Development of Kinetic Monte Carlo Methods for the Stochastic Simulation of Chemical Systems

Thesis Submitted to AcSIR for the Award of the Degree of

DOCTOR OF PHILOSOPHY

in Physical Sciences

By

Shantanu Kadam

Registration Number:10PP13J26046

Dr. Kumar Vanka
(Research Guide)

Physical Chemistry Division
National Chemical Laboratory
Pune 411008

September 2015

सीएसआयआर–राष्ट्रीय रासायनिक प्रयोगशाला
(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)
डॉ. होमी भाभा मार्ग, पुणे – 411 008. भारत
CSIR-NATIONAL CHEMICAL LABORATORY
(Council of Scientific & Industrial Research)
Dr. Homi Bhabha Road, Pune - 411008. India

Certificate

This is to certify that the work incorporated in this Ph.D. thesis entitled **"*Development of Kinetic Monte Carlo Methods for the Stochastic Simulation of Chemical Systems*"** submitted by ***Mr. Shantanu Kadam*** to Academy of Scientific and Innovative Research (AcSIR) in fulfillment of the requirements for the award of the Degree of ***Doctor of Philosophy in Physical Sciences***, embodies original research work under my supervision. I further certify that this work has not been submitted to any other University or Institution in part or full for the award of any degree or diploma. Research material obtained from other sources has been duly acknowledged in the thesis. Any text, illustration, table etc., used in the thesis from other sources, have been duly cited and acknowledged.

Mr. Shantanu Kadam
(Student)

Dr. Kumar Vanka
(Supervisor)

Date: 15th Sep' 2015

Place: Pune

Communications Channels
NCL Level DID : 2590
NCL Board No. : +91-20-25902000
Four PRI Lines : +91-20-25902000
FAX
Director's Office : +91-20-25902601
COA's Office : +91-20-25902660
SPO's Office : +91 20 25902664
WEBSITE
www.ncl-india.org

To My Beloved Father

Acknowledgements

I am glad to present my research work in the form of this thesis. At the same time, this occasion gives me an opportunity to express my heartfelt thankfulness to all the people who have helped me to come this far.

First of all, I take this befitting opportunity to convey my deep gratitude to my thesis supervisor Dr. Kumar Vanka. He introduced me to the exciting world of Monte Carlo simulations. His enduring guidance, valuable advice and critical remarks have helped me to metamorphose into a researcher. His ways of imagining and decoding the complex things and zeal for research rubs off easily. He has always encouraged me by giving a free hand to dauntlessly pursue the things which interest me the most.

I also want to offer my appreciation and gratitude to Dr. Chetan Gadgil for his thought-provoking questions followed by invaluable suggestions during my DAC meetings. The feedback from other DAC members: Dr. Ram Rup Sarkar and Dr. Sarika Bhattacharyya have also helped me to improve my research. I thank all my DAC members for their timely advice, which has lead to the improvement of the quality of my thesis work.

During the course-work, I had a great privilege to attend the masterful lectures on quantum mechanics delivered by Dr. Sourav Pal. His razor-sharp analysis, and clear cut explanation of concepts has provided an aesthetic contour to the subject. In addition, I extend my heartiest appreciation to Dr. Nayana Vaval for her tips on writing constructive programs, which were immensely beneficial in subsequent years. My gratitude to her is ever-lasting. The course-work classes conducted by Dr. Neelanjana Sengupta, Dr. Kavita Joshi and Dr. Durba Sengupta have helped me to understand their respective subjects better. I am thankful to Prof. Anjali Kshirsagar for her guidance during my university days.

The infrastructural facilities provided for research by Dr. Sourav Pal (Former Director, CSIR - NCL) and Dr. Anil Kumar (Head, Physical and Materials Chemistry Division) are highly appreciated. I am grateful to Reliance Industries, DST, CSIR for

funding my research and to Academy of Scientific and Innovative Research (AcSIR) for getting me registered for the doctoral degree. I thank SAO chairman, Deepak Jori and all administrative staff for their help.

I am obliged to my jolly group members and friends Manoj, Nisha, Amrita, Yuvraj, Jugal and Mritunjay for keeping a cheerful atmosphere at my workplace. They have helped me, supported me, and stood by me in every possible way. The mid-afternoon chitter-chatter, leg-pulling and debates on topics ranging from current affairs to movies to cricket will be remembered for a long time.

Over the years, the theory and computational group has continuosly evolved with the arrivals and departures of many research fellows. Indeed it is a Herculean task to thank each one of them individually, though I have had shared a cordial bond with each of them. My special thanks are to the Drs. Sumantra, Deepti, Jitendra, Mudit, Prithvi, Achintya, Arrya, Susanta, Baljinder, Vikas, Yogesh, Atul and others who are in queue to be Drs. Amol, Avinash, Sayali, Debarati, Kamalika, Anagha, Himadri, Manzoor, Jaya, Prathit, Chandan, Souvik, Swagata, Anil, Asis, Sudip and Deepak. I also thank the junior members Sneha, Samik, Rahul, Tamal and Subhrashis. I wish all of them the very best in their future endeavors. A very very special thanks to my roommates and great companions: the jovial Yuvraj and tech-savvy Turbasu. The late night gossips, delicious weekend dinners and all that fun will be definitely cherished for many more years. I would also like to thank my teachers and relatives for their unwavering support. I am grateful to Mrs. Alandkar for her words of encouragement during the inceptive stage of my academic career.

I have been blessed with an extremely supportive and lovable family. My parents (Aai and Papa) made enormous sacrifices to put my career ahead of anything else. Many thanks to Tai and Koya for always putting up with my faults. Both of you are gems as a people. Thanks to Aaji (up and above) for always being there.

Shantanu

Contents

List of Tables

Abstract

In the traditional approach of chemical kinetics, the reaction network is modeled by using a set of ordinary differential equations (ODEs). This set of equations are termed as reaction rate equations (RREs), that are solved subject to given initial conditions. In this approach, the time evolution of the concentrations of constituent species is treated continuously. However, this methodology is found to be valid in the thermodynamic limit. For example, the cellular systems cannot be simulated correctly in the continuum limit due to a smaller number of molecules. Hence, these systems need to be simulated probabilistically by methods that take into account the discrete and random nature of biochemical interactions. The application of kinetic Monte Carlo based stochastic simulation algorithms (SSAs) has gained wide popularity for the simulations of such kind of biochemical systems. However, their drawback is the requirement of huge computational time in performing the simulations.

The problem of excess computational load has been solved by the development of accelerated stochastic simulation algorithms. Unlike the SSAs, these techniques take a large enough leap in time executing several possible reactions. The reaction numbers are modeled by using the Poisson distribution. Further, various attempts have been made to solve the problem of negative molecular numbers during the simulations by proposing new computational methods.

In this thesis, a new computational method for the stochastic simulation of chemical systems has been developed. The newly developed accelerated stochastic simulation method successfully reduces the computational load while maintaining the accuracy of the simulations. The first chapter of this thesis is an introductory chapter. The second chapter discusses different stochastic simulation methodologies. The third chapter deals with the newly developed simulation methodology. The next two chapters are devoted to solving the issue of negative populations, followed by a separate chapter on the accelerated computational method that selects an error control parameter in a logical manner.

The thesis is organized as follows :

Chapter 1: The first chapter of the thesis begins by introducing the notion of chemical reactions followed by their types and corresponding rate equations. A brief discussion about examples of chemical reactions is also provided. Further two different simulation approaches (deterministic and stochastic) to study chemical kinetics have been discussed. The results of deterministic approach to the Lotka-Voltera model shows the necessity of insight through stochastic simulations to get a realistic picture of the behavior of chemical systems.

Chapter 2: Chapter 2 begins with a brief history of Monte Carlo methods followed by the areas of application of the kinetic Monte Carlo method. The rest of the chapter discusses the development of stochastic simulation methods over the years. The tool of kinetic Monte Carlo was launched in the form of SSA as a standard tool for studying chemical kinetics by Daniel Gillespie. In order to overcome the drawback of the computational load during the simulations, several accelerated algorithms (tau leaping methods) were proposed. The hybrid methods combine the deterministic methods along with the stochastic methods. Apart from these, numerical methods analytical techniques which deal with the chemical master equation (CME) in calculating the moments of probability distributions were used.

Chapter 3: This chapter is based on the development of a new computational method which has been named as the representative reaction approach (RRA). In this method, the entire reaction network to be simulated is represented by a single representative reaction (RR). By applying a leap condition to this RR, a large enough time step is derived. The RRA method has been successfully applied for the simulation of several chemical systems. The reported examples show that the first and second moments have been correctly reproduced. The performance of the RRA method in terms of CPU time has also been compared with other state-of-the-art methods.

Chapter 4: It has been found that the simulations of certain chemical systems by accelerated methods is prone to negative molecular numbers. Over the years, a number

of methods have attempted to solve this problem. In this chapter, a new method has been proposed which uses the SSA and the binomial distribution in conjunction with the RRA method. Like other proposed methods, this new approach successfully solves the issue for the given examples. The new methodology is also found to be appealing on the front of the CPU time needed to complete the simulation.

Chapter 5: Chapter 5 deals with the problem addressed in the previous chapter, but with a totally different perspective. Here, a novel idea of noise is used to solve the issue of negative molecular numbers. The failure of the leap condition leads to the firing of excess numbers of reactions, which, in turn, gives rise to negative numbers. This excess part in the occurrence of reaction numbers is treated as noise. It has been found that the removal of this noise part generates correct trajectories of chemical species.

Chapter 6: All the approximate accelerated stochastic simulation methods make use of an arbitrarily chosen error control parameter for the choice of the time step. The choice of the parameter for simulating one particular chemical system may not be a good choice for some other system. Here, a scheme has been developed to choose this parameter in a logical way. This is accomplished by combining the RRA method with the coupled harmonic oscillator (CHO).

Abbreviations

ODEs	ordinary differential equations
PDEs	partial differential equations
SDEs	stochastic differential equations
RREs	reaction rate equations
CME	chemical master equation
kMC	kinetic Monte Carlo
SSA	stochastic simulation algorithm
RR	representative reaction
RRA	representative reaction approach
NRM	next reaction method
DSSA	delay stochastic simulation algorithm
RDM	recycling direct method
GASA	Gillespie's approximate stochastic algorithm
GP	Gillespie-Petzold
CGP	Cao-Gillespie-Petzold
ssSSA	slow-scale stochastic simulation algorithm
LNA	linear noise approximation
PGF	probability generating function
BD-τ	binomial distribution based τ
CV	coefficient of variation
CHO	coupled harmonic oscillator
CPU	central processing unit

List of Publications

1. **Shantanu Kadam** and Kumar Vanka, A new approximate method for the stochastic simulation of chemical systems: The representative reaction approach, *Journal of Computational Chemistry*, 2012, 33, 276

2. **Shantanu Kadam** and Kumar Vanka, Solving the problem of negative populations in approximate accelerated stochastic simulations using the representative reaction approach, *Journal of Computational Chemistry*, 2013, 34, 394

3. **Shantanu Kadam** and Kumar Vanka, The accounting of noise to solve the problem of negative populations in approximate accelerated stochastic simulations, *RSC Advances*, 2014, 4, 58127

4. Nishamole Kuriakose, **Shantanu Kadam** and Kumar Vanka, A Theoretical study of metal-metal cooperativity in the homogeneous water gas shift reaction, *Inorganic Chemistry*, 2011, 51, 377

CHAPTER 1

Introduction

The true logic of this World is in the calculus of probabilities.

-James Clerk Maxwell

1.1 Chemical Reactions

Chemical reactions are processes in which the rearrangement of atoms among reacting molecules leads to the formation of new molecules with different properties. The reacting molecules in a chemical reaction are called reactants, while those getting formed are called products. They can also be interpreted as processes where bonds between atoms in the reactants are broken, and the new molecules are formed as products by forming new bonds. The occurrence of these chemical reactions can be explained on the basis of collision theory. The collisions between the reacting molecules gives rise to reactive collisions and non-reactive collisions. The reactive molecular collisions have enough energy, known as the activation energy, to cross the potential barrier to form the products. The number of reactive collisions can be increased either by increasing the reactant concentration or by increasing the temperature. From recent findings,[1] quantum tunneling is also seen as an another alternative for the occurrence of chemical reactions.

Chemical reactions have been taking place not just in the laboratory, but within us and our surroundings since time immemorial. A few examples of such chemical reactions are photosynthesis, combustion, rusting of iron, digestion of food in our body, and aerobic and anaerobic cellular respiration. Some processes that involve chemical reactions like the making of curd, the souring of milk, the washing of cloths using detergents and striking a match have been associated with our day today lives. The synthesis of ammonia, the water gas shift reaction and the thermite reaction are some examples of industrially important reactions. In the semiconductor industry, chips are manufactured by exposing a photoresist material to ultraviolet light. These chips are further used in electronic devices. On the front of biology, the reactions in a gene regulatory network and metabolic pathways are essential for our cellular activities.

In most of the chemical systems, the transformations of constituent substances (by chemical reactions) are accompanied by their spreading-out process called as diffusion.

These reaction-diffusion systems are also found in physics, biology and ecology. The solutions of the reaction-diffusion equations gives various self-regulated patterns. The formation of such patterns was proposed by Alan Turing[2] as a result of interaction of two chemicals (morphogens) diffusing at different rates. The formation of stripes on tigers and zebras and the spots on the coats of leopards can be explained by using the reaction-diffusion equations.[3,4] He also predicted the notion of oscillating chemical reactions, which are a paradigm for non-equilibrium thermodynamics. Thus, the chemical reactions and their usage have spanned nearly all spheres of our modern civilization.

1.2 Chemical Kinetics

The work on the formulation of law of mass action by Peter Waage, Cato Guldberg and later, independently, by van't Hoff contributed to the development of chemical kinetics. The field of chemical kinetics deals with the study of rates of chemical reactions. The word kinetics is taken from the Greek word 'kinesis'meaning movement.
Consider a chemical reaction[5,6]

$$aA + bB \xrightarrow{k} cC + dD \tag{1.1}$$

In this reaction, *a* molecules of *A* and *b* molecules of *B* react with each other to produce *c* molecules of *C* and *d* molecules of *D*. The lower case letters *a, b, c, d* are the stoichiometric coefficients, while the upper case letters (*A* and *B*) are reactants and (*C* and *D*) products. All the reactions of type (1.1) reported in this thesis are elementary reactions, one that take place in a single step. For such reactions, the rate of a reaction is given by:

$$r = -\frac{1}{a}\frac{d[A]}{dt} = -\frac{1}{b}\frac{d[B]}{dt} = \frac{1}{c}\frac{d[C]}{dt} = \frac{1}{d}\frac{d[D]}{dt} \tag{1.2}$$

where [X] denotes the concentration of substance X.

The rate equation or rate law of a chemical reaction is an expression that relates the rate of a reaction to the concentration of each reactant and the reaction rate constant. For

a chemical reaction like (1.1) the rate equation is :

$$r = k\,[A]^a\,[B]^b \tag{1.3}$$

where k is the reaction rate constant. The sum of powers, $a + b$ of the concentrations of the reactants in the Eq.(1.3) is called as the order of that chemical reaction. The order of a reaction can be zero, first, second, third and even a fraction.

Zero order reaction :

$$\phi \xrightarrow{k_1} A$$

In such type of reactions, the rate of a reaction is independent of the concentration of the reactant. Thus, changing the concentration has no effect on the reaction rate. Its rate equation is :

$$r = k_1 \tag{1.4}$$

The units of the reaction rate constant k_1 are mole $litre^{-1}\ sec^{-1}$.
Some of the enzyme-catalysed reactions and reactions that occur on metal surfaces at high pressure are zero order reactions.

First order reaction :

$$A \xrightarrow{k_2} B$$

The rate of this reaction depends on the concentration of only one reactant. The rate equation is :

$$r = -\frac{d[A]}{dt} = k_2[A] \tag{1.5}$$

The units of the reaction rate constant k_2 are sec^{-1}.
The radioactive decay of unstable nuclei follows first order kinetics.

Second order reaction :

$$A + B \xrightarrow{k_3} C$$

For this reaction, the rate equation depends on the product of the concentrations of two first order reactants. The rate equation is :

$$r = -\frac{d[A]}{dt} = k_3[A][B] \tag{1.6}$$

In case both the reactants in Eq.(1.6) are of the same kind, the rate equation depends on the concentration square of that reactant. Thus, the rate equation becomes :

$$r = -\frac{d[A]}{dt} = k_3[A]^2 \tag{1.7}$$

The units of the reaction rate constant k_3 are $mole^{-1}$ *litre* sec^{-1}.

The reactions of order three occur rarely due to the lower probability of simultaneous collision of three molecules.

1.3 Examples

1.3.1 The Four Reaction Model

This model of decaying-dimerizing reactions[7] had been used by Daniel Gillespie and others for simulations by different methods. The reaction model comprises of following set of four reactions :

$$R_1 : X_1 \xrightarrow{c_1} \phi$$

$$R_2 : 2X_1 \xrightarrow{c_2} X_2$$

$$R_3 : X_2 \xrightarrow{c_3} 2X_1$$

$$R_4 : X_2 \xrightarrow{c_4} X_3$$

where X_1, X_2, X_3 are the different chemical species taking part in this reaction system, with c_1, c_2, c_3, c_4 as the reaction rate constants of the reactions. The chemical system consists of one bimolecular reaction, R_2 and three unimolecular reactions, R_1, R_3 and R_4. In this chemical system, the monomer X_1 decaying through reaction R_1 gets reversibly (through R_2 and R_3) dimerized to an unstable dimer X_2. Further, the reaction R_4 converts X_2 to a stable species X_3.

1.3.2 The Lotka-Volterra Model

The exceptionally remarkable dynamical properties (oscillating behavior) of a set of coupled auto-catalytic reactions were first observed by Alfred Lotka.[8,9,10] Later, Vito Volterra independently applied reaction rate equations to corresponding chemical reactions for the modeling of population dynamics in ecology. Hence, the name Lotka-Volterra model.[11,12] The model consists of following reactions[13] :

$$R_1 : \overline{X} + Y_1 \xrightarrow{c_1} 2Y_1$$

$$R_2 : Y_1 + Y_2 \xrightarrow{c_2} 2Y_2$$

$$R_3 : Y_2 \xrightarrow{c_3} Z$$

where Y_1 and Y_2 are the reactant species that change their molecular numbers with time, while $\overline{X}$ denotes the species whose number of molecules remain constant with time. c_1, c_2, c_3 are the reaction rate constants of the three reactions. The dynamics of these reactions can be understood by using the predator-prey interpretation. Here, Y_1 is treated as a prey and Y_2 as a predator. The reaction R_1 shows the reproduction of species Y_1(prey) by eating the abundantly available food, $\overline{X}$ in nature. The reproduction of species Y_2 by feeding itself on prey species, Y_1 is described in reaction R_2. Finally, the death of the predator, Y_2 is described through reaction R_3.

1.4 Computational Methods

Since the advent of computing machines, the computer simulations of natural phenomena had emerged as one of the key disciplines.[14] An increase in the computing power over the years had further contributed to the field. The methods for the simulations of biochemical reaction networks have been discussed and reviewed in the literature.[15,16,17] For such chemical systems, the choice of a simulation methodology depends on the features that need to be explored, the chemical reactions, the molecule numbers and availability of the computational resources.

Generally, these simulation methods are classified into two main categories of deterministic and stochastic methods. They are further classified depending on the continuous and discrete regime. The deterministic methods with ordinary or partial differential equation (ODE/PDE) based models[18,19] falls in the continuous regime, while those dealing with the Boolean models,[20,21,22] and Petri nets comes in the discrete regime. The continuous stochastic approaches with approximations based on the master equations deals with stochastic differential equations (SDE) like the Fokker-Planck[23,24] and the Langevin equation.[25,26,27] The commonly used kinetic Monte Carlo based Gillespie algorithm[28,13] and other τ-leap based algorithms[7,29] come in the category of discrete stochastic methods.

1.4.1 Deterministic Method

The kinetics of the conversion of sucrose into glucose and fructose was reported by Ludwig Wilhelmy[30] in 1850 by using an ordinary differential equation (ODE). He found that the reaction rate is directly proportional to the concentrations of the reactants. This discovery putforth the future foundations for the modeling of chemical reactions by using ordinary differential equations (ODEs). Thus, for each chemical species an ODE can be written depending on the type of a chemical reaction. This set of ODEs corresponding

to each chemical species are called as the reaction rate equations (RREs). For a reaction network consisting of *N* chemical species $\{S_1, ..., S_N\}$ with X_i as the number of molecules of *i*th species the RREs are of the form :

$$\frac{dX_i}{dt} = f_i(X_1, ..., X_N) \quad (i = 1, ..., N) \tag{1.8}$$

where the functions f_i depend on the types of reactions. Eq.(1.8) is usually expressed in terms of concentration variables, $Z_i \equiv X_i/\Omega$, where Ω is the system volume. The species concentrations are continuous functions of time. This traditional approach is applicable for a well-stirred chemical system where there are no fluctuations or any correlations among the molecules.

The analytical solutions are obtained for simple reaction networks, while for the complex one, it seems unfeasible to solve Eq.(1.8). Hence, numerical iterative methods have had been used to obtain their solutions. In case of the RREs, given the initial value conditions like the species concentrations at *t* = 0 and the values of the rate constants the time evolution trajectories of the species can be obtained. Over a fixed period of time, for given initial conditions, the same set of output values are generated: a fact underlying the deterministic nature of the RREs. In the last few years, the numerical study of ODEs has flourished into a well-established branch of applied mathematics with the availability of sophisticated softwares. But, given a wide range of numerical methods, one must choose a suitable algorithm for integrating such RREs.[18,19]

In case of ODE based numerical schemes, the species concentrations are treated as continuous functions of time. The deterministic approach is adequate for chemical systems with sufficiently large number of molecules (in the thermodynamic limit) giving a continuous description of time evolution. However, in certain chemical systems the molecular populations of some key species becomes so small that it may trigger fluctuations in their concentrations. In such a scenario, the assumption of continuous description of concentration breaks down as the discreteness of molecules comes into the picture.[31,32,33,34,35,36,37]

For example, in case of gene expression, the reactions (transcription, translation and mRNA degradation) are stochastic in nature which gives rise to fluctuations. Here, the lower number of protein and mRNA molecules makes the deterministic simulations inconvenient. This paves the way for stochastic simulations, where the discrete and stochastic nature of molecules is taken into account.

The following example of the Lotka-Volterra Model illustrates the discrepancy between the deterministic and stochastic approach. It can be seen (from Figure 1.1) that the stochastic averages do not match with the results of the deterministic solutions. The RREs for the set of reactions described in the previous section (1.3.2) are provided below[13] :

$$\frac{d[Y_1]}{dt} = c_1\overline{X}Y_1 - c_2Y_1Y_2 \tag{1.9}$$

$$\frac{d[Y_2]}{dt} = c_2Y_1Y_2 - c_3Y_2 \tag{1.10}$$

This predator-prey system can reach a steady state, when the rate of reproduction of prey species, Y_1 becomes the same at which it is getting consumed and the predator species, Y_2 takes birth at a rate equal to the rate of its death. In other words, such a kind of state can be attained by setting the time derivatives of Eq.(1.9) and Eq.(1.10) to zero as follows :

$$\frac{d[Y_1]}{dt} = \frac{d[Y_2]}{dt} = 0 \tag{1.11}$$

The solution of Eq.(1.11) gives $Y_1 = Y_{1s} = c_3/c_2$ and $Y_2 = Y_{2s} = c_1\overline{X}/c_2$. It is found that the deterministic simulation performed by using these values (Y_{1s} and Y_{2s}) gives a constant trajectory of the respective species with time. On the other hand, the physically realistic approach of stochastic simulation[28,13] depicts oscillatory behavior of the corresponding species. The simulations were performed by using the following parameters : $[Y_1]$ = 1000 molecules $litre^{-1}$, $[Y_2]$ = 1000 molecules $litre^{-1}$, c_1 = 10 sec^{-1}/ (molecules $litre^{-1}$), c_2 = 0.01 sec^{-1}/(molecules $litre^{-1}$) and c_3 = 10 sec^{-1}/(molecules $litre^{-1}$)

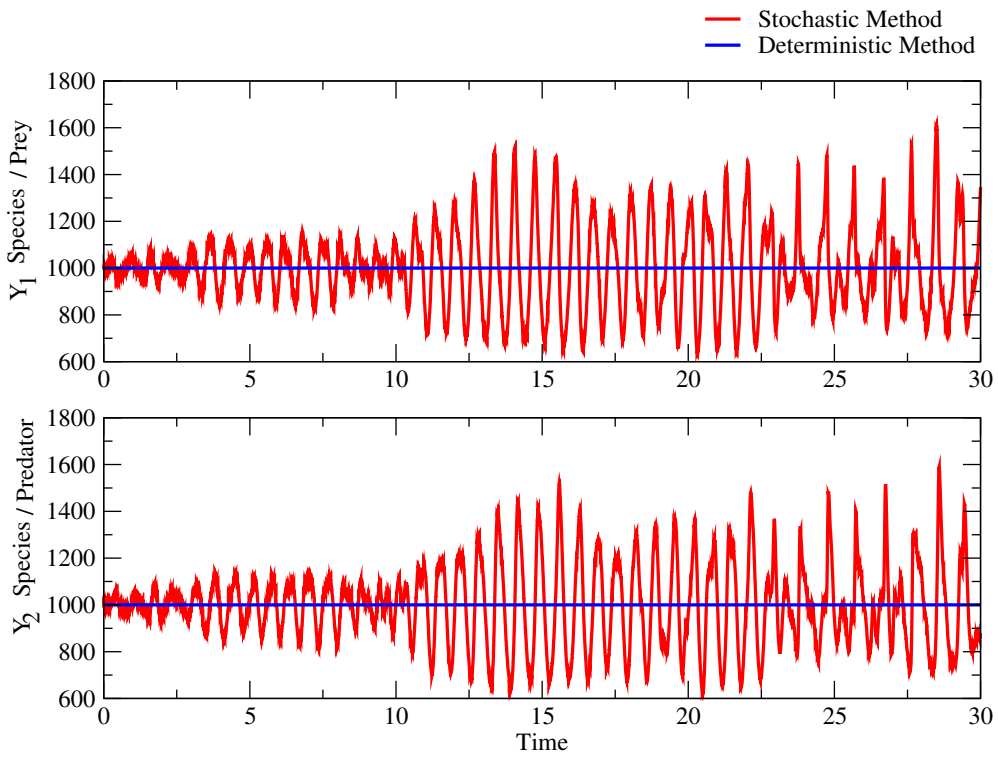


Figure 1.1: The stochastic method produces an oscillatory behavior for the predator and prey species, while the deterministic values remain constant with time.

1.4.2 Stochastic Method

In statistics, probability theory deals with the study of random events taking place in nature. These events are characterized by certain random variables. The evolution of these random variables that are associated with some system over a period of time is treated as a random process or a stochastic process. The successive tossing of a coin, a path traced by a random walker on a lattice, the exchange rate between the US Dollar and the Indian Rupee over a certain period, the emission of photons, the share prices on stock markets and meteorological data are examples of stochastic processes in our day-to-day life. All the aforementioned examples can be studied by constructing stochastic models.

Historically, it was Albert Einstein[38] (1905) and Smoluchowski[39] (1906) who marked the beginning of stochastic modeling. They explained the phenomenon of Brownian

motion where random fluctuations were explained by using statistical concepts.[40] Earlier, Maxwell and Boltzmann had used statistics to give a probabilistic description of the occupation of possible states by gas molecules. But, time evolution was not considered in their theories. In the following years, several scientists like Langevin, Kolmogorov, Doob, Ito and others contributed to the theory of stochastic processes. However, until the 1960s, there were few applications of these ideas to chemical kinetics.

In 1940, it was biophysicist Max Delbruck[41] who took into account the discrete-stochastic nature of molecules. He solved the differential equations with statistical fluctuations for autocatalytic reactions. Around the same time, H. A. Kramers[42] used stochastic ideas to treat chemical reaction like the Brownian motion of particles. Further, Bartholomay derived a stochastic model for unimolecular chemical reactions[43,44] and also for the Michaelis-Menten reaction.[45,46] In the 1960s, in a series of papers, McQuarrie[47,48,49] applied the theory of stochastic processes to study the chemical kinetics of small systems. There, McQuarrie introduced a differential-difference equation, which is nowadays known as the chemical master equation (CME):

$$\frac{\partial P(X,t)}{\partial t} = \sum_{X\prime} \left[W_{XX\prime} P_{X\prime} - W_{X\prime X} P_X \right] \tag{1.12}$$

Here, $P(X, t)$ is the probability of finding the system in state X at time t, while $W_{XX\prime}$ and $W_{X\prime X}$ are the transition probabilities from the states $X\prime$ and X respectively. X represents a vector of species concentrations taking part in a chemical system. The first term on the right hand side of Eq.(1.12) represents a gain in state X with an increase in its probability, while the second term is a loss (accompanied by negative sign) in state X with a decrease in probability.

The CME in Eq.(1.12) is basically a set of coupled ODEs with one equation for each possible reactant combination of molecules. Thus, owing to its complexity, it is only possible to give an analytical solution for a select few simple chemical systems. But, most of the realistic chemical systems consist of large number of reactions. For such complex systems, numerical solutions are required. The numerical approach deals with

the construction of a simulated time profile (trajectory) for each species, i.e. the graph of $X_i(t)$ vs t. The simulation procedures developed for the construction of such trajectories have been discussed in the next chapter.

1.5 Outline of the thesis

The next chapters of this thesis are organized as follows:

Chapter 2 discusses the state-of-the-art stochastic simulation methods that have been developed over the years. Initially a brief historical account of the kinetic Monte Carlo (kMC) methods is provided. The basis of the stochastic simulations is followed with the rigorous theoretical details of the kMC based stochastic simulation algorithm (SSA).[28, 13] Nowadays, the stochastic simulation algorithm (SSA) is widely used for the study of chemical kinetics. Other contemporary algorithms have also been included in the discussion. The expensive computational burden of these aforementioned simulation methods is relieved by the introduction of approximate accelerated simulation methods. The discussion pertaining to a number of such simulation methods is given in the chapter. The last section is based on the treatment of analytical methods.

Chapter 3 is based on our work towards the development of a new accelerated stochastic simulation method. It includes a new methodology that had been proposed in order to increase the computational efficiency of SSA like algorithms. The crux of this new method lies in representing the entire reaction network with a single representative reaction (RR); hence the name 'representative reaction approach (RRA)'.[50] The choice of an appropriate RR, the application of a leap condition to this RR, the calculation of expected number of reactions and the subsequent time step, τ have been discussed in detail. Further, the accuracy of the RRA method is tested by applying it to three different examples. Here, it has been found that the proposed method works well, especially in simulating the behavior of complicated oscillatory chemical reactions.

Chapter 4 reports our work on the development of a computational method that attempts to provide a solution to the occurrence of negative molecular numbers during the simulations. Initially, a brief overview of other algorithms that have tried to solve the problem of negative populations is given. This new algorithm works by combining the newly proposed RRA method with the SSA and also with the ideas of the binomial distribution.[51] The steps for their implementation are outlined in the form of a simple flowchart. The new method has been applied successfully to chemical systems, which are prone to the occurrence of negative numbers. It has been found that the new algorithm is efficient and accurate.

Chapter 5 deals with the same problem that has been mentioned in the previous chapter, but from a totally different perspective. The occurrence of these unrealistic (or negative) numbers can also be seen as a signature of the failure of the leap condition. In this work,[52] it is speculated that a certain *noise* is associated with the choice of reaction numbers of individual reactions. In the simulations, this novel concept of *noise* is used along with the RRA method. This RRA_noise method is validated by simulating different chemical systems, ranging from uni-molecular to oscillatory reactions.

Chapter 6 includes a description of a new accelerated stochastic simulation method, which is based on the idea of choosing an error control parameter, ε in a logical way. Generally, all the accelerated stochastic simulation methods take a time step, τ by using an arbitrarily chosen error control parameter. Here, a model based on a coupled harmonic oscillator is proposed, which can be connected with the representative reaction approach (RRA). Thereafter, ε is chosen based on mathematical considerations for using it further in the evaluation of the time step, τ. The idea has been applicable for efficiently simulating the chemical systems and also for dealing with the issue of negative molecular numbers.

CHAPTER 2

Stochastic Simulation Methods

All Models are wrong, but some are useful.

-George E. P. Box

2.1 Monte Carlo Simulation

Nowadays, the computer simulations stand out as a complementary to the traditional branches of theoretical and experimental sciences. The computer oriented numerical methods are used in the form of simulations. They are used for solving difficult analytical equations or to gain insights into the behavior of complex systems for which designing experiments is demanding. Monte Carlo simulation is one such renowned numerical technique that uses the notion of random numbers for the simulations. This technique was first formulated in the middle of the twentieth century; when Ulam, von Neumann, and Fermi were working on the Manhattan project. The name Monte Carlo was suggested by Metropolis referring to the famous Monte Carlo casino in Monaco. The method of the Metropolis Monte Carlo algorithm[53] is used for studying equilibrium properties by generating configurations according to some desired distribution. However, in Metropolis Monte Carlo, time is not involved, making it unfeasible for studying the time evolution or kinetics.

Over the years, many methods were developed to incorporate time to study simulations of the physical processes. One such attempt was made in 1975 by Bortz *et al.*,[54] when they developed an algorithm for the simulation of Ising spin systems. It was followed by the development of another Monte Carlo based sophisticated algorithm called as stochastic simulation algorithm (SSA)[28,13] for the simulation of chemical kinetics. In the early days, these types of computational methods were called as 'dynamic-Monte Carlo', 'time dependent Monte Carlo'or simply 'Monte Carlo'. The usage of the terminology 'kinetic Monte Carlo (kMC)'was started in the 1990s.[55] The essence of kMC based methods lies in the use of transition rates that depend on the energy barrier followed by time increments. The kMC based algorithms describe exact time evolution of the underlying processes. However, one has to know the processes and their rates in

advance.

The kMC have had been used in the simulations of surface diffusion, thin film growth, crystal growth, vacancy diffusion in alloys, defect mobility and clustering in irradiated solids.[56,57,58,59,60,61,62,63] In recent years, the kMC algorithms have gained wide popularity through their applications to the reaction networks in biological systems. In these aforementioned applications, the processes to be simulated like growth, diffusion and chemical reactions are treated as stochastic processes: hence the resulting simulations are called as stochastic simulations. In this thesis, the theoretical underpinnings followed by applications of the stochastic simulation algorithms to simulate chemical reactions are discussed.

