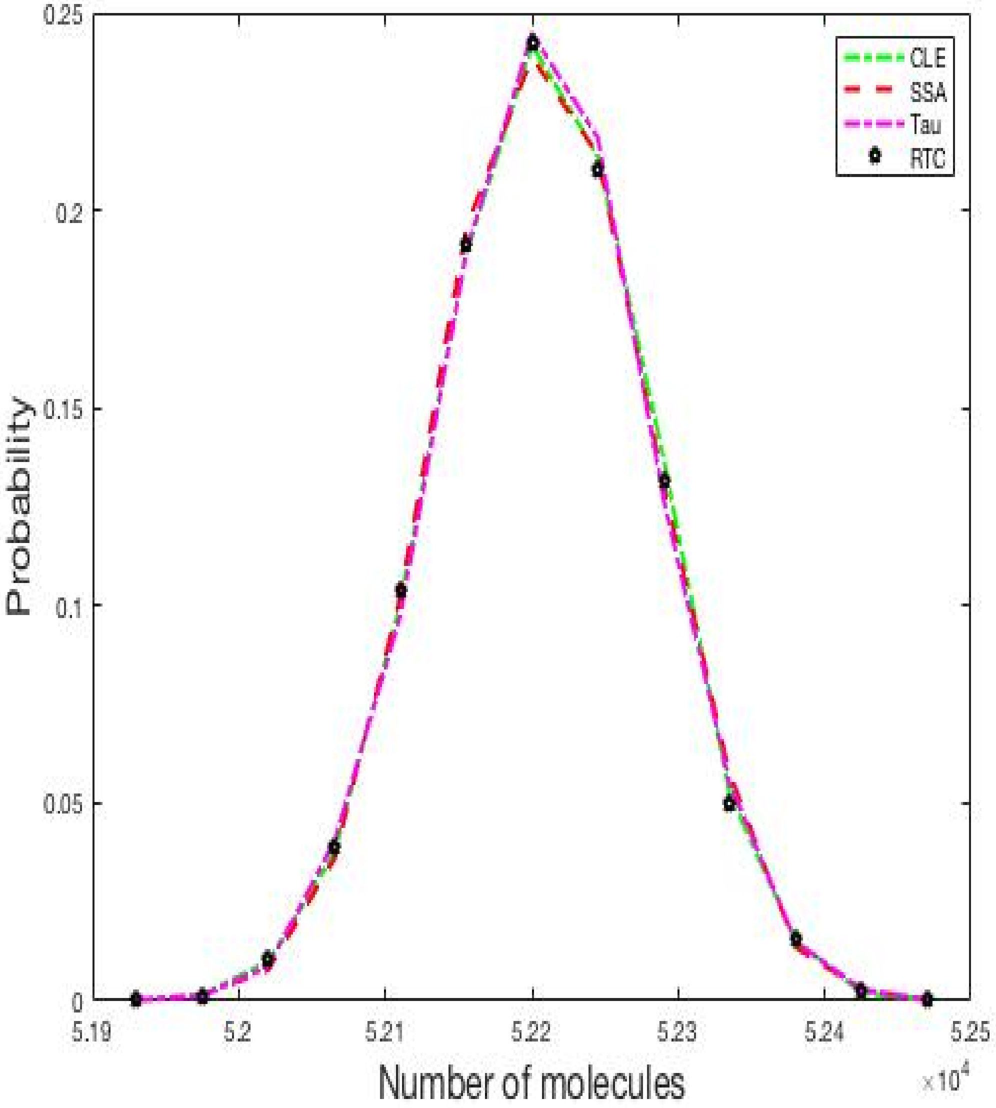

Figure 4.19: EGFR model: The histograms of the number of *Sh*-*G*-*S* molecules at time $t = 1$, using the SSA, RTC, tau-leaping and EM scheme

Chapter 5

Conclusions

Stochastic models are necessary to accurately represent the dynamics of biochemical networks that exhibit randomness. This randomness is often due to the low number of molecules of some reacting species, as is the case for species such as DNA or RNA. One of the most popular stochastic models of well-stirred biochemical networks is the Chemical Master Equation. This is a discrete stochastic model which is often of very high dimension and therefore quite challenging to solve directly. Under certain conditions, the Chemical Master Equation may be reduced to a stochastic continuous model of much lower dimension, namely the Chemical Langevin Equation. In the regime of very large molecular population numbers, the Chemical Langevin Equation reduces to the well-known model of chemical kinetics, that of the (deterministic) reaction rate equations.

The focus of this thesis was on the numerical solution of the stochastic models of homogeneous biochemical networks. Gillespie's algorithm or the

Stochastic Simulation Algorithm is an exact Monte Carlo method for computing the solution of the Chemical Master Equation. This algorithm is widely used for solving numerically stochastic discrete models of well-stirred biochemical kinetics.

However, Gillespie's algorithm is not the best tool for studies, such as sensitivity analysis of the Chemical Master Equation. In this case, the sensitivity estimator based on finite difference approximations using the SSA may give inaccurate results. An exact Monte Carlo method for the Chemical Master Equation that performs better than the SSA, when sensitivity analysis is of interest, is based on the Random Time Change representation of the Markov process modelled by the CME. This representation, due to Kurtz, constitutes the theoretical basis of the exact Random Time Change algorithm.

This thesis studied the Random Time Change algorithm, or the RTC which while less analyzed in the literature than the SSA, provides a powerful tool for sensitivity analysis of stochastic models of biochemical kinetics, when finite-difference approximations are employed. The sensitivity estimator based on the RTC is much more accurate than that based on the SSA.

We studied the advantages of the RTC by comparing its performance with that of SSA and the (approximate) tau-leaping method for the Chemical Master Equation. The algorithms were tested on three biologically relevant models, including a quite complex model of the epidermal growth factor receptor (EGFR) pathway. The RTC was shown to be as accurate as the SSA, and more accurate than the approximate algorithms such as the tau-leaping method or the Euler-Maruyama scheme for the Chemical Langevin Equation. The efficiency of the RTC is also studied.

In conclusion, the RTC algorithm is a promising alternative to the popular stochastic simulation algorithm due to Gillespie. It is particularly useful as an underlying method, for sensitivity analysis, and as a consequence, for identifiability analysis for this model.

In our future work, we plan to study other applications of the RTC representation and algorithm, such as to model reduction of well-stirred biochemical networks.

Bibliography

[1] K. Burrage, T. Tian, P. Burrage, (2004), A multi-scaled approach for simulating chemical reaction systems, *Progress in Biophysics and Molecular Biology*, 85.(2-3),217-234

[2] M. Kerker, (1974), Brownian movements and molecular reality prior to 1900, *J. Chem. Educ.*, 51, 764-768.

[3] E. Farber, (1961), Early Studies concerning time in chemical reactions, *Chymia*, 7, 135-148.

[4] I. R. Epstein, J. A. Pojman, (1998). *An Introduction to Nonlinear Chemical Dynamics: Oscillations, Waves, Patterns, and Chaos,* Oxford University Press, Oxford.

[5] R. Heinrich, S. Schuster, (1996), *The Regulation of Cellular Systems,* Chapman and Hall, New York.

[6] B. G. Cox, (1994), *Modern Liquid Phase Kinetics*, Oxford University Press, Oxford.

[7] M. Delbruck, (1940), Statistical fluctuation in autocatalytic reactions, *J. Phys. Chem.*, 8(1), 120.

[8] H. McAdams and A. Arkin, (1997), Stochastic mechanism in gene expression, *Proceedings of the National Academy of Sciences of the USA*, 94:814- 819.

[9] A. Arkin, J. Ross, and H. McAdams,(1998), Stochstic kinetics analysis of developmental pathway bifurcation in Phage λ-infected Escherichia Coli cell, *Genetics*, 149, 1633.