2.2 Framework : Stochastic Chemical Kinetics

In this section, some of the notations and concepts developed by Gillespie and co-workers have been discussed.[28,13,25,17,16] It is assumed that a well-stirred chemical system of *N* chemical species $\{S_1, ..., S_N\}$ interacts through *M* chemical reactions $\{R_1, ..., R_M\}$. Let X_i denote the integer number of molecules of species S_i. The entire mixture of some fixed volume Ω is kept in thermal equilibrium at some constant temperature T. The dynamical state of the system is specified by the vector: $\mathbf{X}(t) \equiv \{X_1(t), ..., X_N(t)\}$, provided that the chemical system was in state $\mathbf{X}(t_0) = \mathbf{x}$ at some initial time t_0. Each reaction is characterized by the following quantities :

(i) the state change vector, $\boldsymbol{\nu}_j \equiv \{\nu_{1j}, ..., \nu_{Nj}\}$, where ν_{ij} is the change in the S_i molecular population due to reaction channel, R_j

(ii) the propensity function, $a_j(\mathbf{x})$, of a reaction is the product of the rate constant times the number of different reactant combinations. Then,

the probability, that the R_j reaction will take place somewhere inside volume Ω in the next infinitesimal time interval [*t*, *t*+*dt*) is given by :

$$a_j(\mathbf{x})\, dt. \tag{2.1}$$

The above quantities are explained by using some of the commonly occurring different types of reactions $\{R_1, R_2, R_3, R_4\}$ with their respective reaction rate constants $\{c_1, c_2, c_3, c_4\}$:

$$R_1 : \phi \xrightarrow{c_1} X_1$$

$$R_2 : X_1 \xrightarrow{c_2} X_2$$

$$R_3 : X_1 + X_2 \xrightarrow{c_3} X_3$$

$$R_4 : 2X_1 \xrightarrow{c_4} X_2$$

The state change vectors for these reactions are as follows :

$$\boldsymbol{\nu}_1 \equiv \{\nu_{11} = 1\}$$

$$\boldsymbol{\nu}_2 \equiv \{\nu_{12} = -1, \nu_{22} = 1\}$$

$$\boldsymbol{\nu}_3 \equiv \{\nu_{13} = -1, \nu_{23} = -1, \nu_{33} = 1\}$$

$$\boldsymbol{\nu}_4 \equiv \{\nu_{14} = -2, \nu_{24} = 1\}$$

The propensities of the above mentioned reactions are given below :

$$a_1 = c_1$$

$$a_2(x_1) = c_2 x_1$$

$$a_3(x_1, x_2) = c_3 x_1 x_2$$

$$a_4(x_1) = c_4(x_1(x_1 - 1))/2$$

The stochastic evolution of the state of the chemical system, $\mathbf{X}(t)$ is specified by the function $P(\mathbf{x}, t|\mathbf{x}_0, t_0)$. The laws of probability[64] are applied to the definition of $a_j(\mathbf{x})$ in Eq. (2.1) to derive a time evolution equation for the stochastic process. This equation is called as the chemical master equation (CME).[65]

$$\frac{\partial P(\mathbf{x}, t|\mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^{M} [a_j(\mathbf{x} - \boldsymbol{\nu}_j) P(\mathbf{x} - \boldsymbol{\nu}_j, t|\mathbf{x}_0, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t|\mathbf{x}_0, t_0)] \qquad (2.2)$$

The above equation describes the time evolution of the probability, $P(\mathbf{x}, t)$ with respect to some given initial conditions $(\mathbf{x}_0, t_0)$. The solution of the CME gives a complete description of the stochastic dynamics of the underlying chemical system. However, it is a Herculean task to solve the CME for complex chemical systems rather than for some simple one.

2.3 Stochastic Simulation Algorithms

In order to circumvent the difficulties associated with the CMEs, researchers turned towards the development of numerical simulation methods. The essence of these techniques is to find the moments without calculating the corresponding probability distribution functions. These moments are estimated over an ensemble of a sufficiently large number of simulation runs. The decade of 1970s saw various initial attempts[66,67,68,69,70,71] in this direction, but the lack of generalization could not take them far.

In the year 1976, it was Daniel Gillespie, who proposed the kinetic Monte Carlo based elegant, general-purpose yet simple algorithm for the stochastic simulation of chemical reactions. Nowadays, it is known as the stochastic simulation algorithm (SSA).[28] The simulations are carried out by finding out *which* reaction will occur and *when* that particular reaction will occur. In other words, two random variables (μ and τ) are associated with these two events. These two random variables are distributed by the following the joint probability density function, which is a consequence of Eq. (2.1) defined in the previous section.

$$p(\tau, \mu|\mathbf{x}, t) = a_\mu(\mathbf{x}) \exp(-a_0(x)\tau) \tag{2.3}$$

Here, τ is a real random variable with exponential distribution, while μ is an integer with point probabilities. Gillespie introduced the direct method[28] of generating a pair of random variables (τ, μ) in each iteration. These two random variables are selected as

follows :

$$\tau = \frac{1}{a_0(\mathbf{x})} ln\left(\frac{1}{r_1}\right) \tag{2.4a}$$

$$\mu = \text{the smallest integer satisfying} \quad \sum_{\nu=1}^{\mu} a_\nu(\mathbf{x}) > r_2 a_0(\mathbf{x}) \tag{2.4b}$$

where r_1 and r_2 are two uniformly distributed random variables in the interval (0,1).[72] There is another method (called as the first reaction method)[28] of generating the random variables, which is less efficient than the direct method. The elegance of the direct method lies in the execution of a randomly selected *single reaction* event in the next speculated time interval. In this way, different *single reaction* events are simulated in each time interval until the final simulation time is reached. The execution of only *one reaction* in a given time step contributes to the huge computational time for realistic reaction networks.

Thereafter, many attempts have contributed towards the improvement of the SSA. The next reaction method (NRM)[73] of Gibson and Bruck is a significant enhancement over the efficiency of the SSA. This method appears in a modified form of the first reaction method. It uses only a single random number per simulation event as compared to two in the direct method. Here, the index priority queue is used to store the putative reaction times of all the reactions, with the times arranged in an ascending order from top. Thus, the index and the time of the next reaction is always at the top position in the queue. Afterwards, using the idea of dependency graphs, the respective reaction propensities are updated. This avoids the redundant recalculations of all the reaction propensities. In this way, the NRM algorithm saves the computational time relative to the SSA, but its programming is a formidable task.

It was found that summing the propensity functions in Eq. (2.4b) contributes to an increase of the computational load in the SSA. Thus, by taking into account the aforementioned issue the optimized direct method,[74] the sorting direct method[75] and the logarithmic direct method[76] have been developed. The software package *Moleculizer*[77] developed by Lok and Brent makes use of the just-in-time strategy (using the reactions and

species only when required) to simulate the mechanism of yeast. Slepoy *et al.*[78] developed a method in which the next reaction is chosen with a constant-time, independent of the number of reactions. The modified next reaction method of Anderson,[79] which incorporates the time dependent propensities and time delays, is found to be efficient over existing exact simulation methods.

An exact accelerated stochastic simulation algorithm (ER-leap algorithm)[80] simulates multiple reactions in a single step, but with the same probability distribution Eq. (2.3) as the SSA. The delayed reactions[81] have also taken into account in the delay stochastic simulation algorithm (DSSA). More recently, in 2013, Yates *et al.*[82] have proposed an exact stochastic simulation algorithm based on the recycling direct method (RDM). This algorithm uses a single uniform random number with a statistically acceptable recycling of random numbers. This simplicity of the RDM (just one line modification in the SSA) is expected to enhance the applications of the simulation methods in the modeling of biological systems.

2.4 Approximate Accelerated Algorithms

2.4.1 Tau Leaping Methods

There is no doubt that the developments described in the previous section over the years are quite satisfactory. However, the Achilles heel (in spite of accuracy) of all these exact simulation methods is the execution of a *single* reaction in each time step. This, in turn, has lead to an increase in the computational load. Thus, various attempts have been made to sacrifice the accuracy for the speed-up of the simulations. This section discusses all the attempts made in this direction.

In the year 2001, Daniel Gillepsie came up with an approximate accelerated stochastic simulation algorithm. The method is referred as Gillepsie's approximate stochastic algorithm (GASA).[7] Here, the time step, τ is taken sufficiently large so that many more

reaction events are allowed to occur. This time step, τ is chosen in such a way that it fulfills the *leap condition*,[7] which is: the time step has to be small enough so that no propensity function changes by an appreciable amount.

$$a_j(\mathbf{x}) \approx \text{constant in} \quad [t, t+\tau), \forall j \tag{2.5}$$

Further, the reaction events taking place in this time step (leap) are modeled by the Poisson distribution[72] owing to their discrete nature. The random variable (expected reactions), $a_j\tau$, which is provided to the Poisson random number generator gives an integer number of reactions, $\mathcal{P}(a_j(\mathbf{x})\tau)$, that are supposed to occur in the next time step τ. The change in the state of the system is done as follows :

$$\mathbf{x}(t+\tau) = \mathbf{x}(t) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \mathcal{P}(a_j(\mathbf{x})\tau) \tag{2.6}$$

The execution of multiple reactions as compared to a single reaction is the essence of this method. Thus, the multiple reactions and large enough leaps in time makes this algorithm computationally efficient. The more consistent choice of the time step, τ, with the leap condition is provided by the Gillepsie-Petzold (GP)[29] and Cao-Gillepsie-Petzold (CGP)[83] tau leap algorithms.

The multiscale time behavior appears in many realistic chemical systems. There are chemical systems which consist of slow and fast reactions. The fast reactions occur many more times during a typical SSA simulation run than the slow reactions. The fast reactions reach equilibrium more quickly, leaving behind reactions in the slow regime. In fact, the slow reactions are accountable for the dynamics of the chemical system, but much of the time is invested in the simulation of fast reactions. In the context of numerical solutions of ODEs,[18,19] this property is called as the *stiffness*. This issue has been addressed by the implicit tau leap method[84] by using an idea of numerical integration of ODEs. In the slow-scale stochastic simulation algorithm (ssSSA),[85] the insignificant fast reactions are ignored and the slow reactions are simulated using their modified propensities. There are several methods that use the concept of partial equilibrium assumptions[86,87,88,89] to skip the fast reactions.

In recent times, the representative reaction approach (RRA)[50] method is one of the contributions to the approximate accelerated algorithms. It is found that this method is significantly faster for simulating oscillatory reactions, which are difficult to handle using other accelerated methods. The simplicity in coding is yet another advantage over other methods. The new tau leap method proposed by Doraiswami *et al.*[90] based on the Chebyshev's inequality[64] gives a probabilistic view for the assurance of the leap condition. All these methods have increased the speed of the simulations. But, in some cases where there are a small number of species, they have given rise to physically unrealistic numbers (negative numbers) during their simulations.

This difficulty during the simulations is addressed by developing some other simulation methods. One of the main reasons, for the occurrence of negative numbers is the unbound nature of Poisson random variables, which are used for the modeling of reactions. Hence, the Poisson random variables are replaced by the Binomial random variables, $\mathcal{B}(N_j, a_j(\mathbf{x})\tau/N_j)$ using the properties of the later. In this case, the change in the state of the system is given by the following equation :

$$\mathbf{x}(t+\tau) = \mathbf{x}(t) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \mathcal{B}(N_j, a_j(\mathbf{x})\tau/N_j) \tag{2.7}$$

Tian-Burrage[91] and Chatterjee *et al.*[92] have proposed the (bounded) binomial tau leap methods. The multinomial tau leaping (MτL)[93] is an extension of the binomial methods. The Tian-Burrage method fails for the cases of multiple-channel reactant dependencies. This limitation is overcome by the modified binomial Leap method of Peng *et al.*,[94] but with an increase in the computational time. The method of Chatterjee *et al.* solves the problem of negative numbers, but it introduces a bias in the choice of the reaction numbers depending on the order in which they are selected. The algorithm of Cao *et al.*[95] partitions the reaction network into the critical and non-critical sets of reactions. The critical reactions are modeled by the SSA, while the Poisson distribution is used for other. The N-leap method,[96] K-leap method[97] and R-leap method[98] deal with the calculation of the total reactions from the leap condition. These methods have found

dual applications in the speedup as well as solving the issue of negative numbers. The RRA method used in conjunction with the binomial distribution[51] and the idea of noise[52] have also helped to solve the problem of negative numbers.

2.4.2 Hybrid Methods

This is another class of methods that has been used for the simulation of multiscale systems. It employs deterministic methods along with the stochastic. The reaction rate equations or the chemical langevin equation are used with the SSA. Thus, being an amalgam of two different approaches, it has the name 'hybrid methods'.[99,100,101] The deterministic methods are used for the simulation of fast reactions (large species), while the stochastic methods are used for the simulation of slow reactions (small species). Mathematically speaking, this implies the partitioning of the CME. Hellander and Lotsted[102] have given the solution of the CME by coupling the deterministic and stochastic methods. It is found that for certain biological systems, the hybrid solver is more efficient than the SSA. Nevertheless, the hybrid methods lack the required theoretical framework. The reliable criteria of partitioning, handling of the fast reactions with small species, and dynamic re-partitioning are some of the key issues.

2.5 Analytical Methods

The numerical computational schemes discussed earlier require the sampling of a large number of realizations (an ensemble of simulation runs) to estimate authentic statistics in the form of moments of different species of chemical systems. This in turn have contributed towards the computational cost of the simulations. The application of the analytical methods to estimate the mean and variance helps to make the effort computationally less expensive. The use of analytical techniques to chemical kinetics dates back to 1940, when Max Delbruck gave a mathematical reasoning to the fluctuations in the

autocatalytic reactions. Thereafter, researchers have applied the analytical ideas to various simple and specific chemical systems of interest. It must be remembered that the study of chemical kinetics was initiated by using the analytical methods much before the advent of high performance computing. Thus, an increase in the computing power over time has helped in the advancement of such methods.

The CME that gives a probabilistic picture of the underlying dynamics can be solved exactly for very few chemical systems. The systems that involve only uni-molecular reactions,[103, 104] bi-molecular reaction with two reactant species[105] and the closed system of equations are some of them. For the general case of chemical systems, we have to look for some other techniques. The methods of linear noise approximation (LNA),[23] moment closure approximations[106, 107, 108] and probability generating function (PGF)[49, 109, 110] gives an approximate solution of the CME.

The system size expansion[23] introduced by the theoretical physicist Nico van Kampen gives the moments in terms of the power series expansion of the inverse volume of the chemical system of interest. The leading first order term of this expansion is given by the LNA. This term is valid in the limit of large volumes. In this case, the first moment is decoupled from the second moment and are given by the RREs. The time evolution equations of the second moment are decoupled from the third moment. All this works when the species are present in large amounts (large volumes). The limitation of LNA is that it is not good for systems with a small number of species.[108]

For the case of higher order reactions (second order), the equation of moments appear in the form of coupled equations. The moment equations of the first moment include the second and higher moments, while those corresponding to the second moment include the next higher moments. This leads to an infinite hierarchy of coupled equations.[106, 108] At this point, an approximation is required that neglects the moments higher than a certain order by setting them to zero. This is the moment closure approximation. It is more accurate for the intermediate number of species than the LNA. The theoretical validity of these methods has been discussed by Grima *et al.*. Lee *et al.*[106] have shown that the

moment closure approximations are efficient and more accurate than exact methods. The method of the probability generating function (PGF) is used for the numerical estimation of the probability distribution and first and second order moments. It was used for the first time for the numerical computation of the probability distribution by Lee *et al.*. It is used by converting the CME to a partial differential equation (PDE). Further, by taking the derivatives of its solution, the moments are estimated.

2.6 Softwares

The exact and approximate simulation techniques discussed above have been put together in the form of different sophisticated software packages. Some of them are: StochKit2,[111] Kinetikit,[112] BioNets,[113] STOCKS.[114] In the last year, StochSS (Stochastic Simulation Service)[115] has been launched. It incorporates deterministic, stochastic as well as the spatial stochastic simulations.

CHAPTER 3

A New Approximate Method for the Stochastic Simulation of Chemical Systems : The Representative Reaction Approach

Statistical thinking will one day be
as necessary for efficient citizenship
as the ability to read and write.

-H. G. Wells

Abstract

We have developed two new approximate methods for stochastically simulating chemical systems. The methods are based on the idea of representing all the reactions in the chemical system by a single reaction, i.e., by the representative reaction approach (RRA). Discussed in the article are the concepts underlying the new methods along with a flowchart with all the steps required for their implementation. It is shown that the RRA with the reaction $2A \longrightarrow B$ as the representative reaction (RR) performs significantly faster than the exact stochastic simulation algorithm (SSA) developed by Daniel Gillespie and is able to successfully reproduce at least the first two moments of the probability distribution of each species in the systems studied. Moreover, the RRA is shown to be simpler and more effective than Gillespie's approximate stochastic algorithm (GASA) in handling systems where the species fluctuate by different orders of magnitude. As such, the RRA methods represent a promising new method for stochastically simulating chemical systems.

3.1 Introduction

The study of the kinetics of chemical systems has traditionally involved a master equation[49,65] : a set of coupled ordinary differential equations (reaction rate equations) describing the time evolution of the concentration of the different chemical species reacting in the system. To solve such differential equations numerically, with a given set of rate constants and initial concentrations, some integrators[116] have been used to calculate the time profiles of the concentrations. An alternative to this deterministic approach has been put forward by Doob[117,118] and Gillespie[13,28] whose seminal work focused on a stochastic approach, more specifically, the kinetic Monte Carlo[54] - based stochastic simulation algorithm (SSA). Unlike the numerical algorithms, the SSA does not approximate the time increments by specific finite time steps and also takes into account the fluctuations in the system.

Since its introduction, the SSA[28,117,118] has become popular for the study of chemical kinetics of different systems, especially biological systems that involve genetic regulatory networks and cellular processes.[119,120,121,122] However, due to the occurrence of a single reaction event in each time increment, the practical application of the SSA is severely limited with respect to the time scale and the molecular populations of the chemical systems that it can effectively simulate. Several approximate methods have therefore been proposed, so as to speed up the SSA. These include the Poisson τ-leap method,[7] the midpoint τ-leap method,[7] the implicit τ-leap method,[84] the Poisson Runge-Kutta methods,[123] the multinomial τ-leap method,[93] efficient step size selection for the τ leaping,[83] and the binomial τ-leap methods.[91,92] Apart from these, other attempts have also been made to reduce the computational load of the SSA: He *et al.* have used a hybrid Monte Carlo algorithm for polymerization reaction kinetics;[124] Gibson and Bruck have modified the first reaction method (similar to the SSA) such that unused reaction times could be refined for reuse;[73] Rao and Arkin applied the quasi steady-state assumption to

the subset of fast reactions in the system to reduce the computational time to numerically simulate the systems,[89] and Haseltine and Rawlings have tried to improve the computational efficiency by partitioning the system into subsets of fast and slow reactions.[101]

Of these, the most appealing in terms of (relative) simplicity and effectiveness is Gillespie's approximate Poisson τ-leap method[7] (henceforth referred to in the manuscript as Gillespie's approximate stochastic algorithm (GASA)). The principal idea behind this method is that instead of executing a single reaction in every time increment and changing the molecular population accordingly (as in the SSA), a larger leap τ in time is taken, during which all of the reactions in the system are allowed to occur, the number of occurrences of each being determined with the aid of random numbers selected from the Poisson distribution. The size of the τ leap is determined from the leap condition which Gillespie defines as the necessary requirement that the propensity functions (the product of the rate constants with the number of reactant combinations) of the reactions do not change appreciably in value as a result of the leap.

GASA has proved successful in accelerating the simulations of several different types of chemical systems, in comparison to the SSA, while also replicating at least the first two moments of the probability distribution of the species with time. However, for chemical systems where the species amounts fluctuate significantly, or cases where the species amounts vary by different orders of magnitude, GASA has been found wanting in terms of accurately replicating the changes in species amounts with time. Modifications have been proposed to the GASA method to improve its reliability, such as the Gillespie-Petzold (G-P) method,[29] but, to date, most attempts at improvement have also led to increasing complexity of the algorithms employed. A method that can retain the simplicity of GASA, while also being able to reliably and efficiently simulate complex and challenging chemical systems, would therefore be highly desirable. Our objective in this manuscript is to propose just such a method.

This article is organized as follows: (i) first, we have discussed the theoretical basis for our proposed method, which we have termed the representative reaction approach

(RRA), where we have discussed the leap condition for the RRA as well as the choice of the most appropriate RR, along with a flowchart describing the steps in the algorithm; then we have tested our two proposed RRA methods, RRA-τ and RRA-*N*, on (ii) a model system of chemical reactions, on (iii) a more complicated system with chemical oscillations, on (iv) a system having reacting species fluctuating by different orders of magnitude, and finally (v) we have provided an evaluation of our method and presented conclusions.

3.2 Methodology

3.2.1 Chief concepts and notations

This subsection discusses in brief some of the notations and concepts that have been employed by Gillespie[7] and co-workers[92,91,125,86] in developing the approximate stochastic simulation methods. To study the evolution of molecular numbers, a well-mixed reacting system of *N* molecular species $\{S_1, ..., S_N\}$ is considered. Let $X_j(t)$ denotes the number of species of S_j at time t. This entire mixture of chemical species interacts inside some fixed volume Ω at a constant temperature through reaction channels $\{R_1, ..., R_M\}$. For each reaction channel R_i (*i* = 1,...,*M*), a propensity function $a_i(\mathbf{x})$, is defined which, along with a_i**(x)***dt*,[65,25] gives us the probability that the R_i reaction will take place in the infinitesimal time interval [*t*, *t*+*dt*). This propensity function a_i**(x)** is the product of the rate constant with the number of reactant combinations for the given reaction. For a reaction of the type X_1 + $X_2 \longrightarrow 2X_1$, a_i**(x)** = $c_1 x_1 x_2$, where c_1 is the specific reaction probability rate constant for the reaction, being algebraically related to the conventional deterministic rate constant k_1 by $c_1 = \left(\frac{k_1}{N_A \Omega}\right)$ where k_s are in the units of mole inverse second inverse and Ω in liters. N_A is the Avogadro's number. x_1 and x_2 are the amounts of the reactants X_1 and X_2. For a reaction of the type $2X_1 \longrightarrow 2X_2$, a_i**(x)** = $c_2 x_1(x_1 - 1)/2$, where c_2 is the specific reaction probability rate constant for the reaction,

with $c_2 = \left(\frac{2k_2}{N_A\Omega}\right)$

3.2.2 Our approach

For any given step *j* during the simulation, our approach is to first determine the expected number of reactions that would take place for that step *j* for the entire chemical system comprising all of the individual reactions. This is achieved by representing the entire chemical system as a single, representative reaction, and then determining the expected number of reactions for that representative reaction (RR). Our method is therefore termed the representative reaction approach (RRA). The propensity function of the RR is taken to be a_0 : the sum of the propensity functions of all the individual reactions, while C_0 - the specific reaction probability rate constant for RR is determined as $C_0 = \sum_{i=1}^{M}\left(\frac{a_i(\mathbf{x})}{a_0(\mathbf{x})}\right)c_i$: the weighted average of all the c_{is} in the system. For any individual reaction, assuming a Poisson distribution for the possible number of occurrences of that reaction, the expected number of occurrences, n_i, in any given step is equal to $a_i(\mathbf{x})\tau$.[7] Now, one can either determine the value of τ for that particular step by (i) evaluating τ for the representative reaction, or alternatively, (ii) one can evaluate N_0 - the total number of reactions taking place for that step for the representative reaction, and then determine the value of τ for that step from $\tau = N_0/a_0(\mathbf{x})$.

Once τ is determined for that particular step, the expected number of occurrences for the *i*th reaction is determined from $n_i = a_i(\mathbf{x})\tau$.[7] This value $a_i(\mathbf{x})\tau$ is the variable used in poidev the computer algorithm used to generate the corresponding Poisson random number[72] k_i. k_i is the number of times the *i*th reaction will occur in that particular step *j*. Thus, the values of k_i corresponding to all the reactions can be determined for this *j*th step and the amounts of reactants for each reaction adjusted accordingly to afford the propensity functions $a_i(\mathbf{x})$ for the next, *j*+1 th, step, where the process of determining the k_{is} is carried out in exactly the same fashion.

3.2.3 The leap condition for the RRA

Detailed in the previous subsection are the two possible ways, indicated as (i) and (ii), for determining the k_{is} for each of the reactions for a specific step, based on the approach of considering the entire system as being represented by a single reaction. However, the question that has not been answered yet is : how does one determine either the τ for the method (i) or N_0 for the method indicated in (ii) We propose that τ or N_0 be chosen in a manner such that it satisfies the leap condition[7] for RR: that the propensity (in this case, a_0**(x)** of RR is not altered by an appreciable extent by the size of the jump, i.e., by the value chosen for τ or N_0.

For method (i), henceforth, referred to as the RRA-τ method, one can employ the leap condition criteria established by Gillespie[7] to determine the change in τ for the single reaction chosen as the representative reaction. Henceforth, this approach will be called the RRA-τ method. For single reactions, the Gillespie leap condition criteria for choosing τ becomes a simple expression and thus easy to implement. Examples of such expressions will be shown in the next subsection, when discussing the choice of the most appropriate RR.

For method (ii), henceforth, referred to as the RRA-*N* method, the approach employed will be to bound the change in the value of a_0**(x)**, the propensity function for the RR, as follows:

$$|a_0(\mathbf{x} + \lambda) - a_0(\mathbf{x})| \leq \varepsilon a_0(\mathbf{x}) \tag{3.1}$$

or

$$|\Delta a_0(\mathbf{x})| \leq \varepsilon a_0(\mathbf{x}) \tag{3.2}$$

where

$$\Delta a_0(\mathbf{x}) = a_0(\mathbf{x} + \lambda) - a_0(\mathbf{x}) \tag{3.3}$$

Here, ε is a parameter that would remain constant throughout the simulation, and λ is the amount by which we change the state of the system. Now, it has been shown in the past[7] that the term Δa_0**(x)** can be approximated by a first-order Taylor series expansion

as

$$\Delta a_0(\mathbf{x}) = \lambda \nabla a_0(\mathbf{x}) \tag{3.4}$$

Hence,

$$\lambda \nabla a_0(\mathbf{x}) \leq \varepsilon a_0(\mathbf{x}) \tag{3.5}$$

A further approximation is added at this point to further accelerate the system. This can be done by multiplying the function on the right of the inequality above ($\varepsilon a_0(\mathbf{x})$) by a factor. This is the equivalent of dividing $\Delta a_0(\mathbf{x})$ by the same factor, and is therefore justified in that it will lead to a reduction in the $\Delta a_0(\mathbf{x})$ value and thus further comply with the leap condition. This factor is chosen to be 16, so that we can define a new parameter $\varepsilon\prime$, where $\varepsilon\prime = 16\varepsilon$. The somewhat arbitrary nature of the choice of the value of 16 for the factor will be understood in the context of the essential tunability of the values of ε and $\varepsilon\prime$ that will be discussed in the next section: it is found that the algorithm produces the best results in terms of accuracy and speed when such a factor is employed. Hence, after incorporating the factor of 16, we get the new equation:

$$\lambda \nabla a_0(\mathbf{x}) \leq \varepsilon\prime a_0(\mathbf{x}) \tag{3.6}$$

where $\varepsilon\prime = 16\varepsilon$.

By using Eq. (3.6), and using the expression for $\nabla a_0(\mathbf{x})$ for the chosen RR, one can determine the value of N_0, and thereby τ, for a given step for that specific RR. Examples of the values of N_0 thus obtained for different RR cases will be discussed in the next subsection.

3.2.4 What is the most appropriate representative reaction (RR) ?

Described in this subsection are the expressions that can be derived for the methods (i) RRA-τ and (ii) RRA-N for different representative reactions (RRs).

$A \longrightarrow B$ **as the Representative Reaction (RR) :** The unimolecular reaction $A \longrightarrow B$ is the simplest possible choice for the RR, and therefore the one that is considered first.

(i) The RRA-τ method : As discussed in the previous subsection, for this method, the value of τ is determined from the leap condition for the RR (in this case $A \longrightarrow B$), as established in GASA.[7] For $A \longrightarrow B$, the expression for τ is

$$\tau = \frac{\varepsilon}{|-C_0|} \tag{3.7}$$

ε is a parameter that is kept constant throughout the simulation. The value of C_0 for the RR is determined, as discussed earlier from the weighted average of all the c_{is} of the different reactions in the system: C_0 = $\sum_{i=1}^{M} \left(\frac{a_i(\mathbf{x})}{a_0(\mathbf{x})}\right) c_i$. This expression is indeed quite simple, but it suffers from the drawback that it only depends on C_0 and not on x_0 : the amount of the hypothetical species *A*. As will be seen for the subsequent examples, x_0 appears in the denominator of the expression for τ for other RRs, and thus serves to modulate and reduce the value of τ. With the absence of x_0 in Eq. (3.7), the value of τ tends to be somewhat high, especially for the beginning few steps of the simulation. Thus the RR $A \longrightarrow B$, despite its simplicity, is not a good choice for doing simulations with the RRA-τ method.

(ii) The RRA-*N* method : For the RR $A \longrightarrow B$, the expression for $\nabla a_0(\mathbf{x})$ is C_0. Therefore, using Eq. (3.7), one gets the following expression for N_0 :

$$N_0 = \frac{\varepsilon \prime a_0(\mathbf{x})}{C_0} \tag{3.8}$$

Like for the RRA-τ method, this expression suffers from the absence of x_0. For other RRs, as will be shown later, x_0 appears in the denominator in the right-hand side of Eq. (3.8), and serves to regulate the value of N_0. The absence of x_0 in Eq. (3.8) leads to somewhat large changes in N_0 for the beginning few steps of the simulations for any given chemical system. Thus, as for the RRA-τ method, $A \longrightarrow B$ is not a good choice as the RR for the RRA-*N* method.

$2A \longrightarrow B$ as the Representative Reaction (RR) :

(i) The RRA-τ method : Since in case of this RR the propensity function is given by

$$a_0 = \Big(\frac{x_0(x_0-1)}{2!}\Big)C_0 \tag{3.9}$$

to get an expression for x_0, we solve the resulting quadratic equation which yields two roots for x_0, of which we choose the root :

$$x_0 = \frac{C_0 + \sqrt{C_0^2 + 8a_0C_0}}{2C_0} \tag{3.10}$$

to avoid the possibility of negative values for x_0. Here, the value of τ, determined from the GASA leap condition, is

$$\tau = \frac{\varepsilon}{|-C_0(2x_0-1)|} \tag{3.11}$$

This expression gives rise to acceptable values of τ, because of the presence of x_0 in the denominator.

As will be shown in the examples discussed in the next section, the use of the RRA-τ method with $2A \longrightarrow B$ as the RR provides results which equal or improve upon the accuracy and reliability of the other approximate methods.

(ii) The RRA-*N* method : The value of N_0, determined from this method, using Eq. (3.6), is

$$N_0 = \frac{\varepsilon / a_0(\mathbf{x})}{C_0(2x_0-1)} \tag{3.12}$$

where the x_0 is calculated as discussed earlier. As with the RRA-τ method, this value of N_0 provides results that are quite acceptable, as evidenced by the results for the examples discussed in the subsequent sections of the article.

$3A \longrightarrow B$ as the representative reaction (RR) :

(i) The RRA-τ method : Now, as before, considering x_0 as the number of reactant molecules for the hypothetical reactant species *A* for the RR $3A \longrightarrow B$, we have the

propensity,

$$a_0 = \Big(\frac{x_0(x_0-1)(x_0-2)}{3!}\Big)C_0 \tag{3.13}$$

So to get an expression for x_0 for this case, one could use a subroutine to calculate the cubic roots but it makes our algorithm computationally more complicated, losing the simplicity which is one of the criteria for choosing the RR. Instead of that, we can neglect the lower order terms for x_0 in Eq. (3.13) to get a final expression for x_0. The expression of τ for this RR is derived to be

$$\tau = \frac{\varepsilon}{|-3C_0(3x_0^2 - 6x_0 + 2)/6|} \tag{3.14}$$

This expression for τ leads to a problem that is the opposite of the one faced when using $A \longrightarrow B$: here, the quadratic dependence of τ on x_0 in the denominator leads to values that are too small in size as compared to values obtained for the RR $2A \longrightarrow B$.

In other words, not much acceleration over the exact stochastic simulation method (SSA) is observed in this case. The reliability of the algorithm with this RR is thus not in doubt, but the essential purpose of making an accelerated algorithm is lost for this case.

A similar problem occurs for the cases $4A \longrightarrow B$, $5A \longrightarrow B$ and other higher order versions of this type of RR. The term x_0 begins to appear in higher and higher orders in the denominator, thereby making the size of the jumps smaller and smaller, and reducing the efficiency of the accelerated algorithm.

(ii) The RRA-*N* method : The value of N_0, determined from this method, using Eq. (3.6) is

$$N_0 = \frac{2\varepsilon\prime a_0(\mathbf{x})}{C_0(3x_0^2 - 6x_0 + 2)} \tag{3.15}$$

where the x_0 is calculated as discussed in the earlier method. In this case too, as for the RRA-τ method, the presence of x_0^2 in the denominator leads to values of N_0 that are smaller than those obtained from the RR, $2A \longrightarrow B$. Hence the efficiency of the method is reduced. Again, as for the RRA-τ method, higher powers of x_0 appear in the denominator for the cases $4A \longrightarrow B$, $5A \longrightarrow B$ and so on, thereby making the algorithms even less efficient, and thus defeating the purpose of making an accelerated algorithm.

$A + B \longrightarrow C$ **as the representative reaction (RR) :**

In case of this particular RR, we have x_0 as the number of the reactant molecules for the hypothetical reactant species *A* and y_0 as the number of the reactant molecules for the hypothetical reactant species *B*. Thus, the corresponding propensity function is,

$$a_0 = C_0 x_0 y_0 \tag{3.16}$$

Even though the expression looks simple, this is a case of a single equation with two unknown variables, namely x_0 and y_0. Therefore, the values of x_0 and y_0 cannot be determined independently in terms of the known a_0**(x)** and C_0 values.

Naturally, a similar problem also occurs if we take our RR to be $A + 2B \longrightarrow C$ or $2A+B \longrightarrow C$ or $A+B+C \longrightarrow D$ or any other variant of a bimolecular or tri-molecular or any other multimolecular reaction.

Our analysis thus indicates that the most appropriate RR for doing the simulation, the one that combines simplicity and efficiency most effectively, is the RR $2A \longrightarrow B$. This is therefore our chosen reaction for the representative reaction approach. Its efficiency will be revealed in the three illustrative examples that are described later in the article.

3.2.5 Steps for the implementation of the RRA-τ and RRA-*N* methods

Based on the discussion earlier, the implementation steps for the RRA methods, for the RR : $2A \longrightarrow B$, are outlined as follows :

- **Step 1 :** input the initial number of species, rate constants of the constituent reactions; initialize the counters and the random number generators to a seed value.

- **Step 2 :** calculate the propensity functions : $\{a_1, ..., a_M\}$, the sum of the propensity functions : a_0**(x)** = $\sum_{i=1}^{M} a_i$**(x)**, the weighted rate constant : C_0 = $\sum_{i=1}^{M} \left(\frac{a_i\mathbf{(x)}}{a_0\mathbf{(x)}}\right) c_i$

• **Step 3 :** calculate the total number of species present : $x_0 = \frac{C_0+\sqrt{C_0^2+8a_0C_0}}{2C_0}$

• **Step 4 :** calculate the tau step :

(i) For RRA-τ method : $\tau = \frac{\varepsilon}{|-C_0(2x_0-1)|}$ if the tau step is less than or equal to $2/a_0$**(x)** then perform the SSA otherwise continue with the RRA-τ method

(ii) For RRA-*N* method : $\tau = \frac{N_0}{a_0\mathbf{(x)}}$, where the total number of reactions $N_0 = \frac{\varepsilon/a_0\mathbf{(x)}}{C_0(2x_0-1)}$; if N_0 is less than or equal to 1, then perform the SSA otherwise continue with RRA-*N* method

• **Step 5 :** calculate the expected number of occurrences, the n_{is} for the individual reactions $n_i = a_i\tau$

• **Step 6 :** use the Poisson random number generator to find the k_{is}, the actual number of occurrences for the individual reactions; k_i = poidev(n_i, iseed)

• **Step 7 :** make the necessary changes in the species population using the appropriate stoichiometric parameters and reaction numbers

• **Step 8 :** go to Step 2

This algorithm is further outlined in the flowchart below :

Flowchart

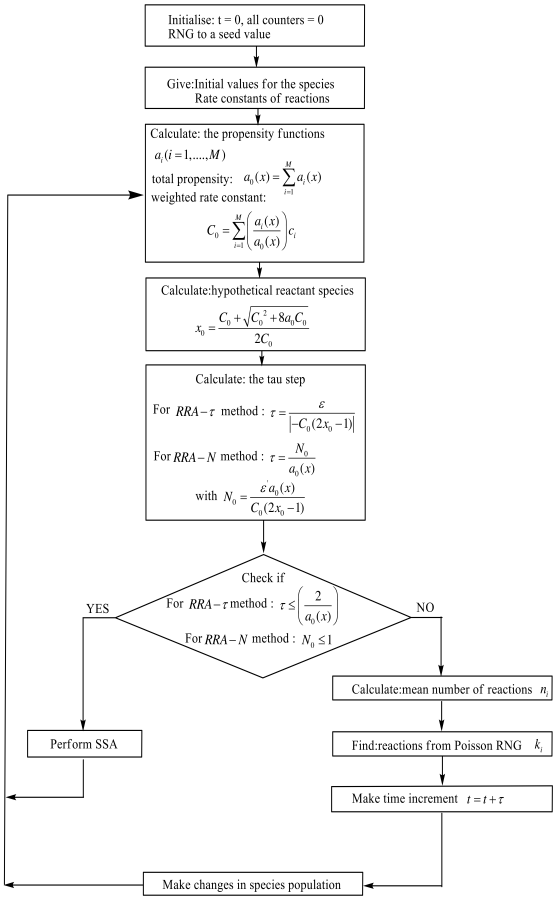


Figure 3.1: The flowchart for the implementation of RRA methods.

3.3 Optimization of the Values of the Parameters ε and $\varepsilon\prime$

It was determined in the previous section that the most appropriate choice of RR for both the RRA-τ and the RRA-*N* approaches is the reaction $2A \longrightarrow B$. For this RR, the expressions derived for τ and N_0 for the two methods depend on the parameters ε and $\varepsilon\prime$ respectively. Discussed in this subsection is the choice of the most appropriate values of ε and $\varepsilon\prime$ for the two methods. To this end, the two methods were tested on a four reaction model : a chemical system that has been demonstrated to be successfully simulated by GASA.[7] The system comprises of the following reactions :

$$R_1 : X_1 \xrightarrow{c_1} \phi$$

$$R_2 : 2X_1 \xrightarrow{c_2} X_2$$

$$R_3 : X_2 \xrightarrow{c_3} 2X_1$$

$$R_4 : X_2 \xrightarrow{c_4} X_3$$

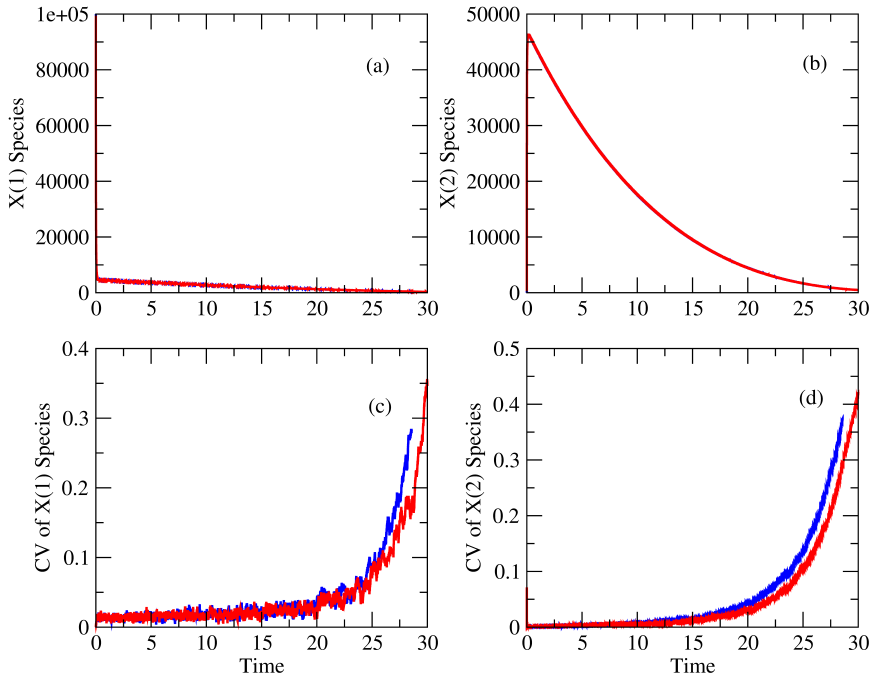
where X_1, X_2, X_3 are the species participating in this reaction system, with c_1, c_2, c_3, c_4 as the rate constants of the corresponding reactions.

Table 3.1: The values of the rate constants of the reactions and the initial species population for the four reaction model.

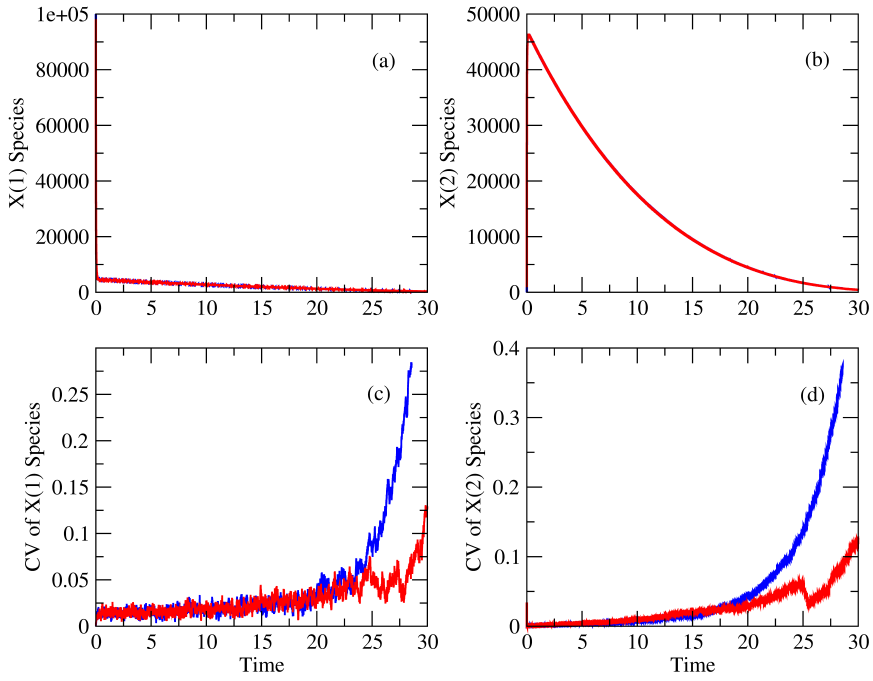
Parameters used in the simulation	Numerical values of the parameters
c_1	1.0
c_2	0.002
c_3	0.5
c_4	0.04
X_1	10000
X_2	0
X_3	0

The RRA-τ and the RRA-N methods were tested for this four reaction model system for different values of ε and $\varepsilon\prime$, respectively. The results are discussed in the following:

(i) The RRA-τ method : In the following Figures (3.2-3.4), the results obtained by the application of the RRA-τ algorithm on the four reaction model, as well as the comparison to the results obtained by the exact stochastic simulation algorithm (SSA) are shown.

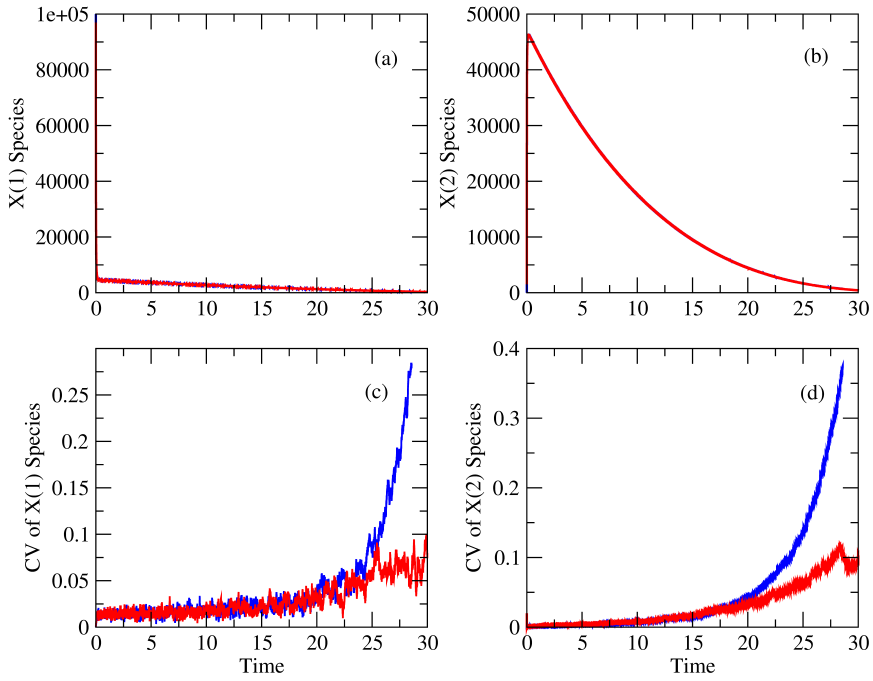


(a) $\varepsilon = 0.03$

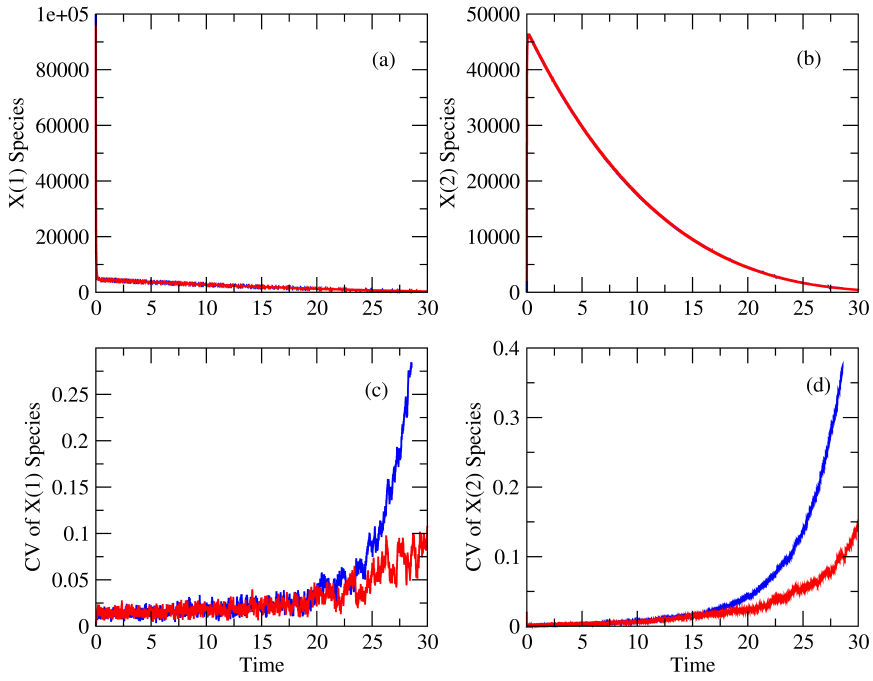


(b) $\varepsilon = 0.09$

Figure 3.2: The trajectories of the means [(a)and(b)] and CVs [(c)and(d)] of X_1 and X_2 species using SSA (blue line) and RRA-τ (red line) with (a) $\varepsilon = 0.03$ and (b) $\varepsilon = 0.09$



(a) $\varepsilon = 0.15$



(b) $\varepsilon = 0.2$

Figure 3.3: The trajectories of the means [(a)and(b)] and CVs [(c)and(d)] of X_1 and X_2 species using SSA (blue line) and RRA-τ (red line) (a) $\varepsilon = 0.15$ and (b) $\varepsilon = 0.2$

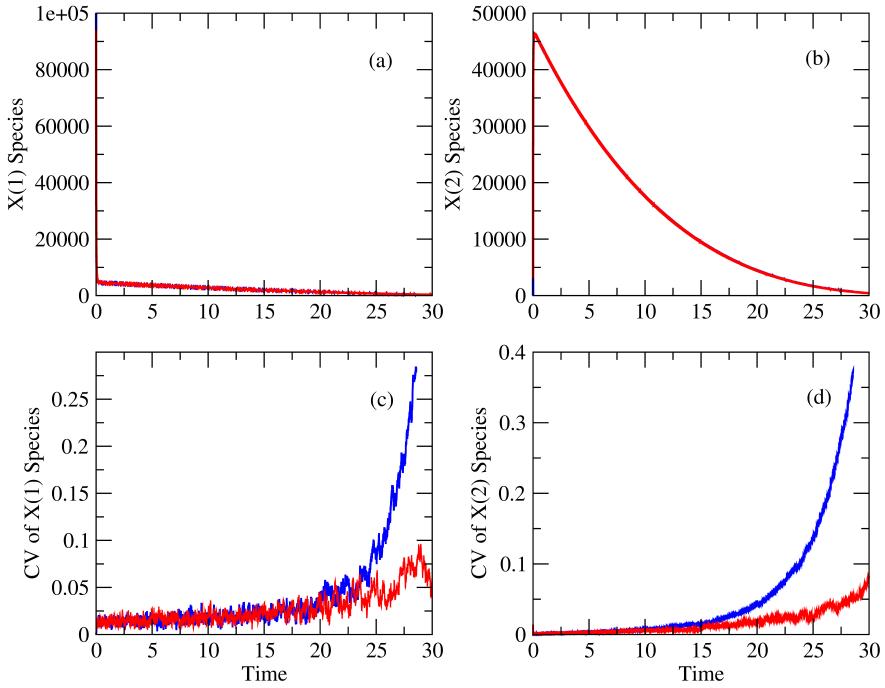
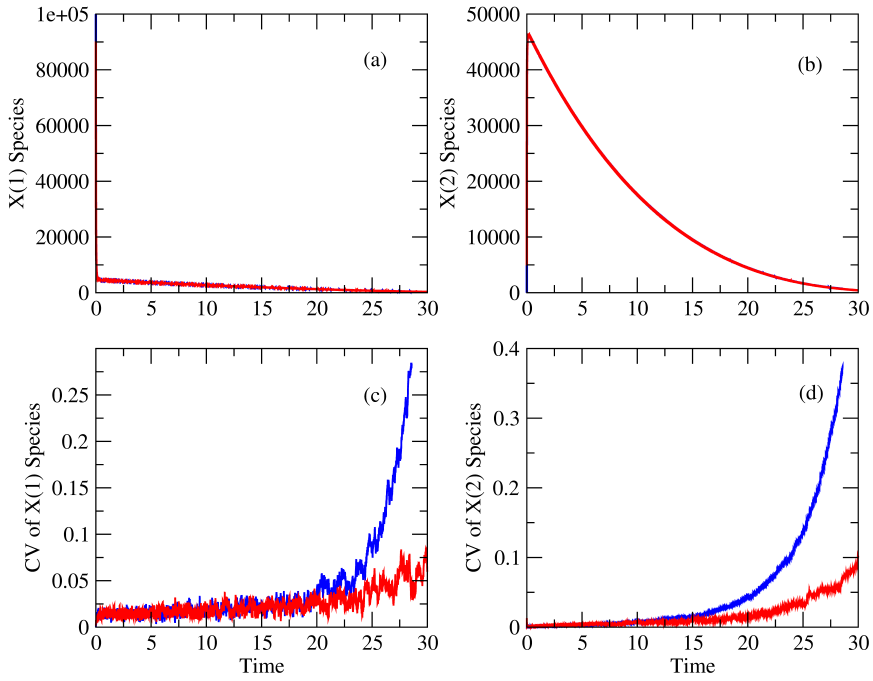


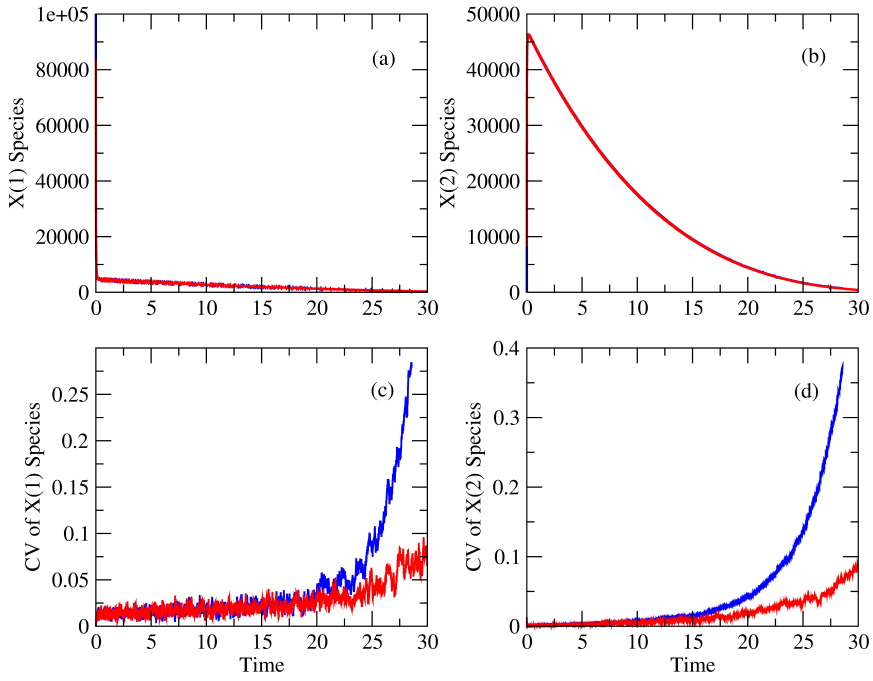
Figure 3.4: The trajectories of the means [(a)and(b)] and CVs [(c)and(d)] of X_1 and X_2 species using SSA (blue line) and RRA-τ (red line) with $\varepsilon = 0.3$.

The simulations were run 10 times for both the SSA and the RRA, using different seed values for the random number generators, and the mean values of the chemical species were calculated. The mean values obtained for the species X_1 and X_2 with time, using the SSA and the RRA-τ, as well as the values of the coefficient of variation (CV), which is defined as the ratio of the standard deviation to the mean for the species, for both the SSA and the RRA-τ are shown in the five figures. The values of ε employed are as follows: 0.03, 0.09, 0.15, 0.2, and 0.3 for the curves shown in Figures respectively. The comparison of the mean and the CV of the results from the RRA-τ method with the SSA indicates that the alteration of the value of ε from 0.09 to 0.3 leads to almost identical results and, as compared to the values obtained from SSA, to fairly accurate results. The accuracy of the results is seen to be marginally more accurate for $\varepsilon = 0.03$, but this comes at the cost of a slower simulation. Hence, keeping the balance of accuracy and efficiency in mind, the value chosen for ε in evaluating the RRA-τ method and comparing it to SSA and the other approximate methods, GASA and G-P,[29] is 0.2.

(ii) The RRA-*N* method : Figures (3.5-3.7) show the results of the simulations done for the four reaction model with different values of $\varepsilon\prime$ employed for the RRA-*N* method. The values of $\varepsilon\prime$ employed are as follows : 0.48, 0.80, 1.28, 1.6, and 3.2 for the curves shown in Figures 3.5-3.7, respectively.

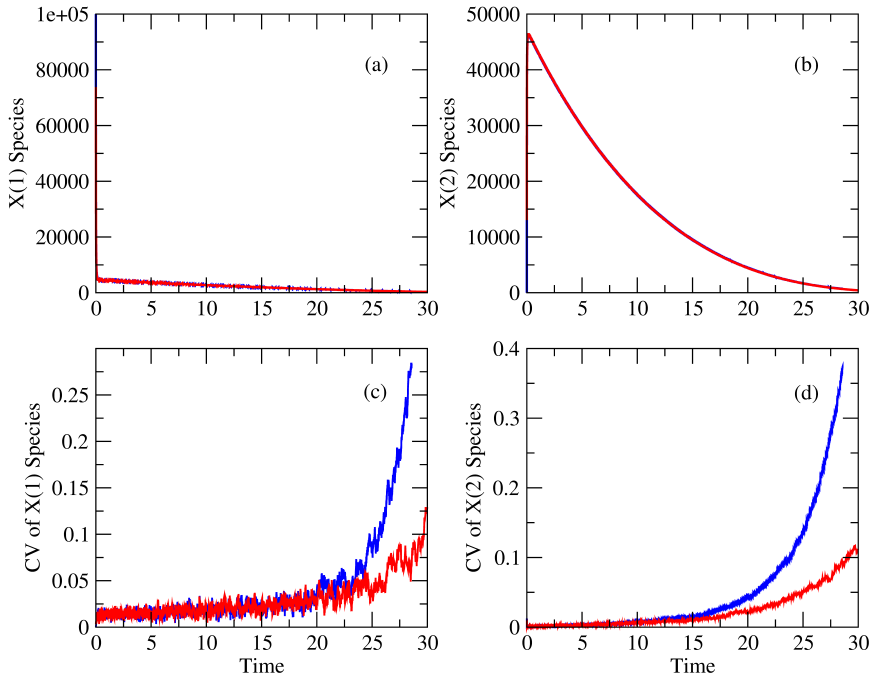


(a) $\varepsilon = 0.48$

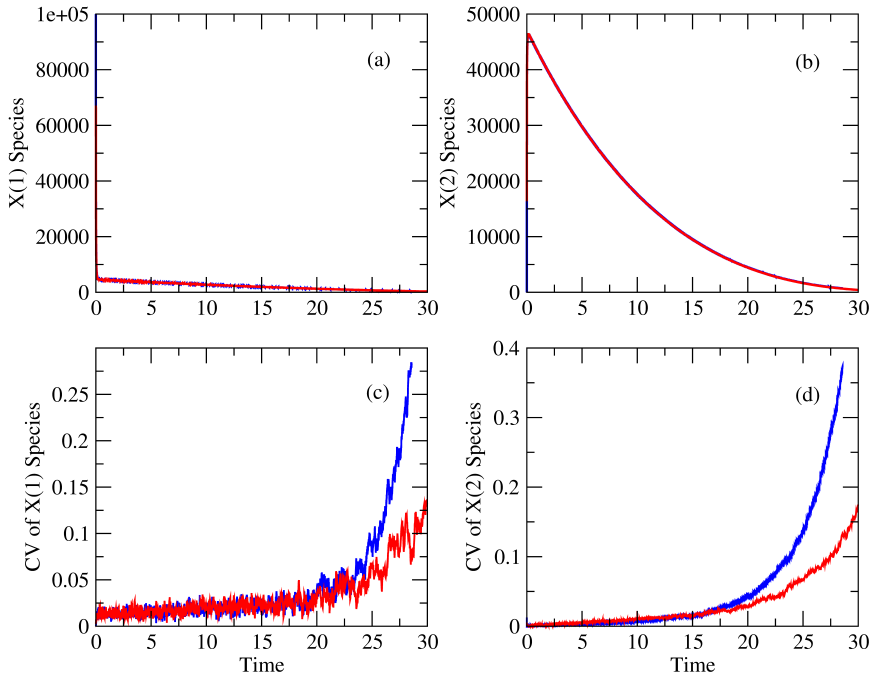


(b) $\varepsilon = 0.80$

Figure 3.5: The trajectories of the means [(a)and(b)] and CVs [(c)and(d)] of X_1 and X_2 species using SSA (blue line) and RRA-N (red line) with (a) $\varepsilon = 0.48$ and (b) $\varepsilon = 0.80$.



(a) $\varepsilon = 1.28$



(b) $\varepsilon = 1.6$

Figure 3.6: The trajectories of the means [(a)and(b)] and CVs [(c)and(d)] of X_1 and X_2 species using SSA (blue line) and RRA-N (red line) with (a) $\varepsilon = 1.28$ and (b) $\varepsilon = 1.6$.

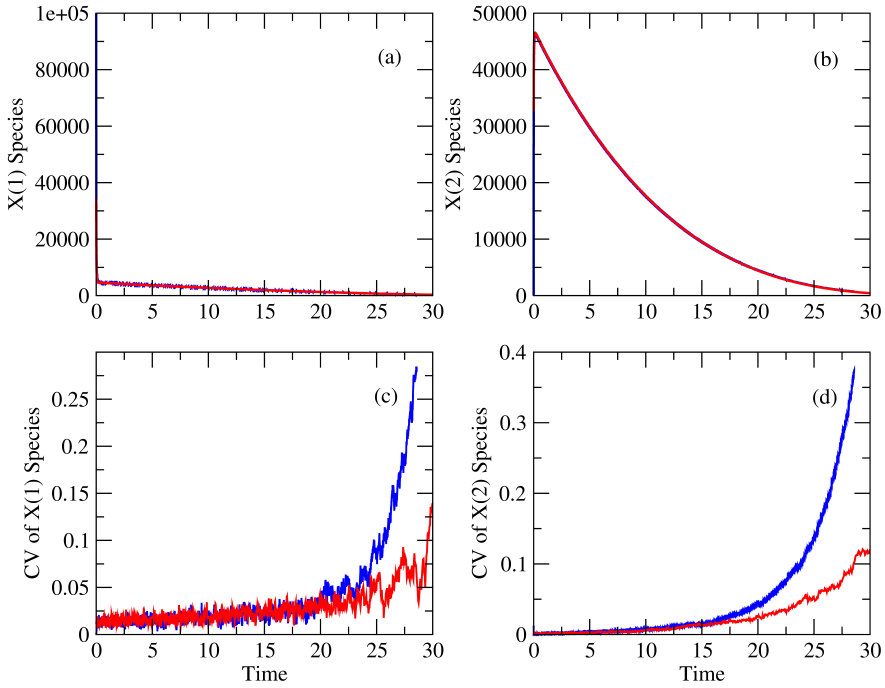


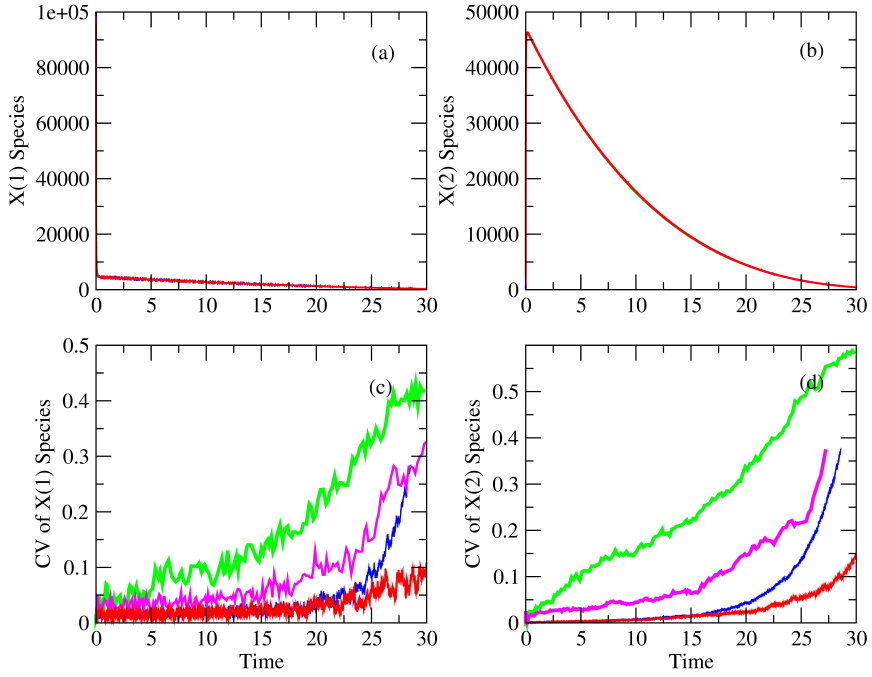
Figure 3.7: The trajectories of the means [(a)and(b)] and CVs [(c)and(d)] of X_1 and X_2 species using SSA (blue line) and RRA-N (red line) with ε = 3.2.

As for the RRA-τ method, both the mean and the CV of the probability distributions for the species X_1 and X_2 were plotted for the different cases, and the values are compared to the SSA. The figures indicate that there is little difference in the curves upon changing the value of $\varepsilon\prime$ from 0.48 to 3.2. However, what is affected is the size of the first jump : from $\varepsilon\prime$ 1.6 onward, it was seen that the change in the values of X_1 and X_2 was quite significant for the first step. Hence, to balance efficiency and reliability of the method, the optimal value of $\varepsilon\prime$ for the RRA-N method was chosen to be 1.28.

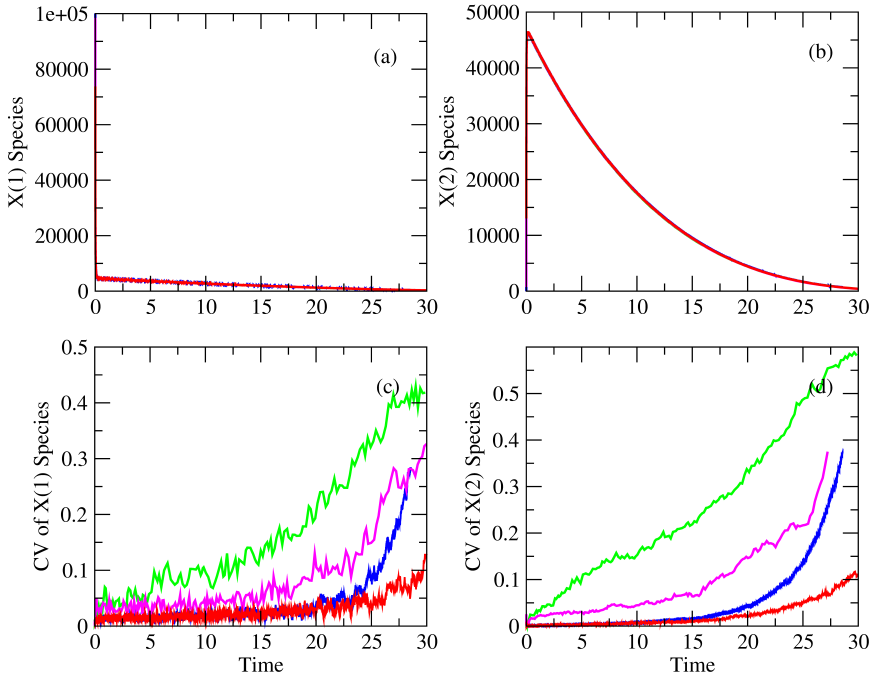
Hence the optimized parameters for the two methods are ε = 0.2 for the RRA-τ method and $\varepsilon\prime$ = 1.28 for the RRA-N method. The efficiency and reliability for the two RRA methods with these optimized values of ε and $\varepsilon\prime$ will be evaluated for three separate examples discussed in the next few sections.

3.4 The Four Reaction Model : Comparison of RRA-τ and RRA-N to Other Accelerated Methods

The four reaction model system was discussed in the previous section in the context of the optimization of the values of ε and $\varepsilon\prime$ for the RRA-τ and the RRA-N methods, respectively. Discussed in this section is the comparison, for the same four reaction model, of the RRA-τ and the RRA-N methods with two other accelerated methods: GASA and the G-P method. It is noted here that both GASA and G-P also employ the parameter $\varepsilon\prime$ and the value that the methods recommend is 0.03. For all the cases discussed here onward, this standard value of ε = 0.03 has been employed for all the simulations using these two approximate methods.



(a) RRA-τ (red curve)



(b) RRA-N (red curve)

Figure 3.8: The trajectories of the means [(a),(b)] and CVs [(c),(d)] for the probability distributions of the species X_1 and X_2 using SSA (blue curve), GASA (green curve), G-P (magenta curve), RRA-τ and RRA-N (both red curves) for the case of the four reaction model.

Figure 3.8 shows the comparison of the mean and the CV of the probability distributions for the species X_1 and X_2 obtained from running simulations with the four different methods : the exact SSA (shown in blue), GASA (shown in green), G-P (shown in magenta),the RRA-τ and the RRA-N (both shown in red). As the figure indicates, all three accelerated methods work quite well in predicting the mean of the probability distributions for the two species X_1 and X_2. However, there is a distinct difference in performance when it comes to the second moment: it is seen from the CV curves (Figs. 3.8a (c) and (d)) that GASA provides results that are less accurate than the RRA-τ method and the G-P. A similar result is obtained when comparing the RRA-*N* method to GASA and G-P as (Figs. 3.8b (c) and (d)) indicate. However, a comparison of the average time taken to run the simulations for the three different algorithms indicates that GASA and G-P hold a distinct advantage over the two RRA methods. The average CPU time taken when employing the different methods is shown in Table 3.2 (which also collects the corresponding information for the other examples that have been studied). For the case

Table 3.2: The averaged values of the CPU time (in seconds) taken by different simulation methods for various chemical systems.

Chemical Systems	SSA	GASA	G-P	RRA-τ	RRA-*N*
Four Reaction Model	11.439	0.036	0.053	1.622	0.300
Oregonator Model	52.524	0.020	1.494	8.893	1.492
Viral Infection Model	140.090	0.074	0.549	147.750	72.509

of the four reaction model, it was determined that the average time taken to do a GASA simulation was 0.036 s; the G-P, 0.053 s; the RRA-*N*, 0.300 s; the RRA-τ, 1.622 s; and the SSA, 11.439 s. Thus, while the two RRA methods do provide a distinct acceleration in comparison to the SSA, as evidenced by the CPU time taken, they are not as efficient as GASA and G-P. The explanation for this is the fact that, while the two RRA methods do accelerate the system, they take smaller jumps in time than GASA and the G-P methods. This is because of the regulating variable x_0 in the denominator for the RR $2A \longrightarrow B$, which serves to decrease the size of the jump for the RRA methods. While

this appears to be a disadvantage, in terms of efficiency, for relatively simple systems such as the four reaction model considered here, this is actually a significant advantage for more complicated systems, two examples of which will be discussed in the next two sections. As will be seen, the sacrifice of a minor loss in efficiency because of smaller step size is more than compensated by the gain in the accuracy of the simulations of the more complicated systems.

3.5 The Model of Oscillatory Reactions: The Oregonator Model

We show in this section the application of our RRA algorithms to the more complicated case of an oscillatory chemical system.[13] To correctly simulate oscillatory chemical systems is a significant challenge for an approximate accelerated method, because of the fluctuations in the species amounts with small changes in time. The model discussed here is the Oregonator which is a theoretical model for autocatalytic reactions. The system consists of the following set of reactions :

$$R_1 : \overline{X_1} + Y_2 \xrightarrow{c_1} Y_1$$

$$R_2 : Y_1 + Y_2 \xrightarrow{c_2} Z_1$$

$$R_3 : \overline{X_2} + Y_1 \xrightarrow{c_3} 2Y_1 + Y_3$$

$$R_4 : 2Y_1 \xrightarrow{c_4} Z_2$$

$$R_5 : \overline{X_3} + Y_3 \xrightarrow{c_5} Y_2$$

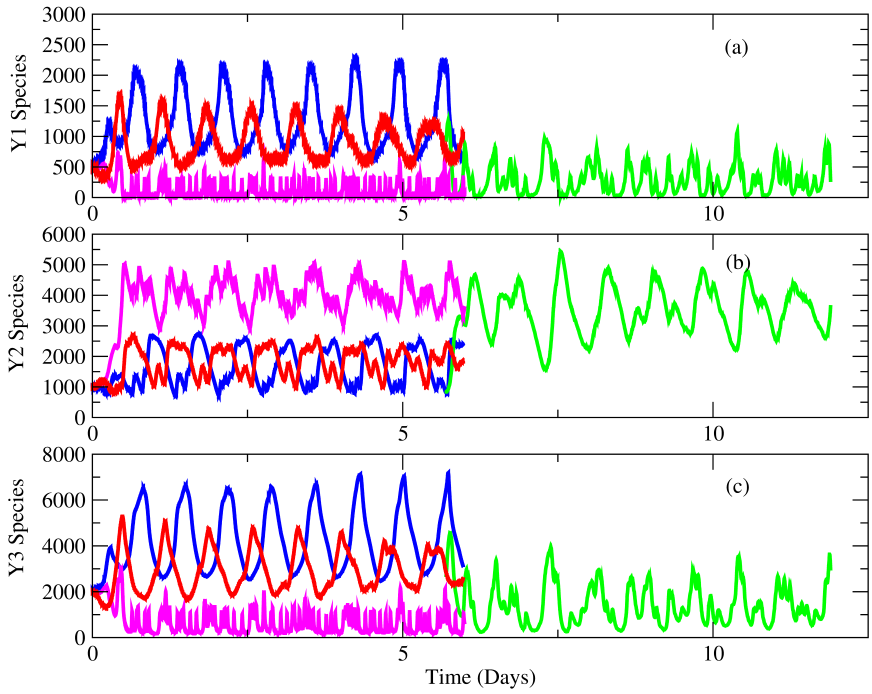
where Y_1, Y_2, Y_3 are the species participating in this reaction system, while $\overline{X_1}$, $\overline{X_2}$, $\overline{X_3}$ indicates that the molecular population level of these species is assumed to remain constant and c_1, c_2, c_3, c_4, c_5, as the rate constants of the above reactions.