[10] D. E. Quelle, F. Zindy, R. A. Ashmun, C. J. Sherr, (1995), Alternative reading frames of the INK4a tumor suppressor gene encode two unrelated proteins capable of inducing cell cycle arrest, *Cell*, 83, 993-1000.

[11] T. Marquardt, (2001), Pax6 is required for the multipotent state of retinal progenitor cells, *Cell*, 105, 43-55.

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

[13] D. T. Gillespie, (1977), Exact stochastic simulation of coupled chemical reactions, *J. Phys. Chem.* 81, 2340-2361.

[14] D. T. Gillespie, (2001), Approximate accelerated stochastic simulation of chemically reacting systems, *J. Chem. Phys.*, 115, 1716-33.

[15] D. T. Gillespie, (2000), The Chemical Langevin equation, *J. Chem. Phys.*, 113, 297-306.

[16] D. J. Wilkinson, (2006), *Stochastic modelling for systems biology*, Chapman and Hall / CRC.

[17] T. G. Kurtz, (1972), The relationship between stochastic and deterministic models for chemical reactions, *J. Chem. Phys.*, 57(7) 2976-2978.

[18] M. Rathinam, P.W. Sheppard and M. Khammash, (2010), Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks, *J. Chem. Phys.*, 132, 034103.

[19] Y. Cao, D. T. Gillespie, L. Petzold, (2005), The slow-scale stochastic simulation algorithm, *J. Chem. Phys.*, 122, 014116

[20] Y. Cao, D. T. Gillespie, L. Petzold, (2006), Efficient step-size selection for the tau-leaping method, *J. Chem. Phys.*, 124, 044109.

[21] Y. Cao, D. T. Gillespie, L. Petzold, (2007), Adaptive explicit-implicit for the tau-leaping with automatic tau-selection, *J. Chem. Phys.* 126, 224101.

[22] D. T. Gillespie, (1992), A rigorous derivation of the chemical master equation, *Physica, A.*, 188, 402-425.

[23] M. Thattai and A. Van Oudenaarden, (2001), Intrinsic noise in gene regulatory networks, *Proc. Natl. Acad. Sci.* U.S.A. 98, 8614.

[24] M.A. Gibson, J. Bruck, (2000), Efficient exact stochastic simulation of chemical systems with many species and many reactions, *J. Chem. Phys.*, 104, 1876-1889.

[25] S.N. Either and T.G. Kurtz, (1986) *Markov Processes: Characterization and Convergence* , Wiley, New York.

[26] S. Ilie, W. H. Enright, K. R. Jackson, (2009), Numerical solution of stochastic models of biochemical kinetics, *Canadian Applied Mathematics Quarterly*, 17(3)523 - 554.

[27] Y. Pu, L. T. Watson, Y. Cao, (2011), Stiffness detection and reduction in discrete stochastic simulation of biochemical systems, *J. Chem. Phys.*, 134, 054105.

[28] B. N. Kholodenko, A. van. Demin, G. Moehren, and J. B. Hoek, (1999), Quantification of short term signalling by the epidermal growth factor receptor, *J. Biol. Chem.*, 274, 30169- 30181.

[29] H. Resat, J. A. Ewald, D. A. Dixon and H. S. Wiley, (2003), An integrated model of epidermal growth factor receptor trafficking and signal transduction, *Biophys. J.*, 85, 730-743.

[30] M. A. Gibson, J. Bruck, (2000), Exact stochastic simulation of chemical systems with many species and many channels, *J. Chem. Phys.*, 105, 1876-1889.

[31] D. F. Anderson, (2012), An efficient finite difference method for parametric sensitivities of continuous time Markov chains, *SIAM J. Numer. Anal.*, 5015, 2237-2258.

[32] T.G. Kurtz, (1978), Strong approximation theorems for density dependent Markov chains, *Stoch. Proc. Appl.*, 6(3), 223-240.