The values of the different model parameters used in the simulation of the above reaction system are provided in the following table.

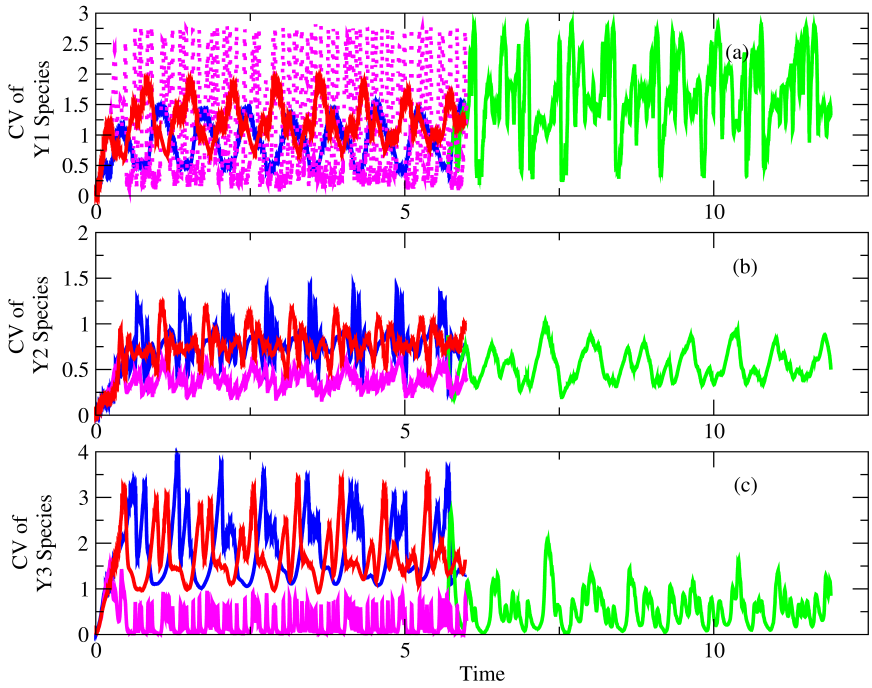
Table 3.3: The values of the rate constants of the reactions and the initial species population for the Oregonator model.

Parameters used in the simulation	Numerical values of the parameters
c_1	2.0
c_2	0.1
c_3	104.0
c_4	0.016
c_5	26.0
Y_1	500
Y_2	1000
Y_3	2000

The Oregonator has been investigated by the SSA, GASA,G-P, and the two RRA methods. Shown in Figures 3.9a (a-c) is the comparison of the mean values obtained for the species Y_1, Y_2 and Y_3 with time, using SSA, GASA, G-P, and RRA-τ. The CV values obtained for the Y_1, Y_2 and Y_3 are shown in Figure 3.9b (a-c). Likewise, Figures 3.10a and 3.10b show the mean and CV values, respectively, for SSA, GASA, G-P, and RRA-*N*.

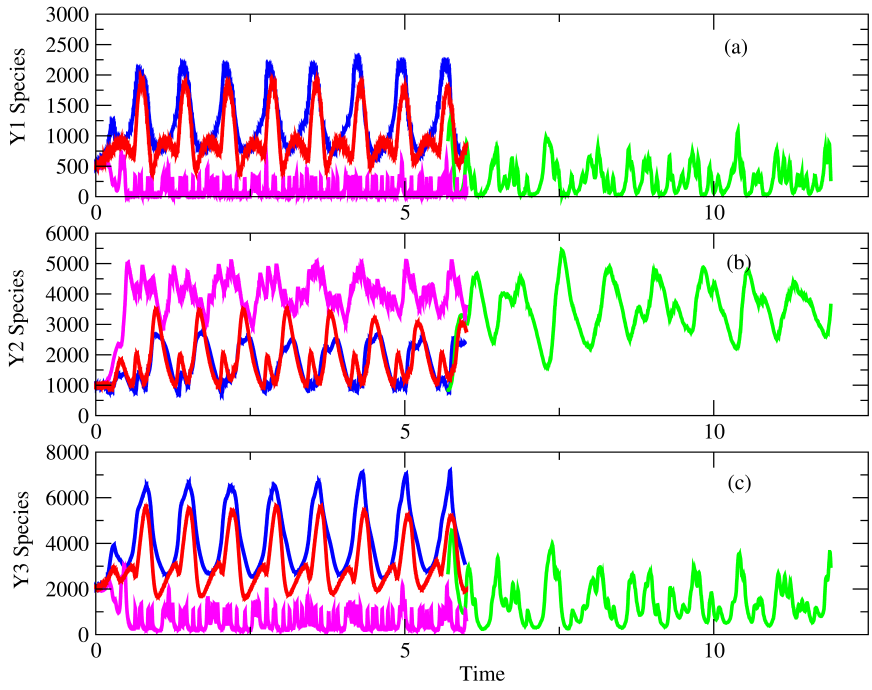


(a) means

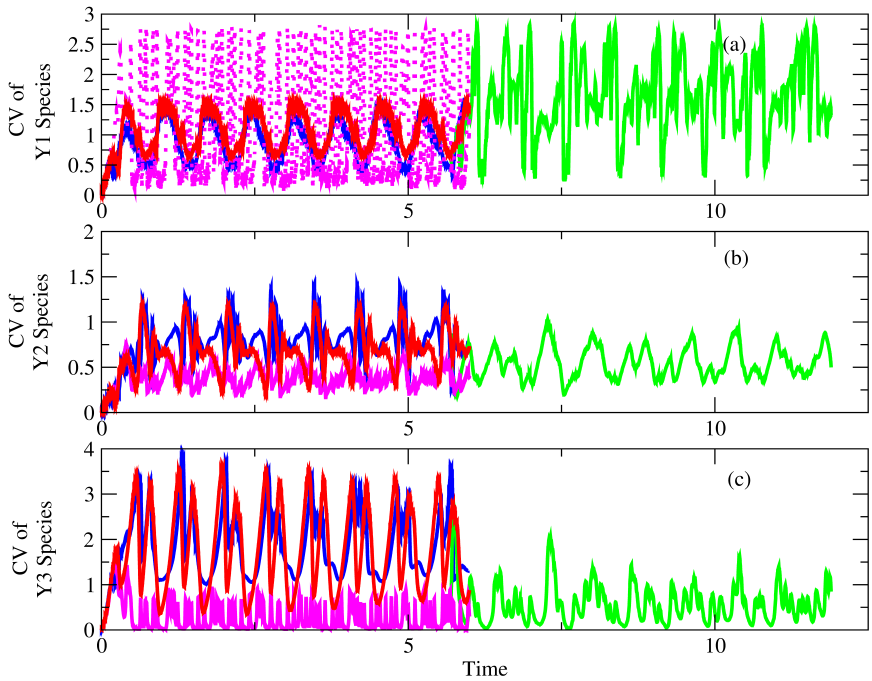


(b) CVs

Figure 3.9: The trajectories of the means and CVs [(a),(b),(c)] for the probability distributions of the species Y_1, Y_2 and Y_3 using SSA (blue curve), GASA (green curve), G-P (magenta curve), and RRA-τ (red curve) for the case of the Oregonator model.



(a) means



(b) CVs

Figure 3.10: The trajectories of the means and CVs [(a),(b),(c)] for the probability distributions of the species Y_1, Y_2 and Y_3 using SSA (blue curve), GASA (green curve), G-P (magenta curve), and RRA-N (red curve) for the case of the Oregonator model.

As is clear from the four figures, the two RRA methods do not give an exact correspondence to the SSA values, which is indeed very difficult to achieve for the challenging oscillatory systems, but it is also clear that the two RRA methods perform significantly better than GASA and G-P, not only in replicating the mean values but even in replicating the second moment (CV) of the probability distributions for the three species in question. In the case of GASA, it is found that the curves showing the mean and CV of the probability distributions for the different species appear to be laterally displaced from the other curves. This is because it was found that the first τ value obtained from the GASA simulations was very large : 5.6969 units, as compared to the first step for the other methods: G-P, 1.2079E-2 units; RRA-τ, 2.7177E-5 units; RRA-*N*, 1.7393E-4 units; and SSA, 1.5315E-6 units. Hence, though the subsequent steps for GASA are smaller in size, the first, large jump violates the leap condition and makes GASA inappropriate for simulating this class of chemical systems.

With regard to the average time taken for the simulations, the values in seconds are GASA 0.020, G-P 1.494, RRA-*N* 1.492, RRA-τ 8.893, and SSA 52.524 (see Table 3.2). GASA, of course, is quite fast in comparison to the others, but, as discussed earlier, it is also completely unreliable. Among the rest, it is seen that the RRA-*N* method performs as efficiently as G-P in terms of time taken, and as the curves indicate, is significantly more reliable in replicating both the first and the second moments of the probability distributions for the different species.

Overall, the comparison of efficiency and accuracy among the different algorithms for the challenging oscillatory model example provides definite evidence of the efficiency and reliability of the RRA methods, especially the RRA-*N* method.

3.6 Example Involving Species Fluctuating by Different Orders of Magnitude

In this section, we discuss the simulation of a chemical system where the chemical species fluctuate by different orders. As mentioned in the section Introduction, this belongs to the class of problems where accelerated methods have been found wanting in terms of combining accuracy and efficiency. The system discussed in this section has been studied and dis- cussed previously by Haseltine and Rawlings.[101] They observed that the three principal species: template, genome, and struct in the system varied by different orders of magnitude as the reactions proceeded over time : the amount of the species template fluctuates between 5 and 25, genome between 0 and 200, and struct between 100 and 12,000.

For this system, the simulations have been done with the SSA, GASA, G-P as well as with the two RRA methods. For each case, the simulations have been repeated ten times, and the values for the mean as well as for the coefficient of variation, CV, have been calculated.

This model, the reactions for which are discussed in the following, shows the mechanism of the infection of a cell by a virus. The chemical reactions are :

$$R_1 : \text{nucleotides} \xrightarrow{template} \text{genome}$$

$$R_2 : \text{nucleotides} + \text{genome} \longrightarrow \text{template}$$

$$R_3 : \text{nucleotides} + \text{aminoacids} \xrightarrow{template} \text{struct}$$

$$R_4 : \text{template} \longrightarrow \text{degraded}$$

$$R_5 : \text{struct} \longrightarrow \text{secreted/degraded}$$

$$R_6 : \text{genome} + \text{struct} \longrightarrow \text{secreted}$$

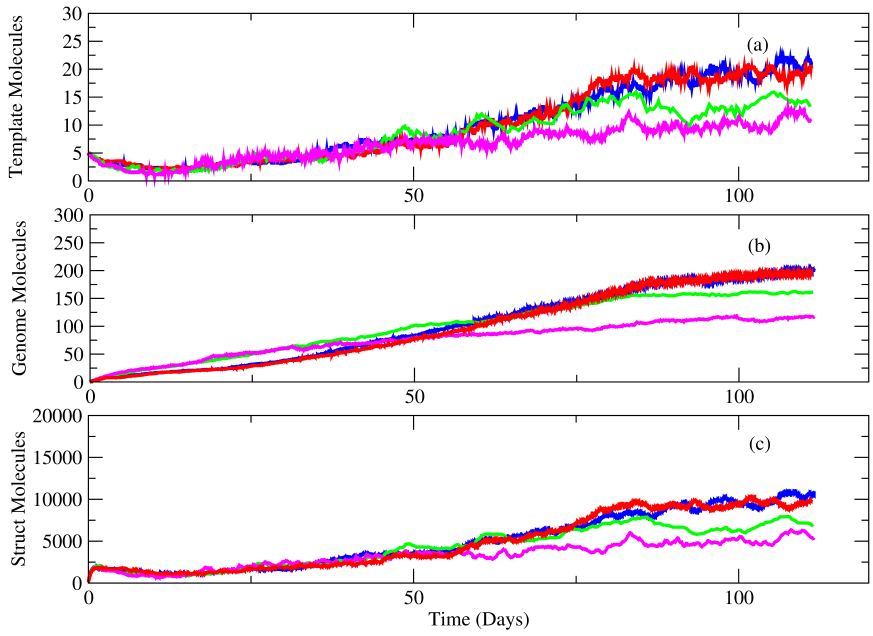
where genome and template are the nucleic acids, and struct is the viral structural protein. In this system, the molecular population number of the nucleotides and amino acids are

assumed to remain constant and the template is considered to act as a catalyst for the reactions R_1 and R_3.

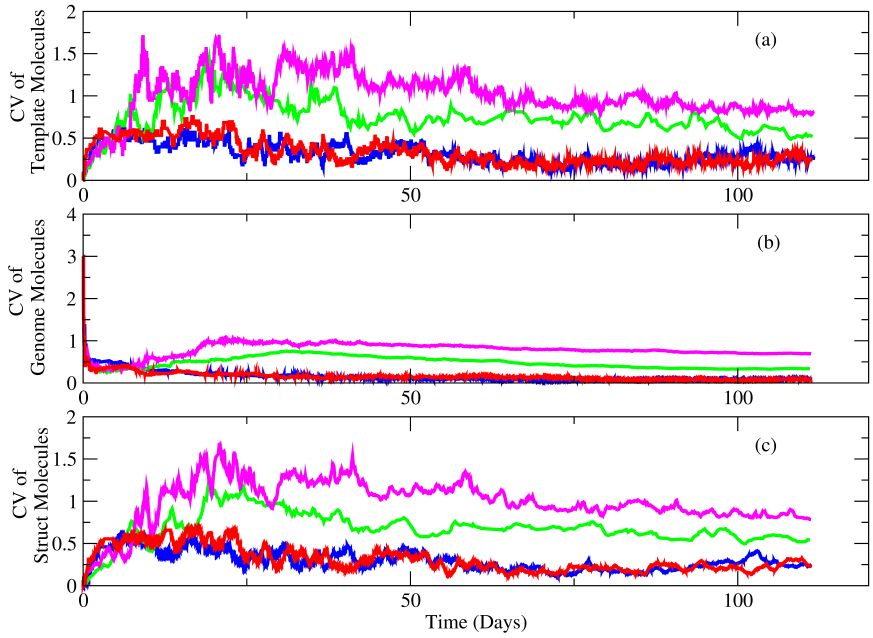
The values of the different model parameters used in the simulation of the above reaction system are provided in the table below.

Table 3.4: The values of the rate constants of the reactions and the initial species population for the viral infection model.

Parameters used in the simulation	Numerical values of the parameters
c_1	1.0 day^{-1}
c_2	0.025 day^{-1}
c_3	1000.0 day^{-1}
c_4	0.25 day^{-1}
c_5	1.9985 day^{-1}
c_6	7.5E-6 day^{-1}
Template	5
Genome	0
Struct	100

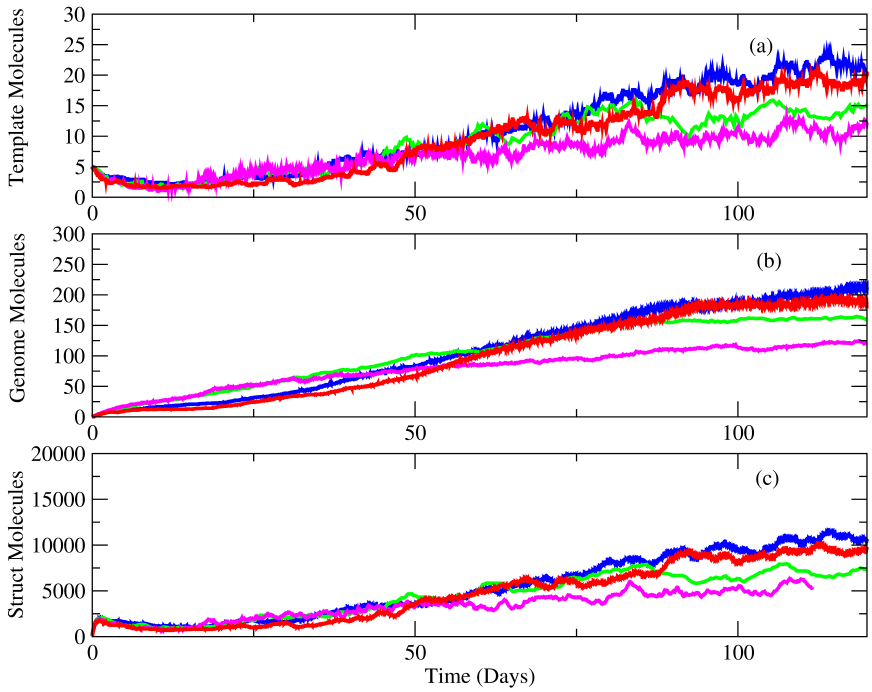


(a) means

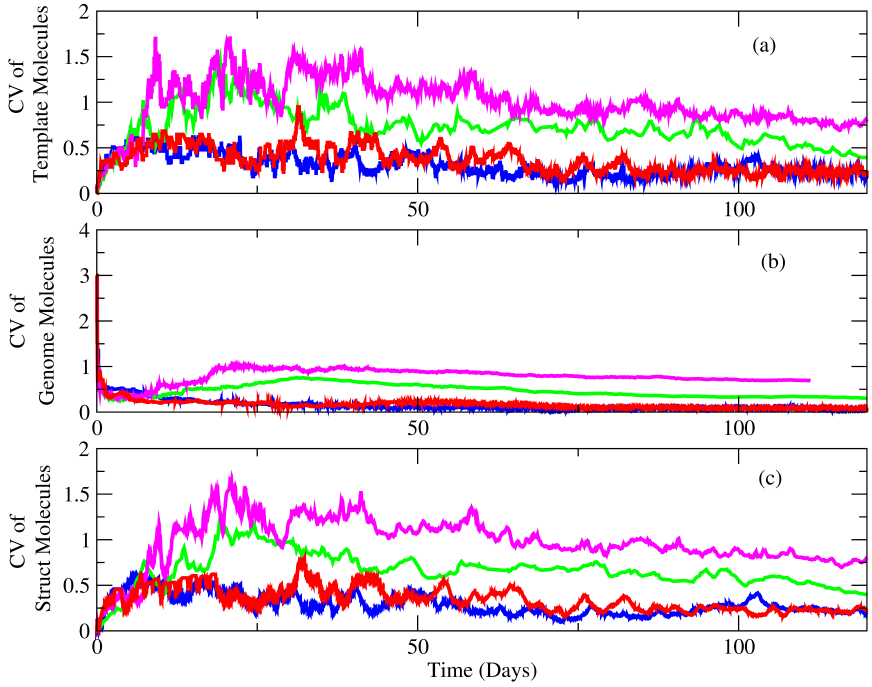


(b) CVs

Figure 3.11: The trajectories of the means and CVs [(a),(b),(c)] for the probability distributions of the species Template, Genome, and Struct using SSA (blue curve), GASA (green curve), G-P (magenta curve), and RRA-τ (red curve) for the case of the viral infection model.



(a) means



(b) CVs

Figure 3.12: The trajectories of the means and CVs [(a),(b),(c)] for the probability distributions of the species Template, Genome, and Struct using SSA (blue curve), GASA (green curve), G-P (magenta curve), and RRA-N (red curve) for the case of the viral infection model.

Figures 3.11a (a-c) show the change in the mean values for the three species: template, genome, and struct with change in time. Each figure has four curves, corresponding to the SSA (blue), GASA (green), the G-P (magenta), and the RRA-τ (red). Figure 3.11b (a-c) shows the corresponding CV values for the three different species as obtained from the four methods. Likewise, Figures 3.12a and 3.12b show the comparison of the SSA, GASA, and G-P with the RRA-*N* method.

As the figures indicate, the two RRA methods outstrip GASA and G-P in replicating both the mean and the CV of the three species in question. With regard to the average time taken, the values in seconds are as follows: GASA: 0.074, G-P: 0.549, RRA-*N* : 72.509, SSA: 140.090, and RRA-τ : 147.750. Hence, while RRA-τ does well to replicate the results of the SSA, it is unable to accelerate the system. However, GASA and G-P, while being much faster than the RRA methods, provide results for both the mean and the CV for the three species that are quite inaccurate. It is only the RRA-*N* method that succeeds in providing accuracy while also managing to accelerate the simulation of the system at least by a factor of 2. Hence, for this difficult and very relevant set of problems, it is demonstrated that the RRA methods, especially the RRA-*N* method, can be effective and reliable substitutes for the exact SSA algorithm.

Overall, what the application of the two RRA methods to the three examples indicates is that the RRA methods, in general, take more steps, and thus slightly longer times, in comparison to the other accelerated methods discussed here : GASA and G-P. However, this, in turn, leads to a considerable improvement in the accuracy and reliability of the methods. The best RRA method that combines the qualities of reliability and efficiency is found to be the RRA-*N* method. This, combined with the overall simplicity of the approach and the resultant algorithm, makes it a promising method for stochastically simulating different types of chemical systems.

3.7 Conclusions

In the work described in this article, we have endeavored to develop new approximate methods for conducting stochastic simulations on chemical systems. The new methods are based on the concept of treating $2A \longrightarrow B$ as a single representative reaction for the system : the representative reaction approach (RRA). Two methods have been proposed based on this approach: RRA-τ and the RRA-*N*, and the application of these methods to three different chemical systems indicates that the two methods, especially the RRA-*N*, perform creditably in combining accuracy and efficiency in simulating the mean and the CV of the probability distributions of the different species, in comparison to other approximate methods such as GASA and G-P method. It is to be noted that the new approximate methods take smaller jumps in time than the other approximate methods that have been proposed. However, the subsequent loss in efficiency is more than compensated for by the concurrent increase in accuracy of the simulations. This is especially relevant when one wants to simulate more challenging and complicated systems, as demonstrated in the article for an oscillatory model system as well as for a system where the species concentrations vary by different orders of magnitude. The methods have the added virtue of giving rise to very simple and straightforward algorithms. Indeed, as can be observed from the flowchart provided for the implementation of the RRA methods, they are simpler algorithms than GASA, for which it is necessary to calculate all the partial derivatives of the propensity functions with respect to the different species present in the system and then to make a stoichiometric matrix to calculate the denominators corresponding to each reaction involved in the system;[7] such calculations are unnecessary in our method. In terms of potential drawbacks for our method, it should be mentioned that for reactions where species approach molecular populations close to zero, the RRA methods, like other approximate accelerated methods[7,84,89,101] may begin to provide negative molecular populations. Work is currently in progress to address this issue. Overall, it is clear that the RRA methods, especially the RRA-*N* method, provides

a simple, easy way to successfully simulate a wide variety of different chemical systems over long periods of time, and, as such, should find wide ranging applicability in the fields of chemistry and bio-chemistry.

CHAPTER 4

Solving the Problem of Negative Populations in Approximate Accelerated Stochastic Simulations Using the Representative Reaction Approach

Twinkle, twinkle, quasi star
Biggest puzzle from afar
How unlike the other ones
Brighter than a billion suns
Twinkle, twinkle, quasi-star
How I wonder what you are

-George Gamow

Abstract

Methods based on the stochastic formulation of chemical kinetics have the potential to accurately reproduce the dynamical behavior of various biochemical systems of interest. However, the computational expense makes them impractical for the study of real systems. Attempts to render these methods practical have led to the development of accelerated methods, where the reaction numbers are modeled by Poisson random numbers. However, for certain systems, such methods give rise to physically unrealistic negative numbers for species populations. The methods which make use of binomial variables, in place of Poisson random numbers, have since become popular, and have been partially successful in addressing this problem. In this manuscript, the development of two new computational methods, based on the representative reaction approach (RRA), has been discussed. The new methods endeavor to solve the problem of negative numbers, by making use of tools like the stochastic simulation algorithm and the binomial method, in conjunction with the RRA. It is found that these newly developed methods perform better than other binomial methods used for stochastic simulations, in resolving the problem of negative populations.

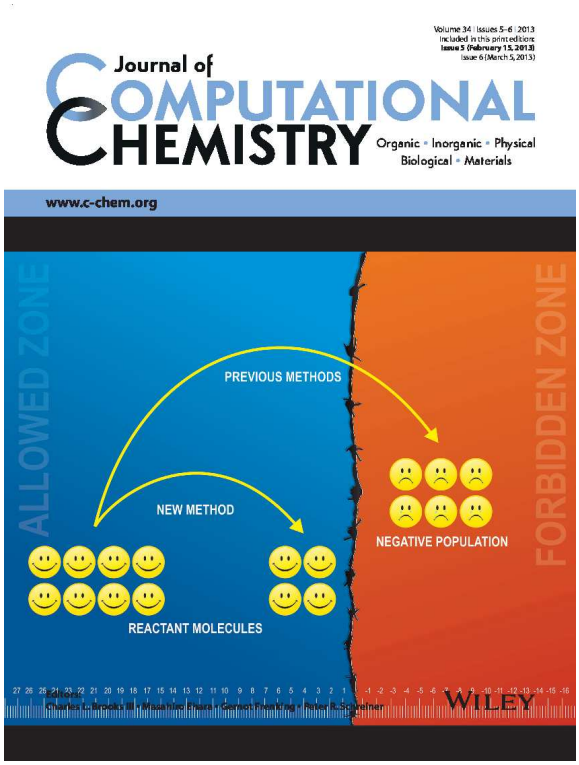


Figure 4.1: Artistic imagination of solving the problem of negative numbers in stochastic simulations. This image appeared on the inner cover of the Journal of Computational Chemistry.

4.1 Introduction

Cellular interactions are of considerable interest for the understanding of chemical dynamics of various biological processes. The state of the art experimental techniques in cell biology attempt to resolve the temporal dynamics in these systems, but, considering the complexity of these processes, computational modeling is of great importance. These interactions can be modeled as a set of chemical reactions. As the interactions are driven by species that are present in low copy numbers, the dynamical behavior is governed by the molecular fluctuations.[33, 120, 126] The deterministic model, based on the reaction rate equations, simulates the time behavior, but fails to capture the intrinsic noise in such systems. Also, the deterministic model treats time evolution as a continuous process and does not take into account the discrete nature of the reactant species. These issues

are effectively handled by the alternative: probabilistic model for simulating biological systems.

The probabilistic kinetic Monte Carlo method[54] : the Doob-Gillespie algorithm,[117, 118, 28, 13] is widely used nowadays for simulating the time behavior of reactants, intermediates and products in a reaction network. In the literature, this algorithm is known as the stochastic simulation algorithm (SSA).[13] In the SSA, a single reaction out of several reactions from the entire reaction network is allowed to occur in each time increment and the simulation is continued until some desired time has been achieved. Although this has proved an effective approach for providing an accurate depiction of the change of the reactant, intermediate and product species during the reaction, the SSA suffers from the significant drawback in that the simulation of any realistic biochemical system, which consists of many more reactions, requires considerable computational time. To address this issue, several improvements to the SSA have been proposed, which include the next reaction method,[73] optimized direct method,[74] and sorting direct method,[75] but these, too, have not been successful at providing sufficient reduction in computational time for the simulations. To incorporate time delays, along with discreteness and noise, the delay SSA[81] has been proposed. This algorithm is essentially a generalized version of the SSA, accounting for time delays that are not considered in the SSA.

To circumvent this difficulty of the SSA, Gillespie designed the Gillespie's approximate stochastic algorithm (GASA),[7] which takes larger time steps to simulate many more reaction events. The reaction events corresponding to every reaction are determined by means of the Poisson random variable. The time step in the GASA is calculated using the leap condition which ensures that there is no significant loss of accuracy in the simulations, by assuming that the propensity function (the product of the rate constant of the reaction with the reactant populations) of each reaction in the system changes by an infinitesimal amount during each time step. This approach has helped to significantly reduce the computational time for the simulations. Several improvements to this method,

that serve to improve the accuracy, have been proposed, which include the Gillespie-Petzold (G-P) method,[29] and methods proposed by Rathinam *et al.*[84] and Cao *et al.*.[83] Along with these methods, the recently proposed representative reaction approach (RRA) of Kadam and Vanka[50] in which all the reactions in a chemical system are represented by a single, representative, reaction, $2A \longrightarrow B$, also presents a viable approach to reducing the computational time while maintaining the required accuracy for chemical systems.

However, one of the major problems with the accelerated methods described in the previous paragraph is that the reaction events corresponding to every reaction are determined by means of the Poisson random variable. The unbounded nature of sample values generated from Poisson random variables may result in negative molecular numbers if the time step is sufficiently large. One way to avoid this negative number problem is to use binomial random variables, since they have a finite range of sample values. Several methods which use the binomial distribution-based random variable have been developed, which include the BD-τ leap methods developed by Tian-Burrage[91] and Chatterjee *et al.*,[92] the multinomial tau leap method[93] which is an extension of the BD-τ leap, the Modified Binomial Leap Method,[94] generalized binomial leap method[127] for delayed reactions, R-leaping method[98] for accelerating the SSA by reaction firings and also the Binomial τ-leap Spatial SSA (Bτ-SSSA)[128] which considers both reaction and diffusion events during each τ leap. In addition to such methods using binomial random numbers, Cao et al.[95] have modified the GASA to solve this problem by classifying the reactions into critical and noncritical reactions, and using Poisson random variables. However, all the approximate methods developed to date that use binomial random numbers are not without their flaws. In case of the BD-τ leap method of Tian-Burrage,[91] the concept of the limiting reactants[91,92] is used to determine the maximum permitted occurrences of a particular reaction. This constraint is over restrictive in cases where there are certain reactions that increase the amount of the consumed reactant population for that particular reaction. The other problem of the Tian-Burrage method is that it fails for cases where there are multiple-channel reactant dependencies. To address these problems, Peng *et*

al.[94] have proposed the modified binomial leap method, but due to the necessity, in this approach, of generating some more binomial random variables for each time step, it is not computationally efficient. Chatterjee *et al.*[92] have also tried to solve the problem of negative populations by first determining the binomial random variable corresponding to a particular reaction and then updating the currently available molecules by subtracting the molecules that have reacted. The problem with this approach, however, is that, if the earlier considered reactions have used up most of the available molecules, then the reactions that are selected afterwards are left with fewer numbers of molecules and have fewer firings than the earlier reactions. The method, therefore, incorporates a bias in the choice of the reaction numbers depending on the order in which the reactions are selected. Chatterjee *et al.*[92] have admitted this problem and have suggested the random selection of the reaction order at each time step. This, however, does not entirely solve the bias at each time step for the reaction numbers.

The intent, in this chapter, is to propose two methods that can successfully address the aforementioned issues while also being computationally efficient and easy to code. The methods are based on the modification of the RRA[50] by (a) the addition of the condition that the SSA be used when negative populations result at a time step : the RRA-*N* with SSA approach and (b) the use of the binomial random variable for every time step : the RRA-*N* with binomial approach. It will be demonstrated that both the methods, which are easy to code, perform creditably in comparison to other methods. This article is organized as follows : (i) first, we have discussed in brief the RRA-based method: RRA-*N*, and the subsequent modifications, mentioned as (a) and (b) in the previous paragraph, which we are proposing to solve the problem of negative molecular numbers. The flowchart of the proposed methods has also been provided. After this, we have tested our proposed methods on (ii) the Carletti-Burrage model,[94] on (iii) the simple isomerization reaction Model, on (iv) the four reaction model and finally on (v) the challenging model of Circadian rhythms.[129,102] Finally, (vi) we have provided some conclusions to our work.

4.2 Overview and Modifications to the RRA

4.2.1 Basis for the stochastic simulation of chemical kinetics

Consider a well-stirred mixture of *N* molecular species $\{S_1, ..., S_N\}$ which are interacting with each other through *M* chemical reactions $\{R_1, ..., R_M\}$. It is assumed that this mixture is in thermal equilibrium at temperature T. The dynamical state of this entire chemical system (or mixture) at any time *t* can be specified by the state vector : $\mathbf{X}(t) \equiv \{X_1(t), ..., X_N(t)\}$. Every reaction $\{R_1, ..., R_M\}$ is characterized by a quantity called the propensity function $a_j(\mathbf{x})$,[65,25] which is the product of the rate constant and the number of reactant combinations for a given reaction. The other property of every reaction is the state change vector $\boldsymbol{\nu}_j \equiv \{\nu_{1j}, ..., \nu_{Nj}\}$. Here, $a_j(\mathbf{x})dt$ gives the probability that R_j reaction will occur in the next infinitesimal time interval [*t*, *t*+*dt*), and ν_{ij} is the change produced by the R_j reaction in the molecular population of the S_i species.

4.2.2 RRA-*N* : an overview

In this recently proposed method,[50] any chemical system consisting of individual chemical reactions is modeled by means of a single reaction, $2A \longrightarrow B$. In other words, the entire chemical system is represented by a single representative reaction (RR), hence the name : RRA. Discussion pertaining to the specific choice of the RR can be found elsewhere. During the course of the simulation in any particular iteration, the expected number of reactions that are supposed to take place in the next time step for the entire chemical system is first determined. For the RR, $2A \longrightarrow B$, the propensity function is taken as the sum of the propensity functions of all individual reactions and the rate constant as the weighted average of all the rate constants associated with the individual reactions. The firings associated with all the reactions in the chemical system are assumed to follow the Poisson distribution; that is, a Poisson random number is used to select the reaction numbers. Here, $a_j(\mathbf{x})\tau$ are the expected number of firings[7] of reaction

R_j in the time step τ.

Using the leap condition, the expected number of reactions (N_0) can be calculated for the concerned RR. This is derived to be :

$$N_0 = \frac{\varepsilon\prime a_0(\mathbf{x})}{C_0(2x_0 - 1)} \tag{4.1}$$

where, $\varepsilon\prime = 16\varepsilon$. The time step can be obtained from $\tau = \frac{N_0}{a_0(\mathbf{x})}$ and next, the expected number of firings of each reaction—the reaction number —can be obtained from the Poisson random number generator.

4.2.3 Modifications to the RRA-*N*

(i) The RRA-*N* with the SSA : As discussed in Introduction, approximate methods such as RRA-*N* that have been developed to speed up the simulations sometimes gives rise to physically unrealistic molecular numbers to negative amounts for the reactant species. When the time step is large and some of the reactants are present in small amounts, there exists the possibility of the molecular amounts being driven to negative values. Also, the simultaneous occurrence of different reactions can lead to negative numbers. When negative molecular numbers are encountered while simulating the chemical system by the RRA-*N* method, we propose to make use of the SSA, where only a single reaction is simulated in each infinitesimal time increment. The procedure is as follows : for any time step during the simulation, first the RRA-*N* method is followed, with the value of error control parameter, ε, taken to be 0.04. If, at the end of the given step, the molecular numbers of any of the species is determined to be negative, then that RRA-*N* step is cancelled. The algorithm does not take the large time step obtained from the RRA-*N* calculations and switches instead to the small time increment from the SSA. Although choosing a single reaction event during this (now) small time step, the SSA approach ensures that the chemical reaction selected for the single firing is usually the one that has relatively greater propensity than other reactions. As the reactions that have low propensity have an extremely small likelihood of being chosen, and since, even if chosen,

the reaction would fire only once, there is no possibility of obtaining negative numbers. The detailed procedure is explained in the flowchart.

(ii) The RRA-*N* with Binomial : In most of the approximate methods, the number of firings of the reactions has been assumed to follow the Poisson distribution. The range of sample values of Poisson random variables goes from zero to infinity. When a larger time step is taken during the simulations, the expected values of the reactions increases (as the expected value is a_j**(x)**τ) and, as a result, we get large values for the Poisson random variables. Thus, when the reaction number exceeds one of the molecular numbers in a particular reaction or when the same reactant is simultaneously involved in different reactions, negative numbers are obtained. As discussed in the Introduction, Tian-Burrage,[91] as well as Chatterjee *et al.*[92] have tried to address this issue by replacing the Poisson random variables with the binomial random variables; since they have a finite range of sample values.

What is proposed here, to address this problem of negative populations, is the application of the idea of binomial random variables to the RRA. To generate a binomial random variable from a computer algorithm,[72] we need to supply two variables to it, namely : (i) the maximum number of reactions (K_j) and (ii) the probability (*p*). The maximum number of reactions, K_j, are determined by using the concept of the limiting reactant. The binomial distribution has the same mean as that of the Poisson; which is a_j**(x)**τ. The probability of occurrence of R_j is given by : $p = \frac{a_j(\mathbf{x})\tau}{K_j}$. The maximum number of reactions K_j for some of the common chemical reactions are defined below :

For the first order reaction,

$$A \longrightarrow B, \quad K_j = A$$

For the second order reactions,

$$A + B \longrightarrow C, \quad K_j = min(A, B)$$

$$2A \longrightarrow B, \quad K_j = A/2$$

In case of RRA-*N* with $A \longrightarrow B$ as the RR, there are x_0 hypothetical species of the reactant *A* with the propensity function a_0**(x)** equal to the sum of the propensities of all the individual reactions; the maximum number of reactions are $K = \frac{x_0}{2}$ and the probability is $p = \frac{a_0(\mathbf{x})\tau}{K}$.

In this modified version of RRA-*N*, we propose to use the binomial approach whenever negative molecular numbers are encountered during the simulation. Thus, if a given RRA-*N* step results in some species having negative populations, we undo that particular step of the RRA-*N* by making a switch to the binomial method. Now, here the number of firings of the reaction, (N_0), is recalculated in exactly the same way as before, but this information is used to obtain the variables for obtaining the binomial random number. Consequently the time step τ, the maximum number of reactions (K_j) and the probability, *p* are all calculated. This recipe will give us the new number of reaction firings (N_0') for this given time step, the value of N_0' being obtained from the binomial distribution. Finally, the reaction numbers of the individual reactions are calculated from N_0' by determining their share of N_0' from the weighted average, that is from $n_j = \text{int}[(\frac{a_j(\mathbf{x})}{a_0(\mathbf{x})})N_0']$. The value of the error control parameter, ε, is the same as that used in the RRA-*N* with SSA (0.04).

The simulation of the chemical system by modifying the RRA-*N* using the binomial approach described here gives an edge over the existing binomial methods of Tian-Burrage[91] as well as that of Chatterjee *et al.*.[92] This will be clear in the subsequent sections where different examples have been discussed.

4.3 Steps and the Flowchart

4.3.1 Steps for the implementation of the modified-RRA

Based on the discussion above, the implementation steps for the RRA-*N* method modified using the SSA and the binomial approaches are outlined as follows :

- **Step 1 :** input the initial number of species, rate constants of the constituent reactions; initialize the counters and the random number generators to a seed value. Transfer the initial number of species to some temporary locations (variables)

- **Step 2 :** calculate the propensity functions : $\{a_1, ..., a_M\}$, the sum of the propensity functions : a_0**(x)** = $\sum_{j=1}^{M} a_j$**(x)**, the weighted rate constant : C_0 = $\sum_{j=1}^{M} \left(\frac{a_j\textbf{(x)}}{a_0\textbf{(x)}}\right) c_j$

- **Step 3 :** calculate the total number of species present.
  $x_0 = \frac{C_0+\sqrt{C_0^2+8a_0C_0}}{2C_0}$

- **Step 4 :** calculate the time step $\tau = \frac{N_0}{a_0\textbf{(x)}}$, where the total number of reactions are : $N_0 = \frac{16\varepsilon a_0\textbf{(x)}}{C_0(2x_0-1)}$, the value of ε being 0.04

- **Step 5 :** calculate the expected number of reactions, the n_{js}, for the individual reactions. $n_j = a_j\tau$

- **Step 6 :** use the Poisson random number generator to find the k_{js}, the actual number of occurrences for the individual reactions.
  k_j = poidev$(n_j, iseed)$

- **Step 7 :** make the necessary changes in the molecular population using the appropriate stoichiometric parameters and reaction numbers

- **Step 8 :** if negative numbers are encountered then discard the step; choose either (a) the SSA or (b) the binomial using the initial species stored in the temporary locations. If negative numbers are not encountered, then continue with the RRA-*N*
  (a) Use the SSA

  (b) Use the binomial

  (i) calculate the total number of reactions : $N_0 = \frac{16\varepsilon a_0\textbf{(x)}}{C_0(2x_0-1)}$
  (ii) calculate the time step : $\tau = \frac{N_0}{a_0\textbf{(x)}}$
  (iii) calculate the maximum number of reactions, *K* = $\frac{x_0}{2}$,

the probability, $p = \frac{a_0(\mathbf{x})\tau}{K}$ and finally, the binomial random variable N_0' = bnldev(p, K, *iseed*)

(iv) calculate the reaction numbers of the individual reactions, $n_j = \text{int}[(\frac{a_j(\mathbf{x})}{a_0(\mathbf{x})})N_0']$

(v) make the necessary changes in the species population using the appropriate stoichiometric parameters and reaction numbers

• **Step 9 :** go to Step 1

4.3.2 Flowchart :

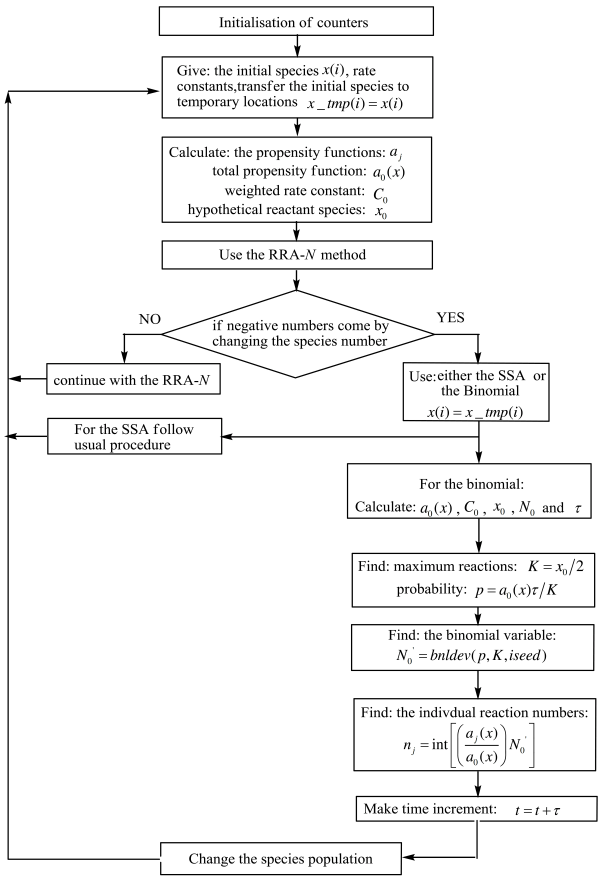


Figure 4.2: The algorithm for the RRA-*N* with SSA and the RRA-*N* with binomial methods.

4.4 The Carletti-Burrage Model

The Carletti-Burrage model[94] consists of the following set of reactions :

$$R_1 : \text{RNA} \xrightarrow{c_1} \text{DNA1}$$

$$R_2 : \text{DNA1} \xrightarrow{c_2} \text{RNA}$$

$$R_3 : \text{m} \xrightarrow{c_3} \text{RNA}$$

$$R_4 : \text{RNA} \xrightarrow{c_4} \text{m}$$

$$R_5 : 2\text{m} \xrightarrow{c_5} \text{D}$$

$$R_6 : \text{D} \xrightarrow{c_6} 2\text{m}$$

$$R_7 : \text{DNA} + \text{D} \xrightarrow{c_7} \text{DNA1}$$

$$R_8 : \text{DNA1} \xrightarrow{c_8} \text{DNA} + \text{D}$$

$$R_9 : \text{DNA1} + \text{D} \xrightarrow{c_9} \text{DNA2}$$

$$R_{10} : \text{DNA2} \xrightarrow{c_{10}} \text{DNA1} + \text{D}$$

where RNA, DNA, DNA1, DNA2, D, and m are the species taking part in the different reactions; and the symbols ($c_1 - c_{10}$) over the arrows indicate the rate constants of the respective reactions.

The values of the different model parameters used in the simulation of this reaction system are provided in the table.

Table 4.1: The values of the rate constants of the reactions and the initial species population for the Carletti-Burrage model.

Parameters used in the simulation	Numerical values of the parameters
c_1	0.078
c_2	3.9E-3
c_3	7.0E-4
c_4	0.043
c_5	0.083
c_6	0.5
c_7	0.020
c_8	0.479
c_9	2.0E-4
c_{10}	8.765E-12
m	200
D	600
DNA	200
DNA1	0
DNA2	0
RNA	0

The different stochastic approaches, discussed in Introduction, were used to simulate the Carletti-Burrage model. It is to be noted that the BD-τ method of Tian-Burrage[91] was inapplicable for this system, due to the fact that there is a bimolecular reaction (R_9), in which the two reactants (D and DNA1) are also the consumed reactants in two other reaction channels, thereby creating the problem of multiple reactions with common consumed reactants. Furthermore, we note here that the simulation of the Carletti-Burrage model using GASA,[7] led to physically unrealistic (negative) numbers. Hence, the results with GASA are also not discussed for this example. Therefore, only the methods that were successful at simulating this model: the RRA-*N* with SSA and the RRA-*N* with Binomial, discussed in the previous section, as well as the SSA,[28] the G-P[29] and BD-τ leap of Chatterjee *et al.*,[92] have been reported here.

The standard value[7] of ε = 0.03 was used for doing the simulations with GASA and G-P. It is also noted here that this is the standard value for ε that has been used for the

G-P method for all the subsequent examples as well. It is further noted that, having now established, with the example of the Carletti-Burrage model, that the GASA and the BD-τ method of Tian-Burrage do not have universal applicability in addressing the problem of negative populations in chemical systems, they will not be considered in the subsequent examples.

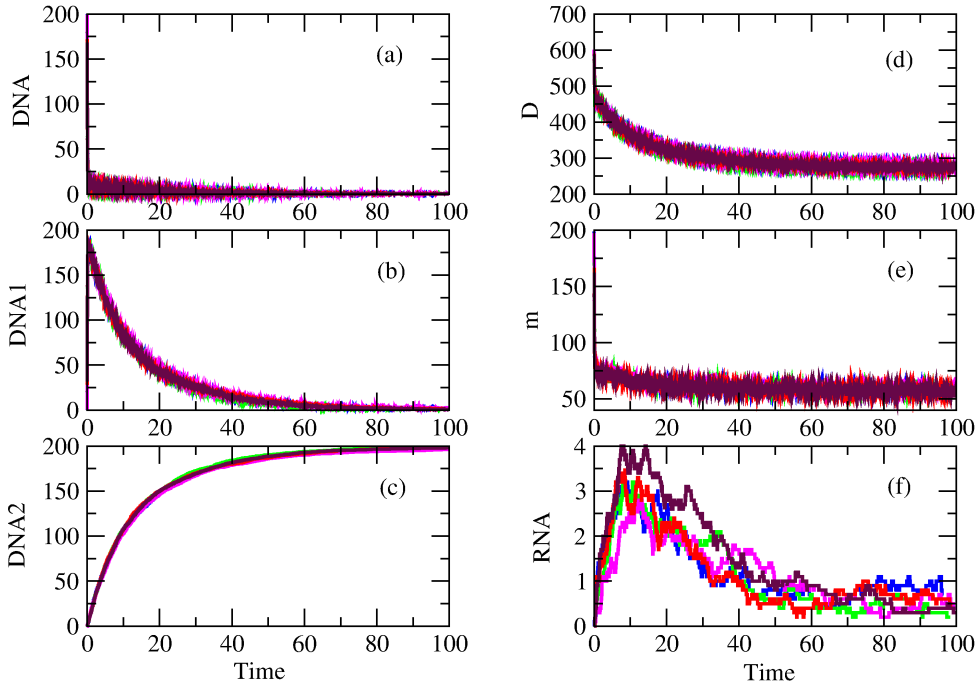


Figure 4.3: The trajectories of the means [(a)-(f)] for the probability distributions of the species DNA, DNA1, DNA2, D, m, RNA using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve), RRA-N with SSA (maroon curve) and RRA-N with binomial (red curve) for the case of the Carletti-Burrage model.

For each case, the simulations were performed 10 times, with the seed value of the random number generator being changed on each occasion, and the mean and the coefficient of variation (CV) calculated from the 10 simulations in each case. Shown in figure 4.3 is the comparison of the means of the probability distributions for the different species using SSA, G-P, the BD-τ method of Chatterjee-Vlachos-Katsoulakis, the RRA-N with SSA and the RRA-N with binomial. The CV for the probability distribution for the species using the respective simulation methods is shown in figure 4.4.

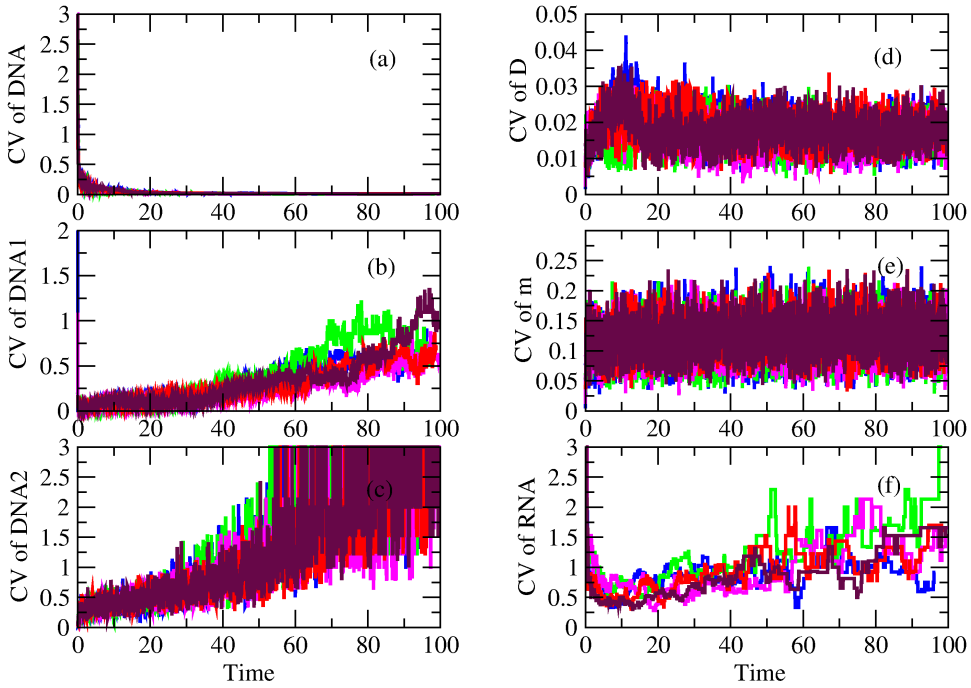


Figure 4.4: The trajectories of the CVs [(a)-(f)] for the probability distributions of the species DNA, DNA1, DNA2, D, m, RNA using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve), RRA-N with SSA (maroon curve) and RRA-N with binomial (red curve) for the case of the Carletti-Burrage model.

Table 4.2: The average values of the CPU time (secs) taken by different simulation methods for the Carletti-Burrage model.

Simulation Methods	SSA	G-P	BD-τ	RRA-SSA	RRA-binomial
CPU time (secs)	2.716	10.580	5.644	0.529	0.529

From the table, it is clear that, for the Carletti-Burrage model, the modified RRA-N methods : the RRA-N with SSA and RRA-N with binomial, provide an edge over the SSA as well as over the other accelerated methods. The Figures 4.3 and 4.4 indicate that there is a near exact correspondence between the simulation methods - the SSA, G-P, BD-τ of Chatterjee-Vlachos-Katsoulakis, RRA-N with SSA and RRA-N with binomial. However, the two RRA-N methods are also faster than all the other methods. In this context, it is to be noted that we have used the coarse grain variable factor f^{92} of 2.0 for the case of the BD-τ method. This causes the simulations to be accurate, but increases the computational time for this method considerably (Table 4.2). Increasing the value of f would speed up the simulations, but this can cause the accuracy of the simulations using the BD-τ method to drop off. This is illustrated in the Figures 4.5 and 4.6, where the means and CVs of the probability distributions for the different species are shown, with the BD-τ simulations having been done for f = 100.0. A perusal of Figures 4.5 and 4.6 shows that, for this case, the results obtained with the BD-τ method begin to diverge from the rest, even though, as shown in Table 4.2, the time required for the BD-τ simulations is reduced. In short, for this particular example, the modified RRA-N methods are seen to provide the proper balance of accuracy and speed as compared to the other approximate methods.

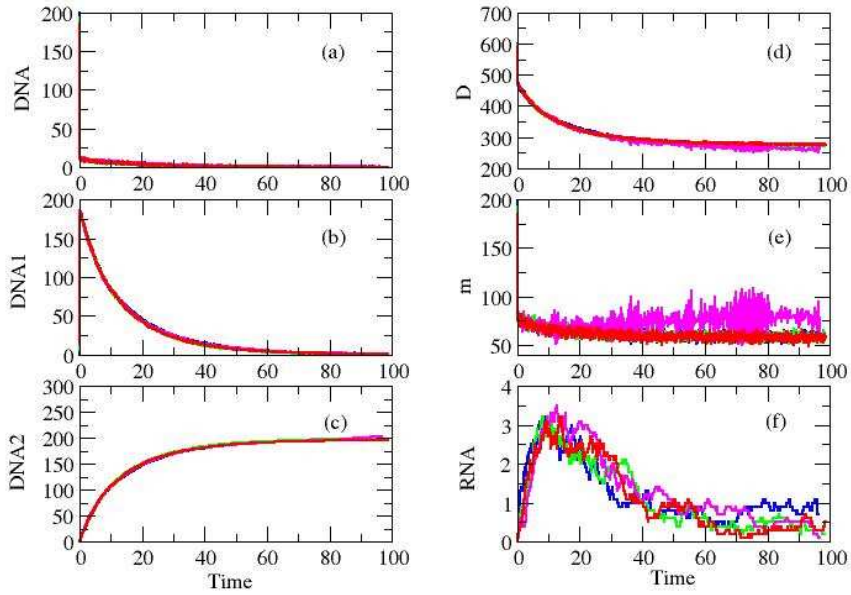


Figure 4.5: The trajectories of the means [(a)-(f)] for the probability distributions of the species DNA, DNA1, DNA2, D, m, RNA using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve), RRA-N with SSA (red curve) for the case of the Carletti-Burrage model.

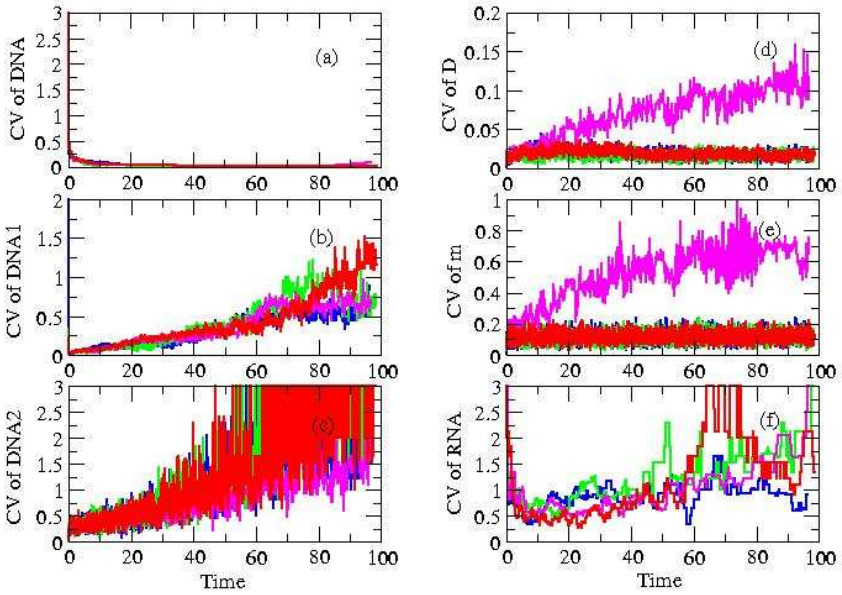


Figure 4.6: The trajectories of the CVs [(a)-(f)] for the probability distributions of the species DNA, DNA1, DNA2, D, m, RNA using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve), RRA-*N* with SSA (red curve) for the case of the Carletti-Burrage model.

Thus, for this example, the modified RRA-*N* methods offer a distinct advantage over the BD-τ methods of Tian-Burrage[91] and Chatterjee *et al.*,[92] using a Binomial distribution, as well as over the GASA[7] and G-P[29] methods, using a Poisson distribution.

4.5 The Simple Isomerization Reaction Model

In this section, we discuss the simple isomerization reaction model consisting of three reactions. This chemical system has the following set of reactions :

$$R_1 : X_1 \xrightarrow{c_1} X_2$$

$$R_2 : X_2 \xrightarrow{c_2} X_1$$

$$R_3 : X_2 \xrightarrow{c_3} X_3$$

where X_1, X_2, X_3 are the reacting species and c_1, c_2 and c_3 are the rate constants of the corresponding reactions.

Table 4.3: The values of the rate constants of the reactions and the initial species population for the simple isomerization reaction model.

Parameters used in the simulation	Numerical values of the parameters
c_1	1.0
c_2	2.0
c_3	0.01
X_1	120
X_2	60
X_3	0

A perusal of the rate constants (c_1 = 1.0, c_2 = 2.0 and c_3 = 0.01) suggests that the population of the species X_2 can become negative during the course of simulation by accelerated methods. The mean and the CV of the probability distributions for the species X_1, X_2, X_3 obtained by means of the SSA, G-P, the BD-τ method of Chatterjee-Vlachos-Katsoulakis and the RRA-*N* with SSA method is shown in Figure 4.7. Likewise, Figure 4.8 shows the corresponding curves for the RRA-*N* with Binomial method along with the other simulation methods.

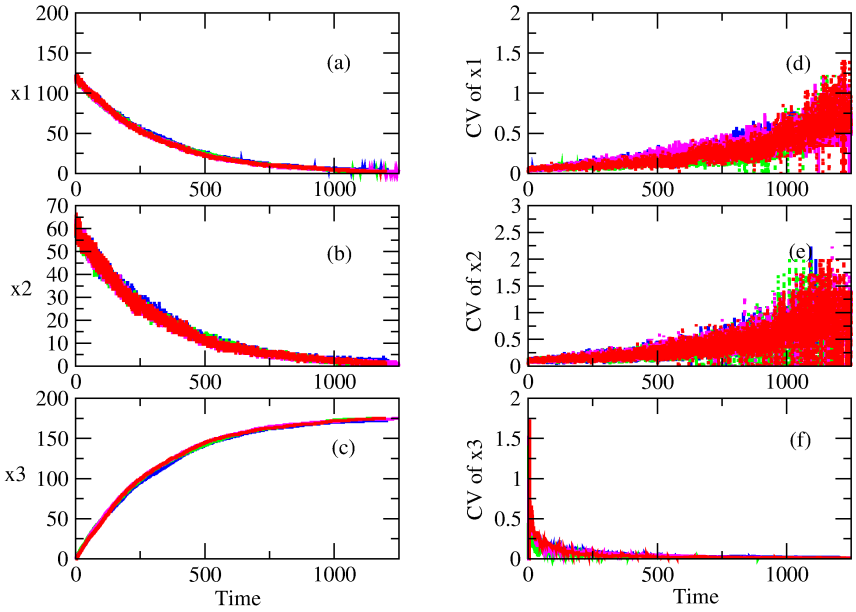


Figure 4.7: The trajectories of the means [(a)-(c)] and CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-N with SSA (red curve) for the case of the simple isomerization reaction model.

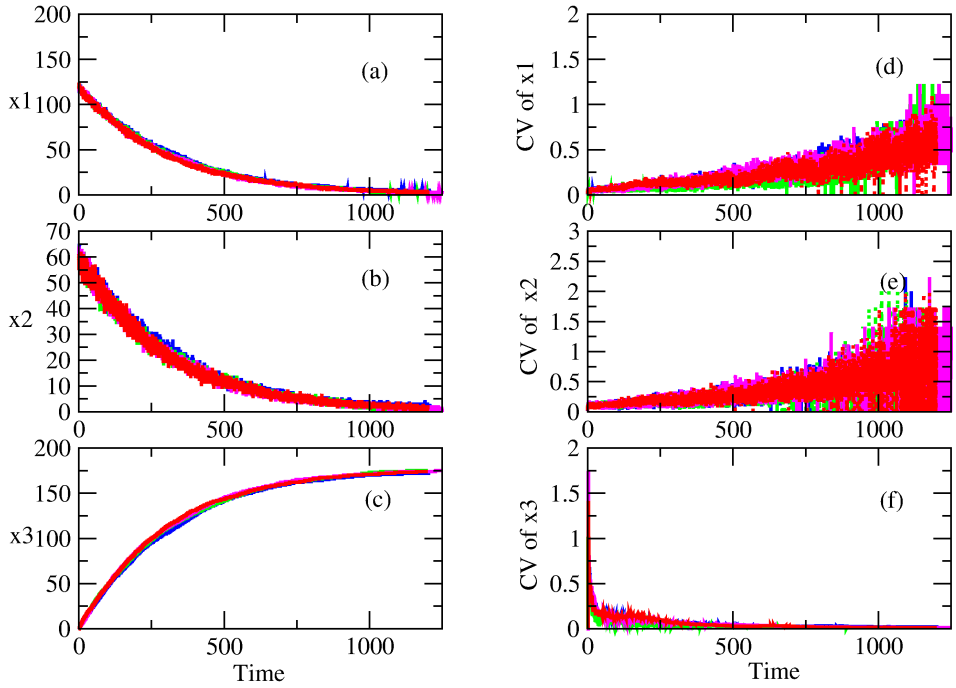


Figure 4.8: The trajectories of the means [(a)-(c)] and CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-N with binomial (red curve) for the case of the simple isomerization reaction model.

The Figures 4.7 and 4.8 indicate that the trajectories of the values for the mean and CV predicted by the SSA, the BD-τ method of Chatterjee-Vlachos-Katsoulakis, G-P, the RRA-N with SSA, and the RRA-N with Binomial are in good agreement. For the BD-τ method of Chatterjee-Vlachos-Katsoulakis, the value of coarse grain factor, f, is taken to be 2. Like in the previous example, increasing the value of f, while reducing the CPU time, leads to the loss of accuracy for this method. The comparison of the CPU time in Table 4.4 also shows that, in terms of efficiency, the RRA-N methods, especially RRA-N with SSA, perform better than the other accelerated methods.

Table 4.4: The average values of the CPU time (secs) taken by different simulation methods for the simple isomerization reaction model.

Simulation Methods	SSA	G-P	BD-τ	RRA-SSA	RRA-Binomial
CPU time (secs)	3.579	3.371	3.270	2.569	3.210

Hence, this example again emphasizes the satisfactory applicability of the new methods to chemical systems where negative population is a potential problem when applying approximate stochastic methods.

4.6 The Four Reaction Model

In this section, we discuss the four reaction model, as proposed by Gillespie.[7] The reactions of this model are given in section 3.3 of the chapter 3 of this thesis. The numerical values used for the simulation are provided in table below.

Table 4.5: The values of the rate constants of the reactions and the initial species population for the four reaction model.

Parameters used in the simulation	Numerical values of the parameters
c_1	1.0
c_2	0.5
c_3	0.002
c_4	0.04
X_1	1000
X_2	0
X_3	0

Figure 4.9 shows the mean and CV of the probability distributions for the species X_1, X_2 and X_3 obtained by means of the SSA, G-P, the BD-τ method of Chatterjee-Vlachos-Katsoulakis and the RRA-*N* with SSA method, while Figure 4.10 shows the same curves for the approximate stochastic methods compared with the RRA-*N* with binomial. Table 4.6 shows the average CPU time that was required to run the different simulations.

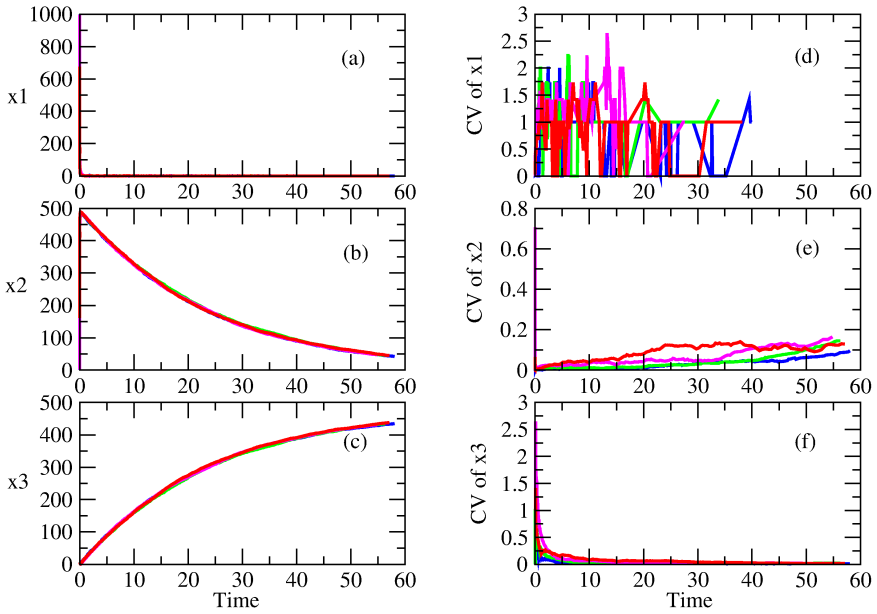


Figure 4.9: The trajectories of the means [(a)-(c)] and CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-*N* with SSA (red curve) for the case of the four reaction model.

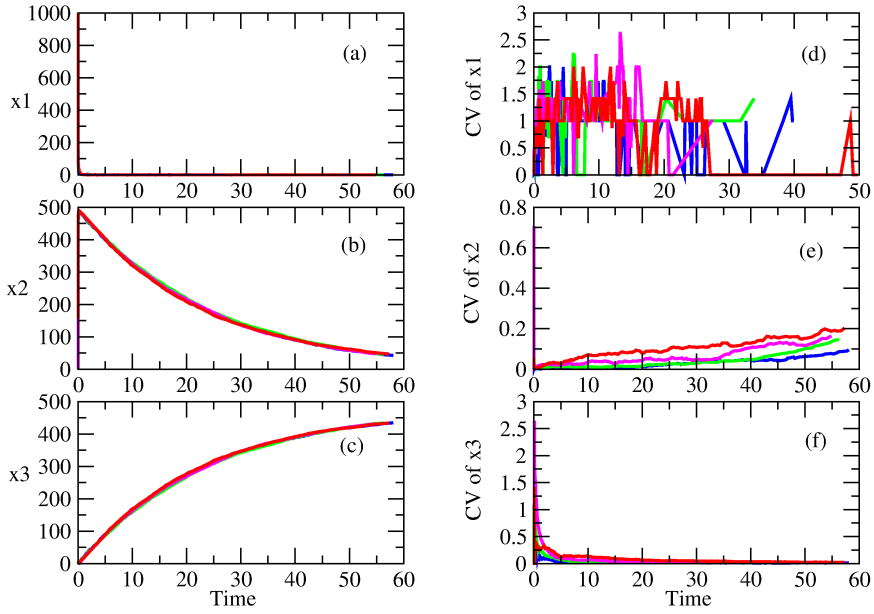


Figure 4.10: The trajectories of the means [(a)-(c)] and CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-N with binomial (red curve) for the case of the four reaction model.

Table 4.6: The average values of the CPU time (secs) taken by different simulation methods for the four reaction model.

Simulation Methods	SSA	G-P	BD-τ	RRA-SSA	RRA-Binomial
CPU time (secs)	4.499E-2	8.798E-2	5.539E-2	1.699E-2	1.999E-2

The Figures 4.9 and 4.10 indicate that there is, again, good agreement between the different methods. However, what is of note is that, as indicated by the average CPU times for the different methods shown in Table 4.6, both the G-P method as well as the BD-τ method of Chatterjee-Vlachos-Katsoulakis are actually slower on average than the SSA. This is due to the higher computational cost of running the more involved algorithms in the G-P and the BD-τ methods. The two RRA-N methods fare better in comparison, requiring about a third of the time in comparison to the other methods, while being as accurate. As before, it is noted here that the CPU time for the BD-τ method of Chatterjee-Vlachos-Katsoulakis can be significantly reduced by increasing the f value,

but this comes at a cost of accuracy of the simulations.

4.7 The Model of Circadian Rhythms

The final model discussed here is that of the Circadian rhythms[129, 102] which is observed in almost all living organisms and which is characterized by a period close to 24 h. These rhythms are cyclic hormonal processes, which have been observed in animals, plants, fungi, and cyanobacteria.

The following set of biochemical reactions have been used to describe the system of Circadian rhythms :

$$R_1 : D_a + A \xrightarrow{\gamma_a A} D'_a$$

$$R_2 : D'_a \xrightarrow{\theta_a} D_a + A$$

$$R_3 : D_r + A \xrightarrow{\gamma_r A} D'_r$$

$$R_4 : D'_r \xrightarrow{\theta_r} D_r + A$$

$$R_5 : \phi \xrightarrow{\alpha'_a D'_a + \alpha_a D_a} M_a$$

$$R_6 : M_a \xrightarrow{\delta_{ma} M_a} \phi$$

$$R_7 : \phi \xrightarrow{\alpha'_r D'_r + \alpha_r D_r} M_r$$

$$R_8 : M_r \xrightarrow{\delta_{mr} M_r} \phi$$

$$R_9 : \phi \xrightarrow{\beta_a M_a} A$$

$$R_{10} : A \xrightarrow{\delta_a A} \phi$$

$$R_{11} : \phi \xrightarrow{\beta_r M_r} R$$

$$R_{12} : R \xrightarrow{\delta_r R} \phi$$

$$R_{13} : A + R \xrightarrow{\gamma_c AR} C$$

$$R_{14} : C \xrightarrow{\delta_a C} R$$

Table 4.7: The values of the rate constants of the reactions and the initial species population for the model of Circadian rhythms.

Parameters used in the simulation	Numerical values of the parameters
α_A	50.0
$\alpha_A^{'}$	500.0
α_R	0.01
$\alpha_R^{'}$	50.0
β_A	50.0
β_R	5.0
δ_{MA}	10.0
δ_{MR}	0.5
δ_A	1.0
δ_R	0.05
γ_A	1.0
γ_R	1.0
γ_C	2.0
θ_A	50.0
θ_R	100.0

Because of the highly oscillatory nature of the rhythms, it is challenging to simulate such systems by approximate accelerated methods. Indeed, for this case, it was found that the G-P method, as well as the RRA-*N* method with Binomial both failed to avoid the problem of negative populations, and so have not been discussed in the figures and the tables pertaining to this system. Also, the BD-τ method of Chatterjee-Vlachos-Katsoulakis cannot be used to simulate this system because it is not able to take into account the constitutive reactions such as ($\phi \rightarrow A$), cases where the reactant (ϕ) amount is not a specified quantity. Therefore, the only approximate method left that both succeeds in avoiding the negative population problem, as well as avoids technical flaws pertaining to constitutive reactions, is the RRA-*N* with SSA method. The comparison of this approximate method with the SSA is shown in figure 4.11 below. Table 4.8 shows the average CPU time that was required for the SSA and RRA-*N* with SSA methods.

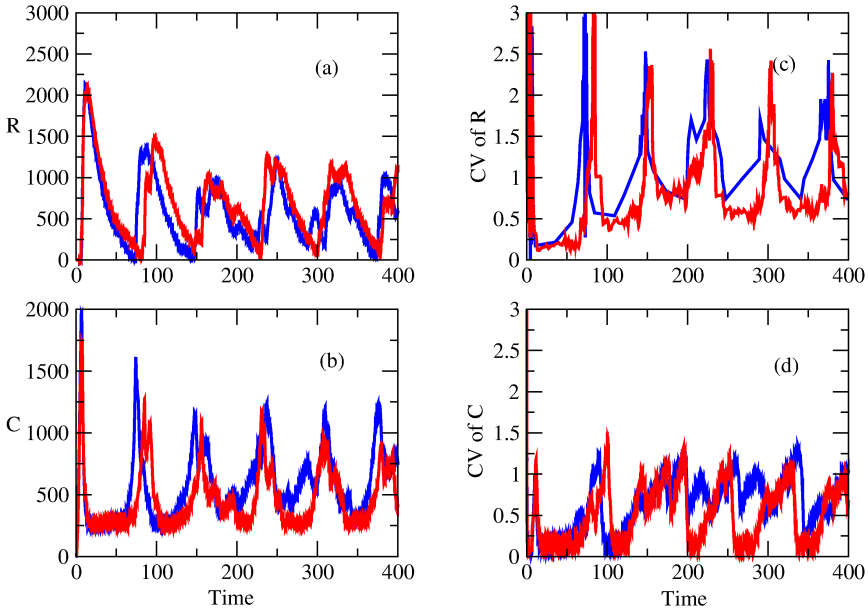


Figure 4.11: The trajectories of the means [(a),(b)] and CVs [(c),(d)] for the probability distributions of the species R and C using SSA (blue curve) and RRA-*N* with SSA (red curve) for the case of Circadian rhythms.

Table 4.8: The average values of the CPU time (secs) taken by different simulation methods for the circadian rhythms.

Simulation Methods	SSA	RRA-SSA
CPU time (secs)	30.103	24.075

From Table 4.8, it is clear that the RRA-*N* with SSA works faster than the SSA in replicating the results. The curves in figure 4.11 indicate a slight mismatch in the accuracy which is attributed to the highly unstable nature of the system over time. However, as figure 4.11 also indicates, the RRA-*N* with SSA method provides a reliable variation in the trend of the change in the means and CVs of the species populations with time.

Overall, for the four examples studied with the different simulation methods, the modified RRA-*N* methods : RRA-*N* with SSA and RRA-*N* with binomial are seen to perform at the same level or better than the other state-of-the-art approximate methods designed to solve the problem of negative populations. The other advantage of the modified RRA-*N* methods is that, unlike some Binomial methods like the BD-τ method of

Chatterjee *et al.*,[92] a standard parameter of 0.04 is specified and used consistently for all the simulations and gives reliable results. This offers an advantage over the BD-τ leap methods where it is actually required to optimize the coarse grain factor, f, for every different chemical system.

4.8 Conclusions

In this manuscript, an attempt has been made to solve the problem of physically unrealistic numbers, that is, negative populations for species present in a chemical system. This is a problem that can appear when approximate, accelerated stochastic methods are used. What has been discussed in the manuscript are modifications to the recently proposed RRA-based RRA-*N* method, to tackle the problem of negative populations. The two modified methods: RRA-*N* with SSA and RRA-*N* with binomial have been tested on various chemical systems and been found to perform at an advantage, in terms of efficiency as well as accuracy, over other state-of-the-art methods that have been proposed to tackle the problem of negative populations. The modified RRA-*N* methods have additional advantages in that they are quite simple in design and relatively straightforward to code. In addition, it is also not necessary to change the modified RRA-*N* methods, by modification of error control parameters, whenever a new system is encountered. As such, the new approaches described in this work present promising alternatives to existing approximate methods for tackling the problem of negative populations in chemical systems.

CHAPTER 5

The accounting of noise to solve the problem of negative populations in approximate accelerated stochastic simulations

Somewhere, something incredible is waiting to be known.

-Carl Sagan

Abstract

The advent of different approximate accelerated stochastic simulation methods has helped considerably in reducing the computational load of the exact simulation algorithms. However, along with the reduction in the computational load comes the risk of driving the molecular numbers to the regime of negative numbers during the simulations. Over the years, various methods have been developed in order to solve the problem by using different strategies. Some methods have employed binomial numbers to model the reactions, while others have tried the partitioning of the reaction network. In this manuscript, we have proposed a new approach where the noise inherent in the choice of the number of firings of a given reaction during a time step is taken into account. This idea of noise accounting is used in conjunction with the accelerated stochastic method: the Representative Reaction Approach (RRA). It is found that the new method is successful at solving the problem of negative numbers, and compares very favorably with other state-of-the-art stochastic simulation methods.

5.1 Introduction

The stochastic time evolution of a chemical system can be described by the Chemical Master Equation (CME).[49,65] But owing to its complexity, solving the CME is a difficult task and one has to rely on Monte Carlo simulation techniques that generate stochastic realizations of the underlying chemical kinetics. One such technique is the kinetic Monte Carlo (kMC)[54] based stochastic simulation algorithm (SSA)[28,13] developed by Daniel Gillespie. This technique simulates a randomly chosen single reaction during each time step giving stochastic realizations until a desired time is reached. However, this approach is demanding for the simulations of realistic systems. Subsequent to the development of the SSA, several methods have been developed in order to improve the performance of the SSA, such as the next reaction method,[73] the optimized direct method,[74] the sorting direct method[75] and the more recent recycling direct method (RDM).[82] In addition to these methods, the delay stochastic simulation algorithm (DSSA),[81] which considers time delays, has also been developed. It has been found that such attempts to increase the computational performance of the SSA have only been marginally successful.

The lack of significant success in improving the SSA with exact simulation approaches has led to the development of new approximate methods, where some of the accuracy of the SSA has been sacrificed. One such approach consists of hybrid methods,[102,130,101] which have been used for the multiscale simulations of chemical systems. In these methods, the chemical Langevin equation[25] or the reaction rate equations are coupled with the SSA. Even though the hybrid methods have succeeded in reducing the computational load of the SSA to some extent, they have lost the simplicity of the SSA. Another approach consists of leaping methods, where larger time steps are taken in order to simulate the occurrence of more reactions. In one such method, which was the first of its kind, Daniel Gillespie proposed Gillespie's approximate stochastic algorithm (GASA).[7] In this method, the time step during the simulations is derived from the leap

condition : a condition wherein a change in the number of reactant molecules in a given reaction is allowed as long as it alters the propensity function (the product of the reactant number of molecules and the rate constant) by an infinitesimal amount for that reaction. This method has helped to reduce the computational load of the simulations. Over the years, several improvements to this approach have been proposed, which includes the Gillespie- Petzold (G-P) method,[29] the implicit tau-leaping method of Rathinam et al.,[84] the efficient step size method of Cao *et al.*,[83] the *N*-leap method of Xu and Lan,[96] the *K*-leap method of Cai and Xu[97] and the recent Representative Reaction Approach (RRA)[50] that we have developed.

In all the approximate accelerated methods mentioned above, the reaction numbers are modeled by a Poisson distribution. Since the range of random variables generated by the Poisson distribution is unlimited, some reactions will fire many more times, thereby giving rise to physically unrealistic or negative numbers during the simulations. In other words, the occurrence of the negative population during the simulations can also be interpreted as the consequence of a violation of the leap condition. One obvious way to avoid the occurrence of negative populations is to model reaction numbers by random variables that have a finite range. Hence, simulation methods which use binomial random variables were developed. They include the BD-τ leap methods of Tian-Burrage,[91] Chatterjee *et al.*,[92] the multinomial τ-leap approach,[93] the efficient binomial leap,[94] the *R*-leap,[98] and the generalized binomial leap for delayed reactions,[127] as well as the RRA used in conjunction with the binomial distribution.[51] Apart from these methods, Cao *et al.*[95] have developed a method where the reaction network is partitioned into critical and noncritical reactions. This same concept of partitioning of a reaction network has been used by Yates *et al.*[131] in their confidence-based method. In case of such methods, the noncritical reactions were modeled by Poisson variables.

However, all the methods mentioned in the above paragraph have their own pros and cons. The BD-τ leap method of Tian- Burrage[91] is based on the concept of limiting reactants, which is used to determine the upper bound on the maximum allowed

firings of each reaction channel during a leap. This constraint seems to be artificial in some situations; for instance, in reaction networks in which there are certain reactions that tend to increase the consumed reactant numbers for that particular reaction. Along with this, this method also fails to simulate the cases where there are multiple channel reactant dependencies, i.e. cases where a single reactant gets consumed in multiple reactions. Chatterjee *et al.*[92] approached this problem by employing the binomial distribution while updating the currently available molecular population. This introduces some bias in the choice of the reaction numbers, as the earlier reactions occur more frequently than those selected afterwards. In other words, it depends on the order in which the reactions are selected. A potential solution to this problem that has been suggested is to choose the reactions randomly at each time step. The efficient binomial leap attempts to solve the problem, but it becomes slow due to the requirement of some more binomial random variables at each time step. Cao *et al.*[95] tried to fix the problem, but their solution became less flexible with the introduction of the second control parameter. Along with this, the use of Poisson random variables keeps alive the risk of physically unrealistic numbers being obtained. The multinomial τ-leaping[93] and R-leaping,[98] which are extensions of the binomial methods, have obtained some success at solving the aforementioned problems, but lose the computational simplicity of these methods.

Therefore, as the paragraphs above indicate, there are difficulties inherent in all the previous methodologies that have been employed to date. In this current work, we report a new approach that we have adopted in order to try and surmount these difficulties. This new method is primarily based on the notion of noise, and works in conjunction with the Representative Reaction Approach (RRA). The new approach is based on the reasoning that approximate stochastic simulation methods fail because the number of firings determined for each reaction during the simulation step by these methods is greater than the appropriate number. In other words, if y is the number of firings determined by an approximate accelerated method for a given reaction during a given time step, it is in excess of the appropriate value, say x, by a value n. By appropriate, what is meant is that x is the

value of the number of firings of that specific reaction that would be perfect in keeping with the leap condition. The implication here is that if x had been chosen as the number of firings by the approximate accelerated method, instead of y, then the simulation would have proceeded perfectly, without encountering problems such as that of negative populations. So, if one could find the correct number of firings (x) for each reaction in each time step, one could proceed with the accelerated simulation. Now, since $y = x + n$, if one could determine the amount (n) by which y exceeds the appropriate number of firings, x, then one could determine x and thus proceed with an accelerated algorithm that would give results faster than SSA but with problems such as negative populations eliminated. But how would one find n ?

We postulate that n is the noise inherent in the determination of the number of firings (y) for a given reaction by the approximate accelerated method. Now, noise in this context in the system can have both positive and negative values, because actual stochastic dynamics can both slow down or accelerate at any stage of the reaction. However, for the leap in question, where the number of firings has led to negative populations, the noise correction can only be a subtraction from the determined number of firings of the given reaction. This is because considering the fluctuations/noise as a positive correction to the number of firings would lead to even more unphysical negative values for the populations. Therefore, such corrections, while they can be calculated, are discarded.

Hence, if the noise n is subtracted from y, *viz.* $x = y - n$, then one would obtain the correct number of firings for each step, in accordance with the leap condition. This is further illustrated in Figure 5.1.

We have tested this idea for the case of the Representative Reaction Approach (RRA), an accelerated stochastic method that we have developed,[50] by incorporating this new concept of subtracting the noise, n, from the number of firings obtained for each step for every reaction. This new approach (termed as RRA-Noise), has been compared to a number of other accelerated methods that have been proposed in the literature, including GASA,[7] G-P,[29] the BD-τ of Chatterjee *et al.*[92] etc. The current approach provides results

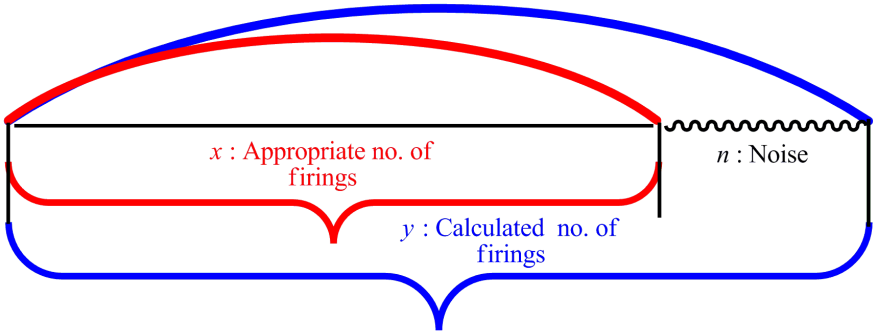


Figure 5.1: The pictorial theme of the concept of noise associated with every reaction; in a particular time step, *x* is the appropriate number of reactions, *y* is the calculated number of reactions and *n* is the associated noise.

which compares very favorably with other approaches, in addition to being relatively simple and easy to implement.

The rest of the paper is organized as follows : in the Methodology section, we have discussed in brief the necessary background required for the theoretical discussion, with the description of the new method followed by the implementation details of the same. In the Results and discussion section, simulations of different examples have been reported that confirm the reliability and efficiency of the newly proposed approach. The conclusions are provided in the last section.

5.2 Methodology

5.2.1 Background

A well-stirred mixture of *N* chemical species $\{S_1, ..., S_N\}$, which are interacting with each other through *M* chemical reactions $\{R_1, ..., R_M\}$ has been considered. The mixture is assumed to be in thermal equilibrium at some finite temperature T. The state of this mixture at any particular time, *t*, is specified by a state change vector : $\mathbf{X}(t) \equiv \{X_1(t), ..., X_N(t)\}$. Our aim is to study the time evolution of this *N* component vector from some given initial conditions, say, $\mathbf{X}(t) \equiv x_0$ Each chemical reaction R_j in the mixture is characterized by a propensity function, a_j and by the state change vector, $\boldsymbol{\nu}_j \equiv$

$\{\nu_{1j}, ..., \nu_{Nj}\}$. Here, ν_{ij} is the change produced by the R_j reaction in the molecular population of the R_j species. The quantity a_j(**x**)*dt* gives the probability that the R_j^{th} reaction will occur somewhere in the next infinitesimal time interval [*t*, *t*+*dt*).

5.2.2 Concept of Noise

As mentioned in the Introduction, the current approach is to determine the value of the number of firings (*n*) that is in excess of the appropriate value (*x*) that would be in accordance with the leap condition. This value *n* is determined as the noise present in the number of firings (*y*) calculated by the approximate accelerated stochastic method. In order to determine the value of the noise, *n*, we calculate the Poisson noise for every reaction firing value calculated for every step.

In an attempt to accelerate the SSA, Gillespie had modeled[7] the occurrences of different chemical reactions by Poisson random variables. Since the number of events (chemical reactions) taking place in a specific time interval are discrete in nature, it is apt to model them by the Poisson probability distribution. In other words, the firings of chemical reactions are treated as Poisson processes. It was further shown by Gillespie that the mean (or expected) value and the variance of the R_j^{th} reaction is $a_j\tau$. In a Poisson process, the actual number of reactions fluctuates about its mean value, $a_j\tau$, with a standard deviation of $\sqrt{a_j\tau}$. These fluctuations in the reaction numbers are treated as Poisson noise. In electronics, similar fluctuations are known as shot noise.[132, 133]

The fluctuation in every individual reaction implies that all reactions in the chemical system are accompanied by noise. The noise also expresses the basic form of uncertainty associated with the occurrence of the reactions. The uncertainty is substantial when the number of molecules participating in such reactions (or the propensity function) is small enough. It can be negligible (or very small), when the number of molecules (or the propensity function) are abundant. This means that the reactions in any chemical system are always accompanied by the noise. The strength of the noise associated with a reaction

varies as the square root of the expected number of firings of the given reaction. Thus, the noise relatively decreases as the expected number of firings of the reaction increases. However, the ratio of expected number of reactions to the noise,i.e., $\frac{a_j\tau}{\sqrt{a_j\tau}}$, increases.

In case of chemical systems that have less number of molecules, the simulations may show unfeasible fluctuations, which, in turn, may give rise to unrealistic (or negative) numbers. Hence, the occurrence of negative numbers can be avoided by removing such unfeasible fluctuations that are in the form of noise. Furthermore, it will be shown in the Results and discussion section that the removal of noise associated with every reaction does not affect the accuracy of the simulations.

5.2.3 Representative reaction approach (RRA) with noise

In order to speed up the SSA simulations, the Representative Reaction Approach (RRA)[50] has been proposed. In this recently proposed method, the chemical system to be simulated is represented by a single representative reaction (RR). The reaction that has been found to be the most effective is $2A \longrightarrow B$. Like any other reaction, the RR is also characterized by the rate constant and the propensity function. The propensity function, a_0**(x)** is the sum of propensities of all the individual reactions and the rate constant, C_0 is a weighted average of all the rate constants. Thereafter, the total number of hypothetical species, x_0 are calculated. Furthermore, by applying the leap condition to this RR, the expected reactions that are supposed to take place in the next time step are determined. The firings of the individual reactions are modeled by using the Poisson random number generator.[72] However, as discussed in the Introduction, such approximate accelerated methods as the RRA are prone to exhibiting negative numbers during the simulations, for certain reaction systems.

The current approach is to employ the notion of noise whenever negative numbers are encountered in a given step during the simulation. Initially, the simulations of any chemical system are carried out in the usual way by the RRA approach. When negative

numbers are obtained at any time step, that step of the RRA is annulled. Working on the assumption that the negative numbers obtained are an indication of excess noise in the leap (see Figure 5.1), the current method attempts to reduce the uncertainty in the fluctuations by removing the noise from the expected number of reactions. The procedure for the noise elimination based approach that is employed along with the RRA, is provided in the next subsection.

5.2.4 Steps for the implementation of RRA-Noise

The implementation details of the new method, RRA-Noise, are outlined below :

- **Step 1 :** input the initial number of species and the rate constants of the constituent reactions; initialize the counters and the random number generators to a seed value and transfer the initial number of species to some temporary locations (variables).

- **Step 2 :** calculate the propensity functions : $\{a_1, ..., a_M\}$ the sum of the propensity functions : a_0**(x)** = $\sum_{j=1}^{M} a_j$**(x)**,
  the weighted rate constant : C_0 = $\sum_{j=1}^{M} \left(\frac{a_j\textbf{(x)}}{a_0\textbf{(x)}}\right) c_j$

- **Step 3 :** calculate the total number of species present :
  $x_0 = \frac{C_0 + \sqrt{C_0^2 + 8a_0C_0}}{2C_0}$

- **Step 4 :** calculate the time step $\tau = \frac{N_0}{a_0\textbf{(x)}}$, where the total number of reactions are : $N_0 = \frac{16\varepsilon a_0\textbf{(x)}}{C_0(2x_0 - 1)}$, the value of ε being 0.06

- **Step 5 :** calculate the expected number of reactions for the individual reactions : $exp_j = a_j\tau$

- **Step 6 :** calculate the actual number of firings of individual reactions : $k_j = \text{poidev}(exp_j, iseed)$

- **Step 7 :** make the necessary changes in the molecular populations using the appropriate stoichiometric parameters and reaction numbers

- **Step 8 :** if negative numbers are not found, continue with the RRA; else discard the step and use the initial species stored for that step in the temporary locations.

- **Step 9 :** calculate the noise : $\sigma_j = \sqrt{a_j \tau}$

- **Step 10 :** calculate the corrected expected number of reactions :

  $exp_j\prime = exp_j - \sigma_j$

- **Step 11 :** calculate the new actual number of reactions :

  $n_j = \text{poidev}(exp_j\prime, iseed)$

- **Step 12 :** make the necessary changes in the molecular populations.

- **Step 13 :** go to Step 1.

5.3 Results and Discussion

Discussed below are the results of simulations done for four different chemical systems. In addition to the simulations done with the newly proposed RRA-Noise method, simulations have also been done with the Stochastic Simulation Algorithm (SSA)[28,13] and the approximate accelerated methods : the Gillespie's Approximate Stochastic Algorithm (GASA),[7] the Gillespie-Petzold (G-P) method[29] and the Binomial distribution based tau (BD-τ) method of Chatterjee *et al.*.[92] This section discusses the results of the simulations for the different systems and provides a comparison of the efficiency and robustness of the RRA-Noise method in comparison to the other methods. Specifically, what was compared was (i) the means and the coefficient of variations (CVs) obtained for 500 simulation runs for each method, and (ii) the average CPU times and the number of steps for a simulation obtained from the CPU times of the 500 simulation runs for each method. This was done for all the five chemical system examples considered. The stability analysis of the newly proposed algorithm has been discussed for the system of first order reactions and the oscillatory reaction model.

5.3.1 The Carletti-Burrage Model

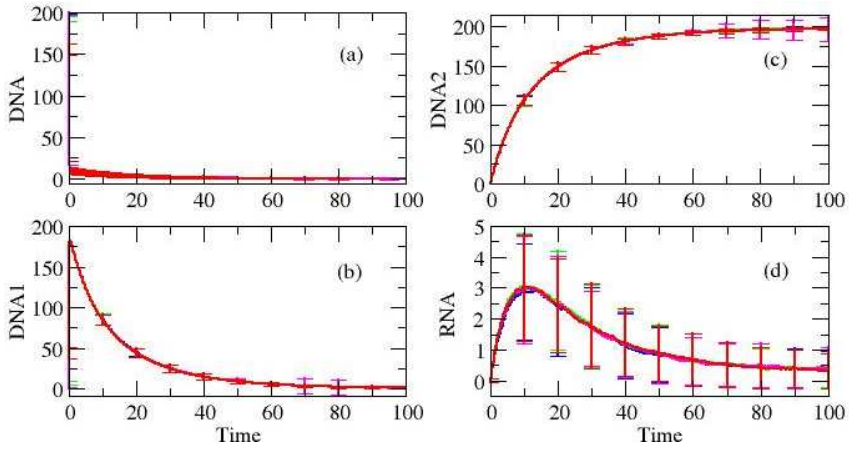
This reaction network was proposed by Carletti and Burrage.[94] It has already been discussed in section 4.4 in chapter 4 of this thesis.

The Carletti-Burrage model is simulated by different methods which have been discussed earlier. In case of simulation by Gillespie's Approximate Stochastic Algorithm (GASA), we found that negative molecular numbers occurred for some species during certain steps of the simulations. Thus, GASA was found to be unsuitable for the simulation of this model. Moreover, the binomial distribution based tau (BD-τ) method of Tian-Burrage[91] could not be applied for this model, since there are some species which take part in multiple reactions: a situation that the BD-τ method of Tian-Burrage is incapable of handling, making it technically non-applicable for such reaction networks. Hence, only the methods that were successfully able to reproduce the simulation trajectories are reported here. They are: the Stochastic Simulation Algorithm (SSA),[28,13] the Gillespie-Petzold (G-P) method,[29] the binomial distribution based tau (BD-τ) method of Chatterjee *et al.*,[92] and our newly proposed method: RRA-Noise.

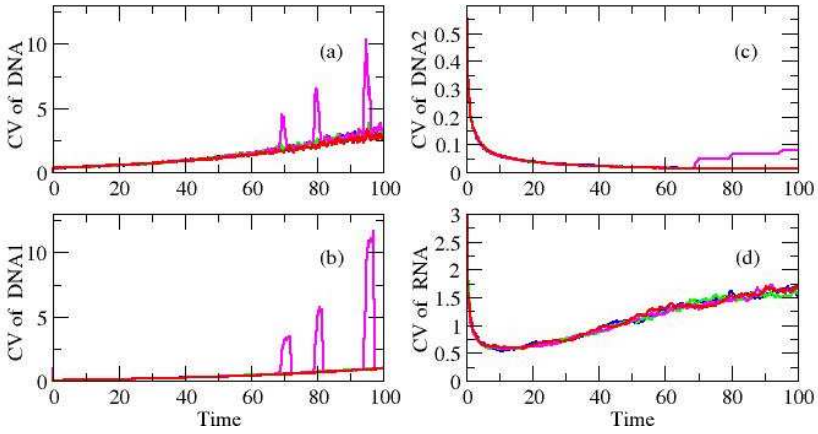
The error control parameter, ε, with a standard value of 0.03 was used for doing the simulations with GASA and G-P (this value of ε has been the standard value employed in previous reports),[7,29] while, in the case of the RRA-Noise, the value of ε has been taken as 0.06. These ε values have been used for all the subsequent examples of the chemical systems simulated by these methods. As mentioned earlier, the values of the means (with their respective error bars) and the CVs reported in Figures 5.2a and 5.2b have been calculated over an ensemble of 500 simulation runs, using a different seed value for the random number for each run.

The comparisons of the means (with $\pm$ 1SD error bars) of the probability distributions of some key species of the Carletti- Burrage model using SSA, G-P, the BD-τ method of Chatterjee- Vlachos-Katsoulakis, and the RRA-Noise is shown in Figure 5.2a. The coefficient of variation (CV) for the same species for the same set of simulation methods

is shown in Figure 5.2b below.



(a) means



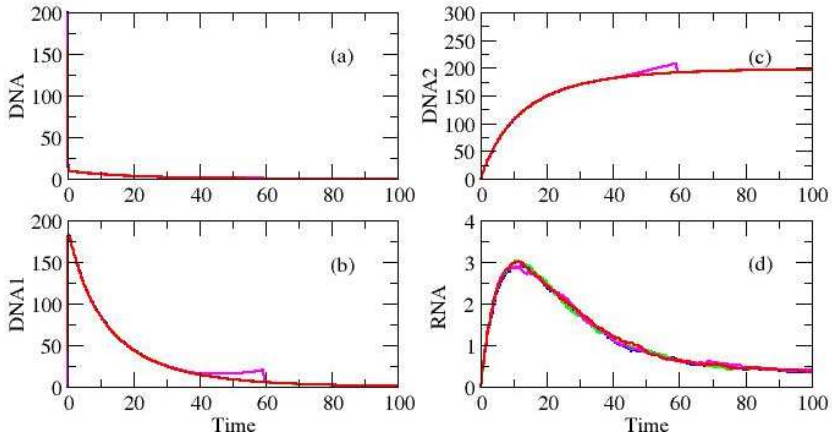
(b) CVs

Figure 5.2: The trajectories of the means and CVs [(a)-(d)] for the probability distributions of the species DNA, DNA1, DNA2 and RNA using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-Noise (red curve) for the case of the Carletti-Burrage model.

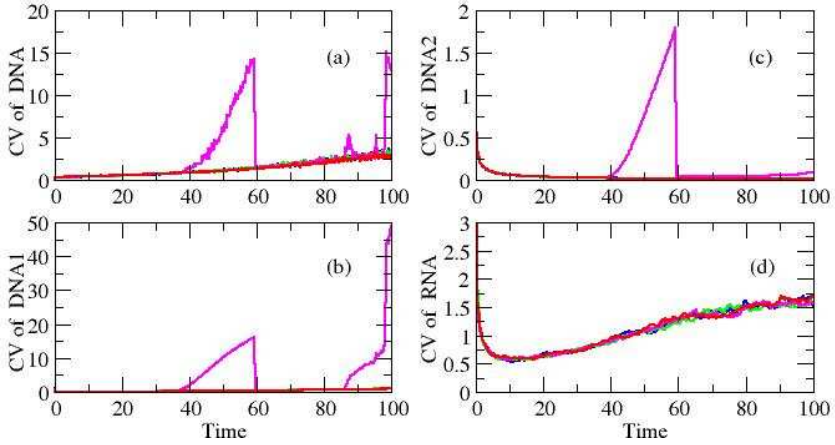
Table 5.1: The average values of the CPU time (secs) and the number of steps for 500 simulations taken by different simulation methods for the case of Carletti-Burrage model.

Simulation Methods	SSA	G-P	BD-τ	RRA-Noise
CPU time (secs)	5.029	14.970	40.327	4.234
Steps	33210	30905	16368	2854

The CPU time values in Table 5.1 show that the newly proposed RRA-Noise is significantly faster than the G-P and the BD-τ methods, and faster than the SSA. This is evident by the less number of steps taken by the RRA-Noise method. The overlap of the trajectories of the means of the respective species in Figure 5.2a is an indicator of good agreement between the different simulation methods. In the case of the BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta), it is found that the tail ends of the simulated trajectories (for DNA, DNA1 and DNA2) are not within the $\pm$ 1SD error bars of the SSA trajectories. The spikes in profiles of CVs for the same species in Figure 5.2b are a signature of this deviation, which are not in agreement with the others. The occurrence of the spikes is attributed to the increase in the standard deviation at the respective time points. In case of the BD-τ method, the time steps are taken by employing a coarse grain factor,[92] f, taken as 2.0. It has been found that the smaller value of the coarse grain factor serves to make the simulations more accurate. However, this also leads to an increase in the CPU time. The increase in the value of f reduces the CPU time, but this now leads to the loss of accuracy in the simulations. This is shown in Figure 5.3a and 5.3b, where the simulations are performed by increasing the coarse grain factor, f, to 4.0. Figure 5.3a and 5.3b depicts different (inaccurate) simulation trajectories obtained from the BD-τ of Chatterjee-Vlachos-Katsoulakis at the reduced CPU time.



(a) means



(b) CVs

Figure 5.3: The trajectories of the means and CVs [(a)-(d)] for the probability distributions of the species DNA, DNA1, DNA2 and RNA using SSA (blue curve), G-P (green curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-Noise (red curve) for the case of the Carletti-Burrage model with an increase in coarse grain factor.

On the other hand, in the case of the G-P method, the choice of SSA during the simulations contributes to the increase in the CPU time. Thus, with the RRA-Noise results lying within the SSA results in terms of $\pm$ 1SD error bar, it turns out that it provides good results in terms of accuracy. This is a heartening result, especially since the RRA-Noise is seen to perform considerably better than the BD-τ method, which had been specifically developed to tackle the problem of negative populations in chemical systems.[92]

5.3.2 The Simple Isomerization Reaction Model

This subsection discusses the simple isomerization reaction model. It has been discussed in section 4.5 in chapter 4 of this thesis. The reaction model alongwith a table of parameters used for the simulation have been provided in chapter 4.

Like for the previous example, the simple isomerization reaction model has also been simulated by different simulation methods. The comparisons of the means with $\pm$ 1SD error bars and the CVs of the probability distributions for the species X_1, X_2, X_3 by using the SSA, the G-P, the BD-τ method of Chatterjee-Vlachos-Katsoulakis, and the RRA-Noise is shown in Figure 5.4. Their trajectories have been calculated over an ensemble of 500 simulation runs. It was found that the simulation of this model by GASA leads to negative numbers, making it inapplicable for comparisons with other methods. The model system has been simulated by the BD-τ method with a coarse grain factor, *f*, equal to 2.0. It was seen that while the simulated profiles are accurate for this coarse grain factor value, the simulation also takes excess CPU time of 59.479 seconds. An attempt to reduce this CPU time by increasing the value of *f* to 50.0 gives totally different trajectories, as shown in Figure 5.5. Furthermore, this also comes at the risk of obtaining negative numbers for the species X_2. Thus, for this example, the RRA-Noise again scores over BD-τ : a method that had been developed specifically in order to sort out the issue of negative numbers.

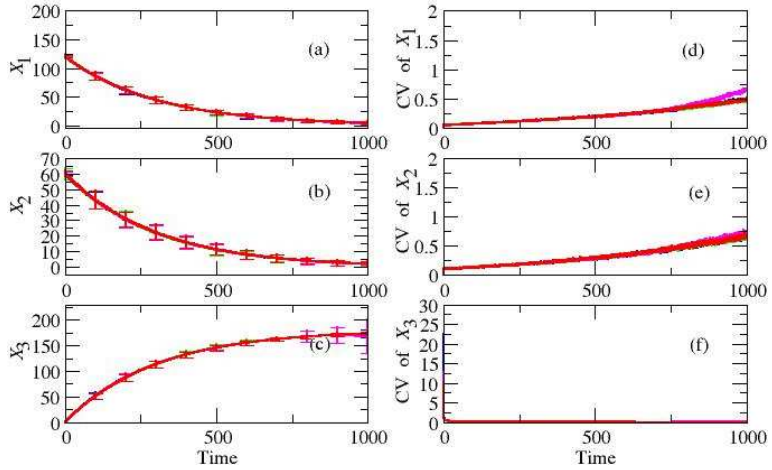


Figure 5.4: The trajectories of the means with $\pm$ 1SD error bars [(a)-(c)] and CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-Noise (red curve) for the case of the simple isomerization model.

Table 5.2: The average values of the CPU time (secs) and the number of steps taken by different simulation methods for the case of the simple isomerization reaction model.

Simulation Methods	SSA	G-P	BD-τ	RRA-Noise
CPU time (secs)	83.919	33.131	59.479	35.151
Steps	68468	36126	34895	27826

Admittedly, the G-P method is marginally better in terms of accuracy in comparison to RRA-Noise, and the two methods are found to be equally accurate, which indicates that the G-P is the most effective method for the simulation of this particular chemical system. However, the RRA-Noise is only slightly less efficient, which indicates that it would be almost as effective as the G-P in simulating this system. It takes less number of steps than observed for all of the other methods (Table 5.2). It is also found that the trajectories obtained from the RRA-Noise simulations are within the SSA results from the viewpoint of $\pm$ 1SD error bars. Therefore, this example also showcases the efficiency and reliability of RRA-Noise at simulating a chemical system that is susceptible to the problem of negative numbers.

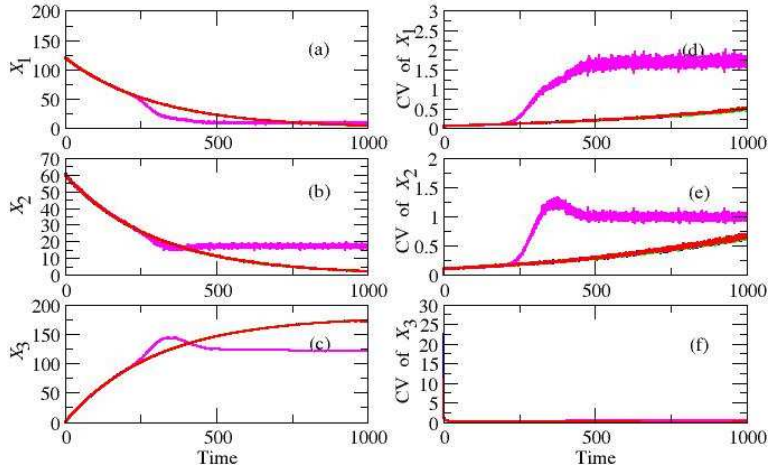


Figure 5.5: The trajectories of the means [(a)-(c)] and CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-Noise (red curve) with f = 50 for the case of the simple isomerization model.

5.3.3 Simple Model System

The simple model system of two reactions discussed here was used by *Cao et al.*[95] to test the reliability and efficiency of their modified Poisson tau leap method. It consists of the following set of reactions:

$$R_1 : X_1 \xrightarrow{c_1} X_2$$

$$R_2 : X_2 \xrightarrow{c_2} X_3$$

where X_1, X_2, X_3 are the reacting species and c_1, c_2 are the rate constants of the corresponding reactions.

Table 5.3: The values of the rate constants of the reactions and the initial species population for a simple model of two reactions.

Parameters used in the simulation	Numerical values of the parameters
c_1	10.0
c_2	0.1
X_1	9
X_2	20000
X_3	0

The study of the associated rate constants of the two reactions (c_1 = 10, c_2 = 0.1) and the corresponding reactant species (X_1 = 9, X_2 = 20 000) indicates that there is a possibility of getting negative numbers for the X_1 species. This fact gets conrmed when the G-P method is seen to drive the X_1 species to unrealistic numbers during the simulations. The same is seen to be true for GASA. Hence, apart from the SSA, only BD-τ and RRA-Noise have been considered. The results are shown in Figure 5.6 and Table 5.4 below.

In case of species X_1, it was observed that it falls off rapidly and afterwards does not demonstrate any fluctuation. Hence, the time trajectories of the X_1 species are not reported in Figure 5.6. What is shown are the time trajectories of the X_2 and the X_3 species. For these two species, the values shown in Figure 5.6 indicate that there is considerable agreement between all the simulation methods.

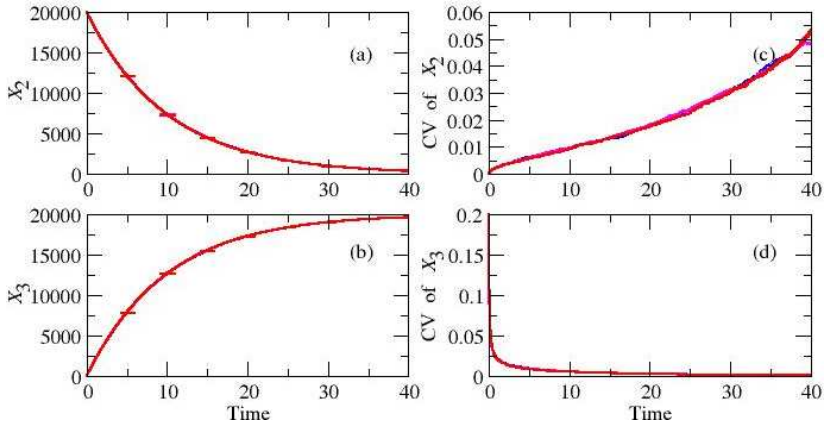


Figure 5.6: The trajectories of the means with $\pm$ 1SD error bars [(a) and (b)] and the CVs [(c) and (d)] for the probability distributions of the species X_2 and X_3 using SSA (blue curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-Noise (red curve) for the case of the simple model system.

Table 5.4: The average values of the CPU time (secs) and the number of steps taken by different simulation methods for the case of the simple model system.

Simulation Methods	SSA	BD-τ	RRA-Noise
CPU time (secs)	8.298	16.794	7.257
Steps	19633	9814	726

The mean trajectories of X_2 and X_3 are within the $\pm$ 1SD error bars of the SSA trajectories. The significantly less number of steps contribute to the CPU time performance of the RRA-Noise. The simulated trajectories and the CPU times tabulated in Table 5.4 indicate the effectiveness of the RRA-Noise. More importantly, it is again seen to be faster than the BD-τ method. Like earlier examples, any attempt to increase the efficiency of BD-τ leads to loss of accuracy in the results.

5.3.4 Model of First Order Reactions : Simulation and Numerical Stability

In this section, the simulation along with its numerical stability of the model of four unimolecular reactions is discussed. This model was used by Chatterjee *et al.*[92] to test their BD-τ method.

$$R_1 : X_1 \xrightarrow{c_1} X_2$$

$$R_2 : X_2 \xrightarrow{c_2} X_3$$

$$R_3 : X_3 \xrightarrow{c_3} X_2$$

$$R_4 : X_2 \xrightarrow{c_4} X_1$$

here, X_1, X_2, and X_3 are the species taking part in four different reactions and c_1, c_2, c_3 and c_4 are the rate constants of these reactions.

Table 5.5: The values of the rate constants of the reactions and the initial species population for first order reactions.

Parameters used in the simulation	Numerical values of the parameters
c_1	2.0
c_2	1.0
c_3	2.0
c_4	1.0
X_1	20000
X_2	0
X_3	0

In this model, which consists of all first order reactions, the species X_2 takes part in several of the reactions. This makes the reaction network more complicated in comparison to the earlier example pertaining to the simple isomerization reaction model.

As shown in Figure 5.7, there is good agreement between the means and CVs of the probability distributions for the species X_1, X_2 and X_3 obtained by using the SSA, the BD-τ method of Chatterjee- Vlachos-Katsoulakis and RRA-Noise. Unlike the previous

two examples, the G-P method gives rise to negative numbers during the simulations as does GASA. Hence, they are not included for the comparative study along with the others. As in all the previous cases, the time profiles of all the species have been calculated over an ensemble of 500 different simulation runs.

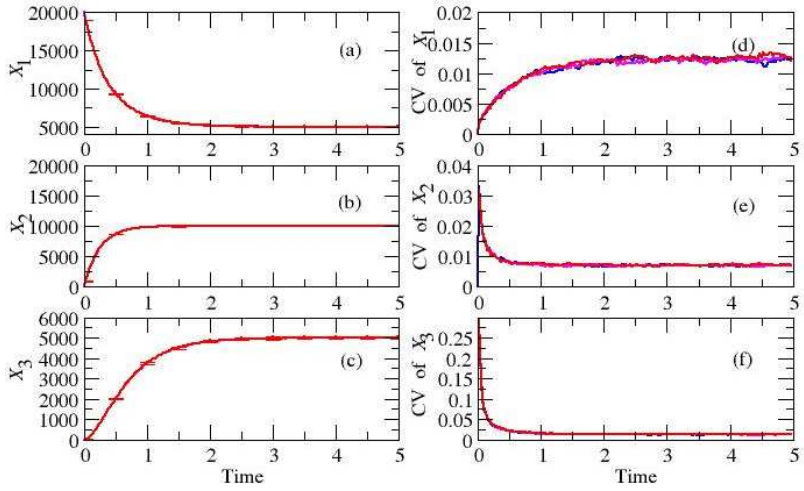


Figure 5.7: The trajectories of the means with $\pm$ 1SD error bars [(a)-(c)] and the CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) and RRA-Noise (red curve) for the case of the model of first order reactions.

Table 5.6: The average values of the CPU time (secs) and the number of steps taken by different simulation methods for the case of the model of first order reactions.

Simulation Methods	SSA	BD-τ	RRA-Noise
CPU time (secs)	18.141	56.224	9.557
Steps	199180	100001	3640

The average CPU times shown in Table 5.6 indicates that RRA-Noise is computationally more efficient than the other methods. It is almost twice as fast as the SSA. It also indicates that the number of steps taken by RRA-Noise are very much less than other methods. The low computational efficiency of BD-τ can be attributed to the small size of the time steps, leading to a large number of steps. As before, this can be changed by increasing the value of the coarse grain factor (2.0), but that, as in the previous examples,

leads to a significant loss of accuracy. This is illustrated in Figure 5.8, showing the results of simulations where the coarse grain factor had been increased to 5000.0. As Figure 5.8 indicates, the mean and the CVs for the different species becomes far less accurate for the BD-τ case in comparison to the other methods. The CPU values corresponding to the aforementioned f values are given in Table 5.7.

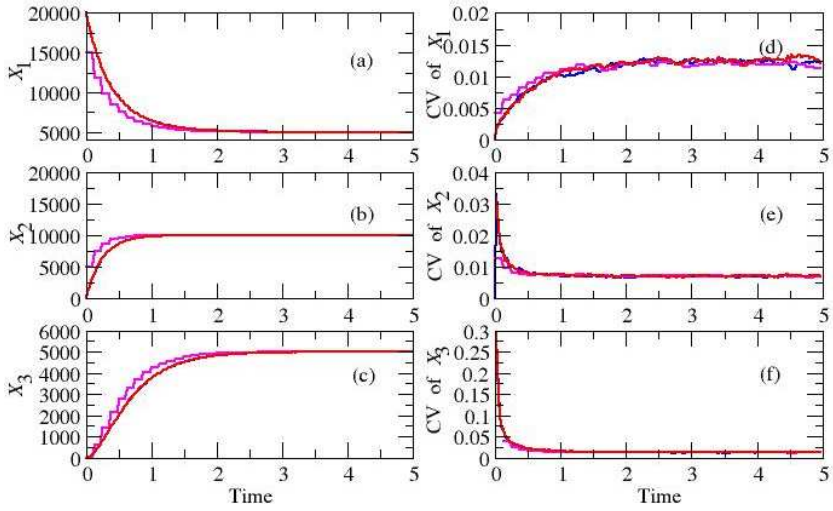


Figure 5.8: The trajectories of the means [(a)-(c)] and the CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (magenta curve) with f = 5000 and RRA-Noise (red curve) for the case of the model of first order reactions.

Table 5.7: The average values of the CPU time (secs) taken by different simulation methods for the case of the model of first order reactions with f = 5000.

Simulation Methods	SSA	BD-τ	RRA-Noise
CPU time (secs)	18.141	1.093	9.557

The results in Figure 5.9 provide a good match of simulation methods with each other in terms of $\pm$ 1SD error bars. Overall this model of first order reactions is a classic example where, in addition to the accuracy, the CPU times of the corresponding methods are of prime importance. And here, as in the previous cases, the newly proposed RRA-Noise again ends as the most favorable approximate simulation method.

The numerical stability of the RRA-Noise is discussed for this example by taking multiple runs and further benchmarking them against the SSA. The SSA is simulated over an ensemble of 20 000 (black curve) simulation runs. The RRA-Noise is simulated over an ensemble of 100 (red curve), 500 (green curve), 1000 (blue curve), 5000 (brown curve) and 10 000 (orange curve) simulation runs. It has been found that with the increase in the number of realizations, the RRA-Noise gets converged to the SSA trajectories with the decrease in the error. The error between the trajectories of the SSA and the RRA-Noise has been calculated at some chosen discrete time points. The Figure 5.9 shows the trajectories ((a)-(c)) of the species along with their closely monitored behavior ((d)-(f)) on a different scale. The Table 5.8 provides the absolute errors for the different realizations at specific time points.

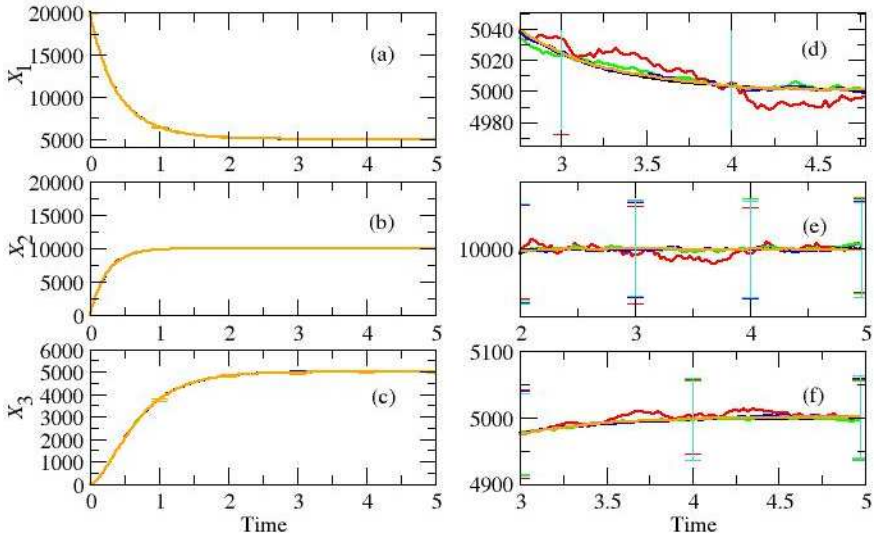


Figure 5.9: The trajectories of the means with $\pm$ 1SD error bars [(a)-(c)] and the same trajectories on a different scale [(d)-(f)] simulated using RRA-Noise with 100 (red curve), 500 (green curve), 1000 (blue curve), 5000 (brown curve), 10000 (orange curve) runs and SSA with 20000 runs (black curve) for the case of the model of first order reactions.

Table 5.8: The absolute errors between the trajectories of the SSA and the RRA-Noise for different runs at discrete time points.

Time points with 100 runs			
1	17.779	2.922	14.856
2	0.294	0.932	0.771
3	9.910	8.100	0.189
4	1.672	5.550	5.878
Time points with 500 runs			
1	9.663	5.020	4.642
2	0.577	2.640	4.063
3	1.826	1.895	1.930
4	0.347	2.411	0.063
Time points with 1000 runs			
1	8.577	4.830	3.746
2	0.649	1.958	4.608
3	0.672	0.032	1.359
4	0.035	0.580	1.455
Time points with 5000 runs			
1	8.379	4.153	4.225
2	1.443	0.923	2.520
3	0.326	2.344	0.672
4	0.212	0.891	1.321
Time points 10000 runs			
1	7.918	3.826	4.092
2	0.728	0.861	1.867
3	0.437	1.230	0.332
4	0.145	0.885	1.260

It is found that with the increase in the number of runs, the absolute error tends to decrease, thereby converging towards the SSA. The trajectory of RRA-Noise with 10 000 runs (orange) gets closer to the SSA profile, relative to the curve which has 100 runs (red). This behavior can be observed in the Figure 5.9 ((d)-(f)), where wide fluctuations are observed for the curve with 100 runs in comparison to those with 10 000 runs.

5.3.5 Oscillatory Model System : Simulation and Numerical Stability

The simulation as well as the numerical stability of the oscillatory reaction model, namely the Oregonator model, will be discussed in this section. This model was simulated by Daniel Gillespie by using the SSA. The simulation of this chemical system by GASA leads to negative numbers, hence the results with the GASA have not been discussed further. The BD-τ of Chatterjee-Vlachos-Katsoulakis has been found to be inapplicable for this particular system. The rest of the methods : SSA, G-P and RRA-Noise have been discussed below. The oscillatory nature of this model poses a challenge to methods that claim to solve the problem of negative numbers.

The mean trajectories are shown in Figure 5.10 (a)-(c), while Figure 5.10 (d)-(f) show the corresponding trajectories of the CVs. The behavior of the trajectories in the Figure 5.10 (a)-(c) by the different simulation methods underlines the oscillatory nature of the chemical system. The trajectories of the G-P (green curve) show a slightly out-of-phase behavior relative to the others. However, the trajectories of SSA (blue curve) and RRA-Noise (red curve) are found to be in good agreement. This is observed to a good extent in all the curves.

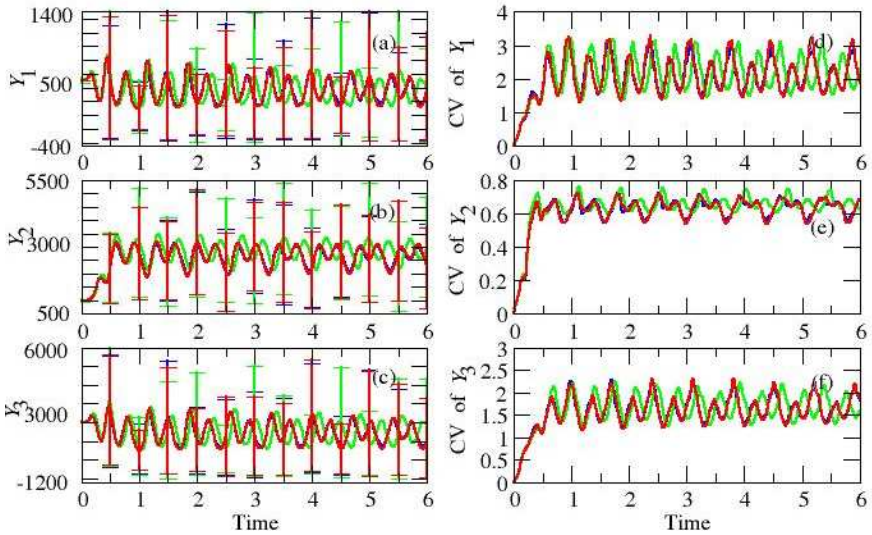


Figure 5.10: The trajectories of the means with $\pm$ 1SD error bars [(a)-(c)] and the CVs [(d)-(f)] for the probability distributions of the species Y_1,Y_1,Y_3 using SSA (blue curve), G-P (green curve) and RRA-Noise (red curve).

Table 5.9: The average values of the CPU time (secs) and the number of steps taken by different simulation methods for the case of the Oregonator model.

Simulation Methods	SSA	G-P	RRA-Noise
CPU time (secs)	70.069	14.976	35.224
Steps	694868	16542	29125

The comparison of the CPU times of different methods in Table 5.9 indicates that the G-P is more efficient than the rest of the methods. However, their simulated profiles indicates a slight outward shift relative to the trajectories of other methods. It is also observed that the G-P trajectories are not within the SSA trajectories in terms of $\pm$ 1SD error bars. On the other hand, the RRA-Noise is found to take relatively more steps, but is able to reproduce the trajectories accurately. More importantly, no negative molecular numbers have been found during the simulations.

The numerical stability of this model system is discussed along the lines similar to

those of the previous example. The given model is simulated by the SSA using an ensemble of 20 000 (black curve) simulation runs, while the RRA-Noise is used for the simulation over an ensemble of 100 (red curve), 500 (green curve), 1000 (blue curve), 5000 (brown curve) and 10 000 (orange curve) simulation runs. Unlike the previous example, this is an oscillatory model with no steady state. It is found that the absolute error at some discrete points decreases with an increase in the number of realizations of the RRA-Noise.

It has been found that with the increase in the number of realizations, the RRA-Noise gets converged to the SSA trajectories with decrease in the error. The error between the trajectories of the SSA and the RRA-Noise has been calculated at some chosen discrete time points. Figure 5.11 shows the mean trajectories ((a)-(c)) of the species, while Table 5.10 provides the absolute errors for the different realizations at specific time points.

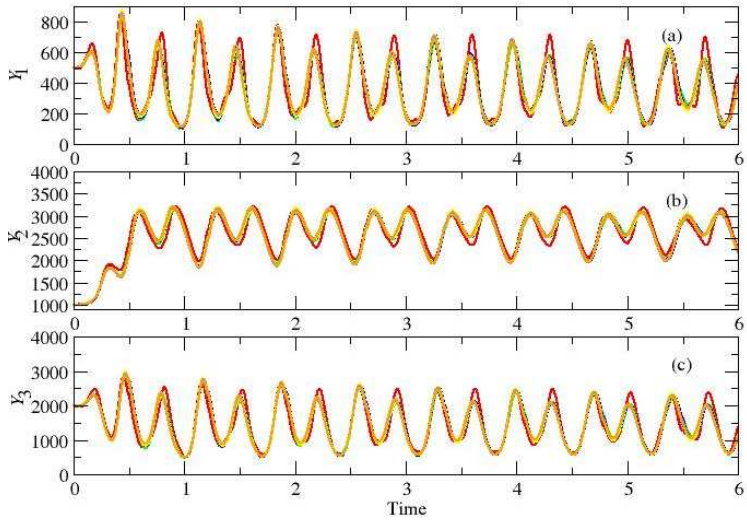


Figure 5.11: The trajectories of the means [(a)-(c)] simulated using RRA-Noise with 100 (red curve), 500 (green curve), 1000 (blue curve), 5000 (brown curve), 10000 (orange curve) runs and the SSA with 20000 runs (black curve), for the case of the Oregonator model.

In Figure 5.11, it is observed that the simulation by the RRA-Noise with 100 runs (red curve) shows a downward and upward shift relative to the SSA trajectory with 20 000 runs (black curve). The trajectory with 10 000 runs (orange curve) gets substantially closer to the exact SSA trajectory. The variation in the absolute error in Table 5.10 is

seen as a signature of the highly oscillatory character of the chemical system.

Table 5.10: The absolute errors between the trajectories of the SSA and the RRA-Noise for different runs at discrete time points.

Time points for 100 runs			
1	11.396	161.357	18.711
2	22.308	48.094	194.045
3	2.155	11.085	160.426
4	0.985	52.447	43.190
Time points for 500 runs			
1	3.174	6.695	17.408
2	41.302	111.648	77.653
3	11.188	72.191	49.520
4	49.485	41.032	29.818
Time points for 1000 runs			
1	2.001	21.669	8.416
2	30.584	109.157	22.778
3	7.388	56.546	34.213
4	49.250	38.105	94.080
Time points for 5000 runs			
1	0.979	47.324	2.136
2	6.969	61.696	5.923
3	17.139	9.866	61.658
4	22.457	24.735	84.764
Time points for 10000 runs			
1	1.304	14.978	8.993
2	3.192	9.973	0.945
3	13.437	26.698	47.993
4	5.367	25.048	10.685

Overall, from the simulations of the five different examples, discussed in the sections above, it can be speculated that the newly developed method, which uses the RRA and accounts for noise during the simulation solves the problem of negative populations. In all the examples considered, it was seen that, unlike in the RRA-Noise case, one has to find the best possible value of the coarse grain factor for the BD-τ method of Chatterjee *et al.* in order to achieve the necessary accuracy, a choice that usually led to a loss of efficiency for the method.

The examples that have been chosen and discussed in the current work were those that highlight difficult cases where the existing state-of-the-art methods either fail or perform with lower efficiency. However, it is to be noted that the current approach is not a general theoretical modification in stochastic simulations for correcting the number of firings for every leap during the simulations, but is a remedy to the negative population problem for specific leaps where the reactant population becomes negative due to the wrongly calculated number of firings, for those leaps, by existing methods. Therefore, it is still possible that the current method, while clearly having been demonstrated to have performed well for the examples considered, might also provide negative numbers for reactant population in certain cases, and thus fail for certain chemical systems. It is, nevertheless, expected, that the current recipe for correcting the problem of negative populations would work in a large majority of cases, as the current set of examples demonstrates.

5.4 Conclusions

In order to achieve a speed-up over the SSA, various approximate accelerated methods have been developed. However, such approaches are fraught with problems of accuracy, problems that become more acute when dealing with chemical systems that deal with low molecular populations. In such cases, there are instances where negative molecular numbers have been obtained during the simulations. In the current work, we have sought to solve this problem by introducing the novel concept of accounting for the noise obtained for the number of firings of each reaction in a given time step. We have tested out this idea by combining the noise accounting with the accelerated method : the Representative Reaction Approach (RRA) that we had developed earlier.[50] This new method, termed as RRA-Noise, has been tested on a number of different examples, ranging from simple unimolecular system to oscillatory chemical system. It has been found that the

RRA-Noise is effective not only in terms of accuracy but also in efficiency, in comparison to state-of-the-art approximate accelerated methods such as Gillespie's Approximate Stochastic Algorithm (GASA),[7] Gillespie-Petzold (G-P),[29] BD-τ of Chatterjee *et al.*.[92] This newly developed method has the added virtue of being quite simple and easy to code. The discussion pertaining to the stability of algorithm emphasizes the robustness of newly proposed method. Furthermore, for the newly proposed method, there is no necessity to change the value of error control parameter for every new chemical system that has to be simulated. Finally, it may also be mentioned that the notion of accounting for noise during a simulation may find applications in other fields of interest as well.

CHAPTER 6

Stochastic Simulation of Chemical Kinetics using a New Strategy to Choose the Error Control Parameter

Anything you dream is fiction,and
anything you accomplish is science,
the whole history of mankind is
nothing but science fiction.

-Ray Bradbury

Abstract

Computational modeling has become an indispensible tool for studying kinetic behavior, especially after advances in approximate accelerated stochastic methods has allowed the combination of efficiency and accuracy in dealing with chemical systems. However, all accelerated approaches rely on the choice of an error control parameter, ε, in order to determine the size of the time step. The choice of ε in these methods is usually arbitrary and restricted, which sometimes leads to an adverse effect on the results of the simulations. In the current work, we propose a novel simulation method by means of which ε can be chosen in both a flexible and logical manner. The simulation results from the methods proposed in the current study have been found to be on par, or better, than the other state-of-the-art accelerated simulation approaches.

6.1 Introduction

The interdisciplinary field of computational biology aims to model the complex interactions within biological systems based on some experimental observations. These interactions are modeled mathematically by a set of coupled, first-order, ordinary differential equations (ODEs), known as the reaction rate equations (RREs). However, with the advent of advanced experimental techniques, it has been confirmed that stochasticity is of paramount importance in biological systems. The stochasticity and the subsequent fluctuations[33, 120, 126] have been attributed to the presence of a very small number of molecules in cellular systems. The RREs, also termed as the deterministic model, which consider the time evolution as a continuous process, fail to take into account the stochastic fluctuations inherent in such systems. This issue has been more succesfully addressed by the kinetic Monte Carlo[54] - based Doob-Gillespie algorithm.[117, 118, 28, 13] This probabilistic model, also known as the stochastic simulation algorithm (SSA),[13] has been widely employed for studying the dynamical behavior of reactants, intermediates and products for a given reaction network. The SSA simulates a single reaction firing from the entire reaction network in each time step until some desired time has been reached. The simulation of every successive reaction event produces accurate realizations of the time trajectories, but in case of realistic cellular systems, the simulations are very slow and computationally expensive. Several different methods have been subsequently developed as potential improvements to the SSA, such as the next reaction method,[73] the optimized direct method[74] and the sorting direct method.[75] Other than this, Barrio *et al.* have generalized the SSA to the delay stochastic simulation algorithm (DSSA),[81] in order to consider the time delays in addition to the discreteness and intrinsic noise. However, all these methods have not been entirely successful in their attempt to reduce the computational load of the SSA. More approximate methods have therefore been developed as alternatives to

the exact SSA approach. One set of such approaches corresponds to the hybrid methods,[101] which have been proposed for the multiscale simulation of chemically reacting systems. These methods couple the RREs, or the chemical Langevin equation,[25] with the SSA. The reaction network is partitioned into the subsets of fast reactions and slow reactions. The fast reactions (or the species with larger molecular numbers) are simulated by deterministic models, while the slow reactions (or the species with smaller molecular numbers) are simulated by the SSA. In order to accommodate the species in the fast reactions having a small number of molecules, the multiscale SSA with the partial equilibrium assumption[86] and the slow-scale SSA[85] have been employed. However, these methods, while having achieved some success at reducing the computational load of the SSA, have also become more unweildy, yielding complicated algorithms, in contrast to the elegant SSA approach.

Another approach attempts to reduce the computational load of the SSA by sacrificing some of the accuracy of the simulations. In one such attempt, Gillespie introduced the idea of Gillespie's Approximate Stochastic Algorithm (GASA).[7] In this method, by taking a large enough pre-selected time step (leap), several firings of each reaction are allowed to occur. The time step is derived from the leap condition, which ensures that there is no significant variation in the propensity functions (the product of rate constant and the number of reactant combinations) of the individual reactions. The individual reaction events are sampled from the Poisson Distribution. Further modifications to the GASA have also been proposed, which include the Gillespie-Petzold (G-P) method,[29] the implicit tau-leaping method of Rathinam *et al.*,[84] the efficient step size method of Cao *et al.*,[104] the N-leap method of Xu and Lan,[96] the representative reaction approach (RRA)[50] and a course-time-step method.[134] Tiejun Li[135] has done the convergence analysis of explicit tau-leaping methods by using the stochastic differential equations. Anderson *et al.*[136] have provided the error analysis of tau-leaping methods. However, the use of Poisson random numbers to sample the reaction events of individual reactions may give rise to negative numbers during the simulations by the above mentioned methods. This

problem is partially solved by methods that employ binomial distribution based random numbers. These include the leaping methods of Tian-Burrage,[91] Chatterjee *et al.*,[92] the multinomial tau leap method,[93] the modified binomial leap method,[94] the generalized binomial leap method[127] and the recently proposed RRA modified methods[51] that use the SSA and the binomial distribution.

In all of the above-mentioned accelerated leaping methods, the calculated time step satisfies the leap condition,[7,29] with ε as an error control parameter. While this approach, in general, increases the size of the time step over the time step obtained with the SSA, it is to be noted that the choice of the value of ε in such methods is completely arbitrary. The ε chosen for the simulation of one particular chemical system might not be a good choice for the simulation of some other system. The choice of some other value of ε may or may not give reliable simulation results. The numerical variation in the value of ε also leads to different simulation results in terms of accuracy and computational time.

Thus, an accelerated stochastic approach which makes use of a sound and flexible way to choose an error control parameter would certainly have an edge over the conventional approximate methods that all employ the ε. In this work, a new approximate accelerated method based on such a notion has been proposed. This has been achieved by taking our recently developed representative reaction approach (RRA)[50] and combining it with the idea of a coupled harmonic oscillator (CHO) as a representation of the chemical system in question. As the following sections of the manuscript will demonstrate, this original approach leads to an approximate simulation algorithm that chooses ε based on mathematical considerations. It provides a viable method for successfully simulating chemical systems with accuracies and speeds comparable to, and, in some cases, better than the existing approximate methods. More importantly, it has also been found to solve the problem of negative populations for certain chemical systems.

The rest of the paper is organized as follows : In the Methodology section, we have

discussed in brief the framework of the representative reaction approach (RRA), the concept of representing the system as a coupled harmonic oscillator, a mathematical strategy to choose the error control parameter and the methodology of the newly proposed method along with the steps of execution. In Results and Discussion section, simulations of different examples which confirm the reliability and efficiency of the newly proposed approach are discussed. The conclusions are provided in the last section.

6.2 Background

In general, the SSA can be used to simulate a well-stirred mixture, in thermal equilibrium at temperature T, of *N* molecular species $\{S_1, ..., S_N\}$ which are interacting with each other through *M* chemical reactions $\{R_1, ..., R_M\}$. Every reaction is characterized by a quantity called the propensity function $a_j(\mathbf{x})$, which is the product of the number of reactant combinations with the rate constant for a given reaction. The other property of every reaction is the state change vector $\boldsymbol{\nu}_i \equiv \{\nu_{1i}, ..., \nu_{Ni}\}$. Here, a_i**(x)***dt* gives the probability that the reaction R_i will occur in the next infinitesimal time interval [*t*, *t*+*dt*), and ν_{ij} is the algebraic change in the number of molecules between the reactant and the product for the R_i reaction. For more details, please see the original papers published by Gillespie and co-workers on this approach.[28,13]

We have recently proposed the representative reaction approach (RRA)[50] in an attempt to reduce the computational load of the SSA. In this simulation technique, the chemical system which consists of many reactions has been represented by a single, representative reaction (RR) $2A \longrightarrow B$. Hence, this method has been termed as the representative reaction approach (RRA). Like other individual reactions, the RR $2A \longrightarrow B$ is characterized by a propensity function, a_0**(x)**, which is taken as the sum of the propensity functions of all the individual reactions, and the rate constant, C_0, which is the weighted average of all the rate constants. Since a_0**(x)** is the propensity function of the RR $2A \longrightarrow B$, and the rate constant of this reaction is also known, one can then also determine

the amount of the hypothetical species x_0 corresponding to the reactant A in the RR *2A* $\longrightarrow$ *B*. Considering the system to be comprised of this single representative reaction, the time step, τ, for any given iteration can then be determined. The reaction numbers associated with each individual reaction in the system are then sampled from the Poisson distribution, with a_i**(x)**τ as the expected number of reactions of reaction R_i in the time step τ.

The RRA has been found to be successful in simulating all types of chemical systems.[50] Furthermore, in more recent work, we have demonstrated that employing the RRA in conjunction with the Binomial distribution leads to a potential solution to the problem of negative populations.[51]

6.3 R_CHO : Representing the System as a Coupled Harmonic Oscillator

As mentioned in the previous section, in our recently proposed method, the RRA, the chemical system to be simulated is characterized by two fundamental quantities, namely - the propensity function, a_0**(x)**, and the reaction rate constant, C_0. What is proposed as a new approach is to further represent the system as two bodies of equal masses, *m*, attached to three springs, each with a spring constant, *k* : the coupled harmonic oscillator (CHO). Representing the system as a CHO will, as shown in subsequent sections, allow the determination of the time step, τ , for any given iteration.

For this representative CHO, the spring constant, *k* is correlated to a_0**(x)**, the propensity function of the chemical system. The reasoning behind this is as follows : the value of *k* governs the dynamical behavior of the coupled mass-spring system, just as the value of a_0**(x)** is instrumental in deciding the size of the time step, τ, in the chemical system. Furthermore, the mass, *m*, which measures the amount of matter contained in the system,

is a fundamental property of the CHO. In continuing the analogy with the chemical system, one can then make the correlation of *m* with the reaction rate constant, C_0, which is a specific, characteristic property of the chemical system.

Therefore, in order for this CHO to represent the chemical system, the fundamental quantities that characterize the CHO : its *k* and *m*, can be correlated to the quantities a_0**(x)** and C_0 respectively of the chemical system.

Now, the angular frequency, ω, of the CHO is as follows :

$$\omega = \sqrt{\frac{k}{m}} \tag{6.1}$$

For the specific CHO that represents the chemical system, the angular frequency, ω now is equated to

$$\omega = \sqrt{\frac{a_0}{C_0}} \tag{6.2}$$

This new representation of the chemical system by a CHO, and the corresponding correlation between the different quantities in the system, is illustrated in Figure 6.1.

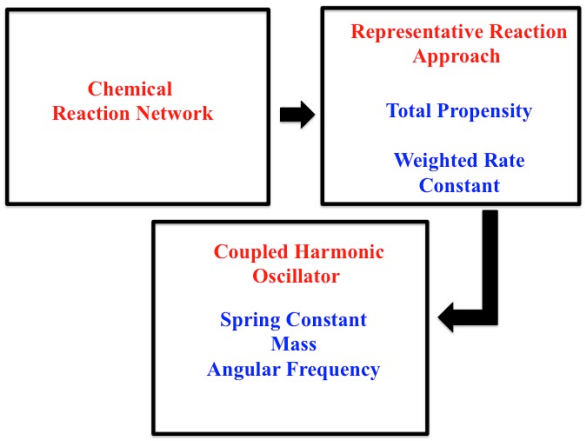


Figure 6.1: A block diagram showing the correlations between the chemical system to be simulated, the Representative Reaction Approach (RRA) where the entire chemical system is represented as a single reaction, and the relation of the fundamental quantities of this representative single reaction chemical system with the quantities of the system when represented as a coupled harmonic oscillator (CHO).

The value of having a CHO represent the chemical system is that this provides an

avenue for determining the time step of the system, for any given iteration. This is elaborated upon in the next section.

6.4 Mathematical Recipe

In case of the Coupled Harmonic Oscillator (CHO), the displacement, *x*, is given by :

$$x(t) = Acos(\theta + \phi_1) - Bsin(\theta + \phi_2) \tag{6.3}$$

Where, $\theta = \omega t$

Here, *A* and *B* are the amplitudes of the oscillations, ω is the angular frequency, ϕ_1 and ϕ_2 are the phase angles and *t* is the cumulative elapsed time.

Considering a representative CHO that has *A* = 1, *B* = 1 and $\phi_1 = \phi_2 = 0$, we have

$$x(t) = cos(\theta) - sin(\theta) \tag{6.4}$$

Now, $\theta = \omega t$, and, ω, if we were to employ a CHO as a representative of the system, would be given by the equation : $\omega = \sqrt{\frac{a_0}{C_0}}$, as discussed in the previous section.

Hence,

$$\theta = \sqrt{\frac{a_0}{C_0}}t \tag{6.5}$$

Therefore, for any given iteration, there would be a specific value for ω, depending on the values of a_0**(x)** and C_0 at that given iteration. Therefore, at every time-step, the system would be represented by a specific CHO, with a specific angular velocity, ω, different from the preceding and the succeeding time step.

The objective of representing the chemical system by a CHO is to obtain the value of the time step, τ, or the size of the leap, for that given iteration in the chemical system. This is done by determining a specific value of θ, by following the mathematical procedure described in the next paragraph. Once θ is determined, the value of the time *t* corresponding to that value of θ can be determined from Eq. (6.5), as $t = \sqrt{\frac{C_0}{a_0}}\theta$. This *t* value can be considered to be τ if the determined value of θ represents a small change

in the dynamical behavior of the system consistent with Gillespie's leap condition:[7] the time evolution during the stochastic simulation of a chemical system should proceed in such a way that the propensity function of each chemical reaction in the system remains nearly unchanged while going from one step to another.

How the specific value of θ, discussed above, is obtained is described in the following few paragraphs.

In dynamical systems theory, the concept of an Arnold tongue[137,138] is used to study the region in the space of parameters. It was used earlier to study the dynamical systems on the circle. The resulting equation of the dynamical behavior is termed as the circle map equation.[137,138] The circle map is given by :

$$\theta_{n+1} = \theta_n + \Omega - \frac{K}{2\pi} sin(2\pi\theta_n) \tag{6.6}$$

Here, θ is the polar angle whose value lies between 0 and 1, *K* is the coupling strength and Ω is the externally applied frequency.

The Jacobian of the circle map is given by :

$$\frac{\partial\theta_{n+1}}{\partial\theta_n} = 1 - Kcos(2\pi\theta_n) \tag{6.7}$$

In our case, the idea of Arnold tongue[137,138] is applied to the system of the coupled harmonic oscillator (CHO). The parameters in the circle map equation, namely, the coupling strength, *K* and the externally applied frequency, Ω are replaced by the parameters : total propensity function, a_0**(x)** and the angular frequency, ω of the system in question. Thus, the resulting equations become :

$$\theta_{n+1} = \theta_n + \omega - \frac{a_0\mathbf{(x)}}{2\pi} sin(2\pi\theta_n) \tag{6.8}$$

$$\frac{\partial\theta_{n+1}}{\partial\theta_n} = 1 - a_0\mathbf{(x)}cos(2\pi\theta_n) \tag{6.9}$$

In order to ensure that the Leap Condition is satisfied, it is hypothesized that the subsequent values of θ change by a minuscule amount. It turns out that the small change ($\theta_{n+1} \approx \theta_n$) in their values introduces an error, viz. $\delta_1 = |\theta_{n+1} - \theta_n|$. The proximity in

the values of θ makes the change in θ_{n+1} w. r. t. θ_n, i. e., $\delta_2 = \frac{\partial\theta_{n+1}}{\partial\theta_n}$, small. These small changes introduces errors. Since these random errors follows the normal distribution, the errors δ_1 and δ_2 are approximated by random deviates that follow a normal distribution.[72] Hence, the modified equations become :

$$\theta_n = \frac{1}{2\pi}\sin^{-1}\left[\frac{2\pi}{a_0\mathbf{(x)}}(\omega + |\delta_1|\right] \tag{6.10}$$

$$\theta_n = \frac{1}{2\pi}\cos^{-1}\left[\frac{1-\delta_2}{a_0\mathbf{(x)}}\right] \tag{6.11}$$

These values of θ_n are further used for calculating the time step, τ for any iteration as given below :

$$\tau = \sqrt{\frac{C_0}{a_0}}\theta_n \tag{6.12}$$

6.5 Methodology

For any given iteration, the time step, τ, is calculated from Eq. (6.12), with θ_n calculated from Eq. (6.11) for the specific a_0**(x)** and the C_0 values determined for that particular step. Next, the time step, τ with θ_n from Eq. (6.11), is used for the calculations of the expected number of reaction events, $a_i\tau$, of each reaction, followed by the sampling of the reaction numbers by using the Poisson probability distribution.[72] Finally, the amount of the reactants in the chemical system is updated according to the stoichiometry. In case any negative molecular numbers are obtained during the simulations, the time step, τ is recalculated using the θ_n value obtained from Eq. (6.10) for the same values of a_0**(x)** and C_0 for that particular step. The rationale behind using Eq. (6.10) is that it gives relatively small values of the time step than that obtained using Eq. (6.11). Thus, it shields the algorithm form violating the Leap Condition and also ensures that only positive molecular numbers are obtained during the simulations. The time steps obtained from Eq. (6.11) are fairly large, which helps to increase the efficiency of the simulations. In case of occurrence of negative numbers (or subsequent loss of accuracy) a switch to Eq. (6.10) gives time steps relatively small. These small time steps ensures the simulations are

accurate. Thus, the occasional shifts to these two regimes maintain the efficiency and accuracy of the simulations. More importantly, unlike other accelerated methods this has been achieved without any arbitrary choice of the error control parameter.

6.6 Steps for the Implementation of RRA_CHO

Based on the discussion above, the implementation steps for RRA_CHO are outlined as follows :

- **Step 1 :** input the initial number of species and the rate constants of the constituent reactions; initialize the counters and the random number generators to a seed value and transfer the initial number of species to some temporary locations (variables).
- **Step 2 :** calculate the propensity functions : $\{a_1, ..., a_M\}$ the sum of the propensity functions : a_0**(x)** = $\sum_{j=1}^{M} a_j$**(x)**
- **Step 3 :** calculate the rate constant : C_0 = $\sum_{j=1}^{M} \left(\frac{a_j\textbf{(x)}}{a_0\textbf{(x)}}\right) c_j$
- **Step 4 :** calculate the parameter : $\theta_n = \frac{1}{2\pi} \cos^{-1}\left[\frac{1-\delta_2}{a_0\textbf{(x)}}\right]$,
  where δ_2 = *gasdev*(*iseed*) is a normal (Gaussian) random deviate
- **Step 5 :** calculate the time step : $\tau = \sqrt{\frac{C_0}{a_0}}\theta_n$
- **Step 6 :** calculate the expected number of reactions : $n_i = a_i\tau$
- **Step 7 :** use the Poisson random number generator to find the k_{is}, the actual number of reactions for the individual reactions :
  $k_i = poidev(n_i, iseed)$
- **Step 8 :** make the necessary changes in the molecular population using the appropriate stoichiometric parameters and reaction numbers
- **Step 9 :** if negative numbers are not found, go to step (2); else discard the step and use the initial species stored for that step in the temporary locations.

• **Step 10 :** calculate the parameter : $\theta_n = \frac{1}{2\pi}\sin^{-1}\left[\frac{2\pi}{a_0(\mathbf{x})}(\omega + |\delta_1|\right]$, where δ_2 = *gasdev*(*iseed*) is a normal (Gaussian) random deviate

• **Step 11 :** follow the steps from (5)-(9)

• **Step 12 :** go to step (2)

6.7 Results and Discussion

The newly proposed method discussed in the previous section has been employed for three different examples that are discussed below. In order to compare their accuracy and efficiency with other state-of-the-art methods, the three examples have also been studied with the Stochastic Simulation Algorithm (SSA), Gillespie's Approximate Stochastic Algorithm (GASA),[7] Gillespie-Petzold (G-P) method[29] and the Binomial distribution based tau leap (BD-τ) method of Chatterjee *et al.*.[92]

6.7.1 Four Reaction Model

The Four Reaction Model[7] consists of the following set of reactions :

$$R_1 : X_1 \xrightarrow{c_1} \phi$$

$$R_2 : 2X_1 \xrightarrow{c_2} X_2$$

$$R_3 : X_2 \xrightarrow{c_3} 2X_1$$

$$R_4 : X_2 \xrightarrow{c_4} X_3$$

Here, X_1, X_2 and X_3 are the species taking part in the different reactions; while c_1, c_2, c_3 and c_4 the rate constants for the respective reactions.

The numerical values of the initial molecular species and the rate constants of the reactions are given in Table 6.1.

Table 6.1: The values of the rate constants of the reactions and the initial species population for the four reaction model.

Parameters used in the simulation	Numerical values of the parameters
c_1	1.0
c_2	0.002
c_3	0.5
c_4	0.04
X_1	10000
X_2	0
X_3	0

The simulated trajectories of the first two moments of the probability distributions by different stochastic simulation methods : the SSA (Blue), the Gillespie's Approximate Stochastic Algorithm (GASA) (Green), the Gillespie-Petzold (G-P) (Magenta) as well as the newly proposed RRA_CHO (Red) are depicted in Figure 6.1.

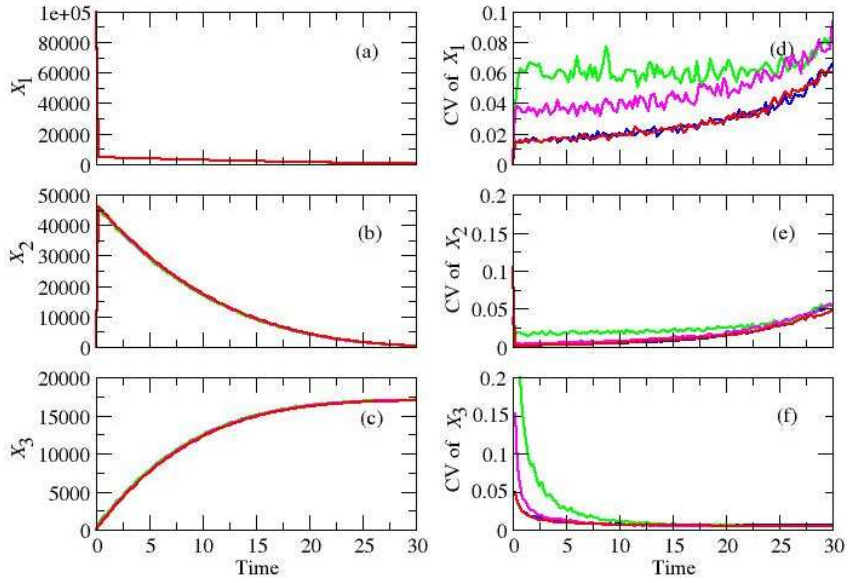


Figure 6.2: The trajectories of the means [(a)-(c)] and of the CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 obtained over 100 simulation runs of (blue curve), GASA (green curve), G-P (magenta curve), and RRA_CHO (red curve).

The error control parameter, ε, with a standard value of 0.03 was used for doing the simulations with GASA and G-P. The values of ε are the standard values that have been employed by the respective methods.[7,29] These ε values have been used for all the subsequent examples simulated by these methods in this manuscript. The values of the means and the CVs in Figure 6.2 are calculated over an ensemble of 100 simulation runs.

Figures 6.2 (a-c) show the time evolution of the first moment of the probability distribution : the mean for the different species, using the simulation methods mentioned above. Likewise, Figures 6.2 (d-f) show the time evolution of the second moment of the probability distribution : the Coefficient of Variation (CV). The trajectories of the means in the Figures 6.2 (a-c) obtained by different simulation methods are in good agreement with each other. However, the trajectories of the CVs in Figures 6.2 (d-f), obtained by GASA and G-P, are not on par with those obtained by other simulation methods, namely - the SSA and RRA_CHO. This loss of accuracy in simulations by the GASA and G-P methods is attributed to the larger time steps for the iterations, and thus, in their failure

Table 6.2: The average values of the CPU time (secs) taken by different simulation methods for the four reaction model.

Simulation Methods	SSA	GASA	G-P	RRA_CHO
CPU time (secs)	9.906	0.270	0.311	4.935

to obey the leap condition.

The CPU times of different simulation methods, collected in Table 6.2, show that for this example, GASA and G-P offer a distinct advantage relative to the other simulation methods. However, as discussed above, this comes at the cost of a loss of accuracy for these two methods. On the other hand, the simulated trajectories (i.e. means and CVs) of the newly developed RRA_CHO method have a proper fit with their counterpart trajectories of the SSA. This indicates that the handicap of the CPU time performance of the RRA_CHO has been overcome by the superior accuracy of the CVs for this method. Apart from this, it is also to be noted that the existing methods use ε values that have been tuned to be effective for such chemical systems. It is possible that the same ε values might prove less effective for more challenging chemical systems, such as the Carletti-Burrage Reaction Model, which, being prone to the problem of negative molecular numbers, is much more difficult to simulate. This model is discussed in the next subsection.

6.7.2 The Carletti-Burrage Model

The following reaction network model was proposed by Carletti and Burrage,[94] which consists of ten reactions will be discussed here. This chemical system has the following set of reactions :

$$R_1 : \text{RNA} \xrightarrow{c_1} \text{DNA1}$$

$$R_2 : \text{DNA1} \xrightarrow{c_2} \text{RNA}$$

$$R_3 : \text{m} \xrightarrow{c_3} \text{RNA}$$

$$R_4 : \text{RNA} \xrightarrow{c_4} \text{m}$$

$$R_5 : \text{2m} \xrightarrow{c_5} \text{D}$$

$$R_6 : \text{D} \xrightarrow{c_6} \text{2m}$$

$$R_7 : \text{DNA} + \text{D} \xrightarrow{c_7} \text{DNA1}$$

$$R_8 : \text{DNA1} \xrightarrow{c_8} \text{DNA} + \text{D}$$

$$R_9 : \text{DNA1} + \text{D} \xrightarrow{c_9} \text{DNA2}$$

$$R_{10} : \text{DNA2} \xrightarrow{c_{10}} \text{DNA1} + \text{D}$$

where RNA, DNA, DNA1, DNA2, D and m are the species taking part in the different reactions; and the symbols ($c_1 - c_{10}$) over the arrows indicate the rate constants of the respective reactions.

The numerical values of the initial species population and the rate constants of the reactions are given in Table 6.3.

Table 6.3: The values of the rate constants of the reactions and the initial species population for the Carletti-Burrage model.

Parameters used in the simulation	Numerical values of the parameters
c_1	0.078
c_2	3.9E-3
c_3	7.0E-4
c_4	0.043
c_5	0.083
c_6	0.5
c_7	0.020
c_8	0.479
c_9	2.0E-4
c_{10}	8.765E-12
m	200
D	600
DNA	200
DNA1	0
DNA2	0
RNA	0

It is to be noted that the simulation of this particular system by GASA was found to lead to physically unrealistic (negative) numbers. Hence the results with the GASA have not been discussed further. Moreover, the binomial distribution based tau (BD-τ) method of Tian-Burrage[91] could not be applied for this model, since there are some species which take part in multiple reactions: a situation that the BD-τ method of Tian-Burrage is incapable of handling, making it technically non-applicable for such reaction networks. Hence, only the methods that were successfully able to reproduce the simulation trajectories are reported here. Hence the SSA, G-P, the BD - τ method of Chatterjee-Vlachos-Katsoulakis and the newly proposed RRA_CHO have been discussed in the figure and in the paragraphs below. The colors pertaining to the different methods remain the same.

Figures 6.3 (a-c) show the trajectories of the means of the species involved in the system, while Figures 6.3 (d-f) shows the corresponding trajectories of the CVs.

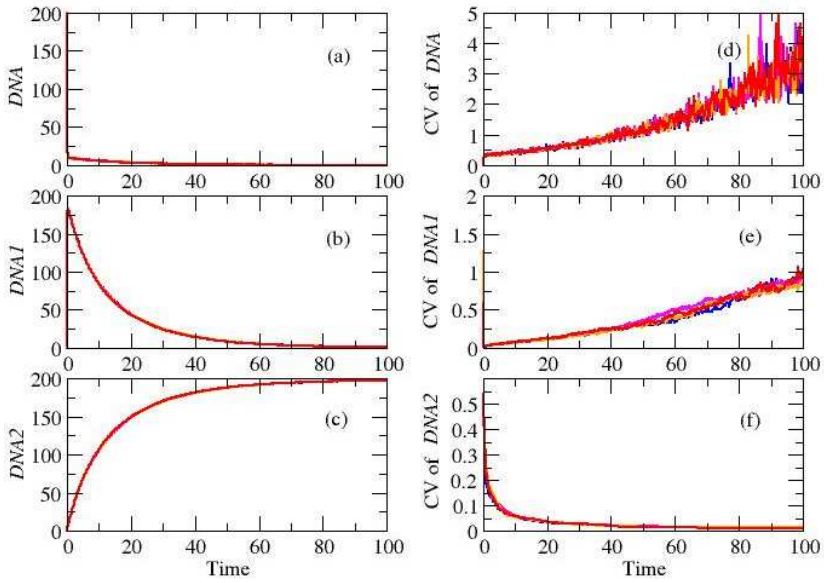


Figure 6.3: The trajectories of the means [(a)-(c)] and of the CVs [(d)-(f)] for the probability distributions of the species DNA, DNA1, DNA2 obtained over 100 simulation runs of (blue curve), G-P (magenta curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (orange curve) and RRA_CHO (red curve).

Table 6.4: The average values of CPU time (secs) for 100 simulations taken by different simulation methods for the Carletti-Burrage Model.

Simulation Methods	SSA	G-P	BD-τ	RRA_CHO
CPU time (secs)	1.065	3.107	8.222	3.242

The CPU time values in Table 6.4 show that the stochastic simulation algorithm is faster than all the other methods used for the simulation. All the methods agree with each other to a good extent, as shown in Figure 6.3. In case of the BD-τ method, the time steps are taken with a coarse grain factor,[92] *f*, taken as 2.0. The accuracy of the simulations is more guaranteed by using the smaller values of the coarse grain factor. However, this also contributes to an increase in the CPU time of simulations. The increase in the value of f reduces the CPU time, but this now leads to the loss of accuracy in the simulations. This is observed in the simulations performed by taking a coarse grain factor, *f*, as 4.0. The CPU time of G-P is slightly less than that of newly developed RRA_CHO. The CPU time of G-P is contributed by the share of SSA during the simulations. This example is prone to the problem of negative numbers emphasizes that the RRA_CHO method tackles the problem quite efficiently than the BD-τ method, which had been specifically developed for the problems of negative populations.

Thus, this example shows that the currently proposed approach can also be used to solve the problem of negative numbers during the simulations.

6.7.3 Model of First Order Reactions

In this sub-section, the simulation of the model of four unimolecular reactions is discussed. This model was used by Chatterjee *et al.*[92] to test their BD-τ method. This model has more complexity, as the species X_2 takes part in several of the reactions. The chemical system is made up of the following reactions :

$$R_1 : X_1 \xrightarrow{c_1} X_2$$

$$R_2 : X_2 \xrightarrow{c_2} X_3$$

$$R_3 : X_3 \xrightarrow{c_3} X_2$$

$$R_4 : X_2 \xrightarrow{c_4} X_1$$

Here, X_1, X_2, and X_3 are the species taking part in four different reactions and c_1, c_2, c_3 and c_4 are the rate constants of these reactions.

The numerical values of the model parameters used for the simulations are given in Table 6.5.

Table 6.5: The values of the rate constants of the reactions and the initial species population for the model of first order reactions.

Parameters used in the simulation	Numerical values of the parameters
c_1	2.0
c_2	1.0
c_3	2.0
c_4	1.0
X_1	20000
X_2	0
X_3	0

The simulation of this chemical system by GASA and G-P with the usual error control parameter gives very poor trajectories. This behavior is attributed to the failure of the leap condition when employing these two methods. Hence, GASA and G-P have not come in the discussion of this chemical system. Thus, discussed below are the results of simulating the chemical system by SSA, BD-τ of Chatterjee-Vlachos-Katsoulakis and the newly developed RRA_CHO.

Figures 6.4 (a-c) show the time evolution of the mean values of the species: X_1, X_2, X_3 by the different simulation methods, while the Figures 6.4 (d-f) shows the CVs of

the corresponding species. From the Figure 6.4, it is clear that the different simulation methods have successfully simulated the time behavior of different molecular species in this system. The BD-τ of Chatterjee-Vlachos-Katsoulakis (orange curve) and the RRA_CHO (red curve) are as good as the SSA (blue curve).

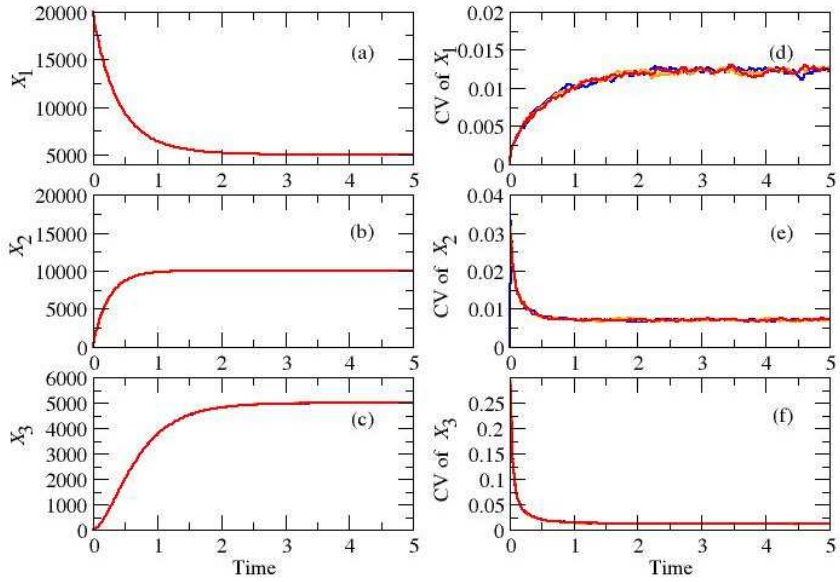


Figure 6.4: The trajectories of the means [(a)-(c)] and the CVs [(d)-(f)] for the probability distributions of the species X_1, X_2, X_3 using SSA (blue curve), BD-τ of Chatterjee-Vlachos-Katsoulakis (orange curve) and RRA_CHO (red curve) for the case of First Order Reactions.

Table 6.6: The average values of CPU time (secs) taken for 500 simulations by different simulation methods for the First Order Reactions.

Simulation Methods	SSA	BD-τ	RRA_CHO
CPU time (secs)	18.141	59.208	9.687

The average CPU times shown in Table 6.6 indicates that RRA_CHO is computationally more efficient than the other methods. It is almost two times as fast as the SSA. The low computational efficiency of BD-τ can be imputed to the small size of the time steps, leading to a large number of steps. It has been found that, the increase in the value of coarse grain factor leads to a significant loss of accuracy. This is found that with the coarse grain factor, f = 5000.0, the means and the CVs for the different species becomes far less accurate for the BD-τ case in comparison to the other methods. Further increase in the coarse grain factor to 10000.0 also leads to similar results.

As the results shown in Figure 6.4 and Table 6.6 indicate, this model of first order reactions is a classic example for cases where, in addition to the accuracy, the CPU times of the corresponding methods are of prime importance. The newly proposed RRA_CHO turns up as the most favorable approximate simulation method for this case.

6.8 Conclusions

The approach of investigating chemical kinetics through stochastic simulations, instead of through deterministic methods, has seen important developments, with new accelerated, approximate approaches[7,29,50] having been developed in order to study all types of chemical systems. However, all the accelerated approaches developed to date rely on the use of an error control parameter, ε, for modulating the speed of the simulations, a parameter whose value is arbitrarily assigned. The current work attempts to overcome this dependence on an arbitrary parameter by providing flexibility for the choice of ε in a logical and mathematical fashion. This has been achieved by representing the chemical

system as a Coupled Harmonic Oscillator (CHO), and by evaluating the time step by following the time evolution of the CHO, for any given iteration, over a period of time. This novel accelerated method has been found to be effective not only for successfully simulating chemical systems but also for simulating those that have the problem of negative molecular numbers.

CHAPTER 7

Conclusions and Future Work

Science, my lad, is made up
of mistakes, but they are
mistakes which it is useful
to make, because they lead
little by little to the truth.

-Jules Verne

The use of the state-of-the-art experimental techniques have underlined the inherent stochastic nature of the interactions in biological systems. The complexity of these interactions makes the aspect of mathematical modeling essential for the study of such systems. However, the of modeling systems having discrete integer species population turns out to be a demanding task with the conventional methods, thus paving the way for an alternative approach, that of - stochastic simulation methods. The kinetic Monte Carlo based stochastic simulation algorithms have been successfully applied for simulations of biological systems. One of the pitfalls of these algorithms is the enormous computational time taken for the simulations of realistic reaction networks.

The aforementioned issue of computational time has been resolved by the development of approximate accelerated stochastic simulation algorithms. This dissertation introduces one such approximate accelerated stochastic simulation method. We have named it as the representative reaction approach (RRA). The crux of this method lies in representing the whole system of chemical reactions by a single representative reaction. The objective is to find the expected number of reactions and, further, a leap in time. This newly developed methodology has been successfully applied for the simulation of a number of chemical systems. The RRA method takes smaller steps in time relative to other approximate accelerated methods. Thus, it has found to be useful for the simulation of complicated oscillatory chemical reactions. The resulting algorithm of the RRA method is simple and easy to code. Like other contemporary methods, the RRA method also suffers from the drawback of negative numbers during simulations.

In this dissertation, attempts have been made to deal with the problem of occurrence of negative numbers by developing two new computational methods using the RRA method. In one simulation method, a binomial distribution based idea is applied to the RRA method. The other one with the concept of noise in the occurrence of reaction numbers. The excess amount of noise in the reaction numbers is responsible for driving the molecular numbers to a negative regime. The problem is solved by removing this noise

from the reaction numbers. It is found that these new methods are on par with the other methods developed for solving the problem of negative numbers. All these methods have the advantage of being simple and relatively straightforward to code.

All these approximate accelerated simulation methods rely on the use of an error control parameter. This parameter is instrumental for choosing the size of the time step. In all these simulations, the choice of the parameter is completely arbitrary. Hence, a parameter chosen for one particular chemical system may not be good enough for accurately predicting the dynamics of some other system. We have developed a method that provides flexibility for the choice of the error control parameter in a logical and mathematical manner. The proposed method successfully simulates the chemical systems, but also solves the problem of negative numbers during the simulations.

The work presented in this dissertation can be extended by applying the developed computational methods to some realistic biochemical reaction networks. It would be interesting to apply the proposed methods to the metabolic and signaling pathways and also to gene regulatory networks. These biochemical networks are in the form of models. The simulation data generated by our methods can be verified by testing with the experimental data.

The difficulties in solving the CME for complex chemical systems have made people go for the estimation of statistics by using the exact stochastic simulation algorithms. These algorithms are accurate, but their simulations are computationally time consuming. While the simulations performed with a large enough time step using the accelerated methods can be computationally efficient, there could be a significant bias in the results. The recenly proposed Multi-level method solves this problem. Hence, a future goal could be the development of new Multi-level Monte Carlo method with applications to gene regulatory networks. Finally, I want to conclude by quoting the following lines of Daniel T. Gillespie - 'I have come to believe that one's knowledge of any dynamical system is deficient unless one knows a valid way to numerically simulate that system on a computer. '

Bibliography

[1] Robert J. McMahon. Chemical reactions involving quantum tunneling. *Science*, 299(5608):833–834, 2003.

[2] A. M. Turing. The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 237(641):37–72, 1952.

[3] Shigeru Kondo and Takashi Miura. Reaction-diffusion model as a framework for understanding biological pattern formation. *Science*, 329(5999):1616–1620, 2010.

[4] J. D. Murray. *Mathematical Biology:I An Introduction Third Edition (Springer)*, 2008.

[5] Keith J. Laidler. *Chemical Kinetics, Third Edition (Pearson Education)*, 2008.

[6] Ira N. Levine. *Physical Chemistry, Sixth Edition (McGraw-Hill Education)*, 2008.

[7] Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[8] Alfred J. Lotka. Contribution to the theory of periodic reactions. *The Journal of Physical Chemistry*, 14(3):271–274, 1909.

[9] Alfred J. Lotka. Analytical note on certain rhythmic relations in organic systems. *Proceedings of the National Academy of Sciences of the United States of America*, 6(7):pp. 410–415, 1920.

[10] Alfred J. Lotka. *Elements of Mathematical Biology (Dover, New York)*, 1956.

[11] Narendra S. Goel, Samaresh C. Maitra, and Elliott W. Montroll. On the volterra and other nonlinear models of interacting populations. *Rev. Mod. Phys.*, 43:231–276, Apr 1971.

[12] R. N. Chapman. *Lecon sur la theorie mathematique de la lutte pour la vie (Gauthier-Villars,Paris)*, 1931.

[13] Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[14] David Edelson. Computer simulation in chemical kinetics. *Science*, 214(4524):981–986, 1981.

[15] Chetan Gadgil. Stochastic modeling of biological reactions. *Journal of the Indian Institute of Science*, 88(1):45–55, 2012.

[16] Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007.

[17] Daniel T. Gillespie, Andreas Hellander, and Linda R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17):170901, 2013.

[18] Uri M Ascher and Linda R Petzold. *Computer methods for ordinary differential equations and differential-algebraic equations*, volume 61. Siam, 1998.

[19] C William Gear and Linda R Petzold. Ode methods for the solution of differential/algebraic systems. *SIAM Journal on Numerical Analysis*, 21(4):716–728, 1984.

[20] Réka Albert and Hans G Othmer. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. *Journal of Theoretical Biology*, 223(1):1–18, 2003.

[21] Madalena Chaves, Réka Albert, and Eduardo D. Sontag. Robustness and fragility of boolean models for genetic regulatory networks. *Journal of Theoretical Biology*, 235(3):431–449, 2005.

[22] Kai Willadsen and Janet Wiles. Robustness and state-space structure of boolean gene regulatory models. *Journal of Theoretical Biology*, 249(4):749 – 765, 2007.

[23] Nicolaas Godfried van Kampen. *Stochastic Processes in Physics and Chemistry (North-Holland, 1992)*, 1981.

[24] C. W. Gardiner. *Handbook of Stochastic Methods: For Physics, Chemistry, and the Natural Sciences*, 1986.

[25] Daniel T. Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[26] Silvana Ilie. Variable time-stepping in the pathwise numerical solution of the chemical langevin equation. *The Journal of Chemical Physics*, 137(23), 2012.

[27] Silvana Ilie and Alexandra Teslya. An adaptive stepsize method for the chemical langevin equation. *The Journal of Chemical Physics*, 136(18), 2012.

[28] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[29] Daniel T. Gillespie and Linda R. Petzold. Improved leap-size selection for accelerated stochastic simulation. *The Journal of Chemical Physics*, 119(16):8229–8234, 2003.

[30] Ludwig F Wilhelmy. Ueber das gesetz, nach welchem die einwirkung der säuren auf den rohrzucker stattfindet. *Annalen der Physik und Chemie*, 81:413–433, 1850.

[31] Stanislas Leibler Naama Barkai. Biological rhythms: Circadian clocks limited by noise. *Nature*, 403:267–268, 2000.

[32] Johan Elf, Johan Paulsson, Otto G. Berg, and Måns Ehrenberg. Near-critical phenomena in intracellular metabolite pools. *Biophysical Journal*, 84(1):154 – 170, 2003.

[33] HarleyH. McAdams and Adam Arkin. Stochastic mechanisms in geneexpression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[34] Johan Paulsson, Otto G. Berg, and Måns Ehrenberg. Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences*, 97(13):7148–7153, 2000.

[35] Christopher V. Rao and Adam P. Wolf, Denise M.and Arkin. Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237, Nov 2002.

[36] Lucia Santoso, Hilary S. Booth, Conrad J. Burden, and Markus Hegland. A stochastic model of gene switches. In Rob May and A. J. Roberts, editors, *Proc. of 12th Computational Techniques and Applications Conference CTAC-2004*, volume 46, pages C530–C543, June 2005.

[37] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15):8614–8619, 2001.

[38] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322:549–560, 1905.

[39] M. von Smoluchowski. Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Annalen der Physik*, 326:756–780, 1906.

[40] S. Chandrasekhar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15:1–89, Jan 1943.

[41] Max Delbrück. Statistical fluctuations in autocatalytic reactions. *The Journal of Chemical Physics*, 8(1):120–124, 1940.

[42] H.A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284 – 304, 1940.

[43] AnthonyF. Bartholomay. On the linear birth and death processes of biology as markov chains. *The bulletin of mathematical biophysics*, 20(3):97–118, 1958.

[44] AnthonyF. Bartholomay. Stochastic models for chemical reactions: I. theory of the unimolecular reaction process. *The bulletin of mathematical biophysics*, 20(3):175–190, 1958.

[45] AnthonyF. Bartholomay. Enzymatic reaction rate theory: a stochastic approach. *Annals of the New York Academy of Sciences*, 96:897–912, 1962.

[46] AnthonyF. Bartholomay. Enzymatic reaction rate theory: a stochastic approach. *Biochemistry*, 1:223–230, 1962.

[47] Donald A. McQuarrie. Kinetics of small systems. i. *The Journal of Chemical Physics*, 38(2):433–436, 1963.

[48] D. A. McQuarrie, C. J. Jachimowski, and M. E. Russell. Kinetics of small systems. ii. *The Journal of Chemical Physics*, 40(10):2914–2921, 1964.

[49] Donald A. McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4(3):413–478, 1967.

[50] Shantanu Kadam and Kumar Vanka. A new approximate method for the stochastic simulation of chemical systems: The representative reaction approach. *Journal of computational chemistry*, 33(3):276–285, 2012.

[51] Shantanu Kadam and Kumar Vanka. Solving the problem of negative populations in approximate accelerated stochastic simulations using the representative reaction approach. *Journal of computational chemistry*, 34(5):394–404, 2013.

[52] Shantanu Kadam and Kumar Vanka. The accounting of noise to solve the problem of negative populations in approximate accelerated stochastic simulations. *RSC Advances*, 4(102):58127–58136, 2014.

[53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[54] A.B. Bortz, M.H. Kalos, and J.L. Lebowitz. A new algorithm for monte carlo simulation of ising spin systems. *Journal of Computational Physics*, 17(1):10–18, 1975.

[55] Arthur F. Voter. Introduction to the kinetic monte carlo method. *Radiation Effects in Solids (Springer Netherlands)*, 235:1–23, 2007.

[56] Richard Gordon. Adsorption isotherms of lattice gases by computer simulation. *The Journal of Chemical Physics*, 48(3):1408–1409, 1968.

[57] Farid F. Abraham and George M. White. Computer simulation of vapor deposition on twodimensional lattices. *Journal of Applied Physics*, 41(4):1841–1849, 1970.

[58] C.S. Kohli and M.B. Ives. Computer simulation of crystal dissolution morphology. *Journal of Crystal Growth*, 16(2):123–130, 1972.

[59] G.H. Gilmer. Growth on imperfect crystal faces: I. monte-carlo growth rates. *Journal of Crystal Growth*, 36(1):15 – 28, 1976.

[60] Per Arne Rikvold. Simulations of a stochastic model for cluster growth on a square lattice. *Phys. Rev. A*, 26:647–650, Jul 1982.

[61] Eric S. Hood, Brian H. Toby, and W. H. Weinberg. Precursor-mediated molecular chemisorption and thermal desorption: The interrelationships among energetics, kinetics, and adsorbate lattice structure. *Phys. Rev. Lett.*, 55:2437–2440, Nov 1985.

[62] Arthur F. Voter. Classically exact overlayer dynamics: Diffusion of rhodium clusters on rh(100). *Phys. Rev. B*, 34:6819–6829, Nov 1986.

[63] G. Barcaro, A. Fortunelli, F. Nita, and R. Ferrando. Diffusion of palladium clusters on magnesium oxide. *Phys. Rev. Lett.*, 95:246103, Dec 2005.

[64] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, 1968.

[65] Daniel T. Gillespie. A rigorous derivation of the chemical master equation. *Physica A: Statistical Mechanics and its Applications*, 188(1–3):404 – 425, 1992.

[66] Tetsuo Nakanishi. Stochastic analysis of an oscillating chemical reaction. *Journal of the Physical Society of Japan*, 32(5):1313–1322, 1972.

[67] M. Solc and I. Horsak. Simulation of first-order chemical reaction as a stochastic process on a digital computer. *Collect. Czech. Chem. Commun.*, 37(9):2994–3000, 1972.

[68] I. Horsak and M. Solc. Simulation study of a stochastic model of reversible first-order reaction equilibrium. *Collect. Czech. Chem. Commun.*, 38(8):2200–2203, 1973.

[69] Don L. Bunker, Bruce Garrett, Tadeusz Kleindienst, and George Stevenson Long III. Discrete simulation methods in combustion kinetics. *Combustion and Flame*, 23(3):373 – 379, 1974.

[70] I. Horsak and M. Solc. Simulation of stochastic models of higher-order reactions. *Collect. Czech. Chem. Commun.*, 40(2):321–325, 1975.

[71] Tetsuo Nakanishi. Time dependent fluctuation in a nonlinear chemical system. *Journal of the Physical Society of Japan*, 40(5):1232–1239, 1976.

[72] Saul A. Teukolsky William Press, Brian P. Flannery and William T. Vetterling. *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 1992.

[73] Michael A. Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The Journal of Physical Chemistry A*, 104(9):1876–1889, 2000.

[74] Yang Cao, Hong Li, and Linda Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The Journal of Chemical Physics*, 121(9):4059–4067, 2004.

[75] James M. McCollum, Gregory D. Peterson, Chris D. Cox, Michael L. Simpson, and Nagiza F. Samatova. The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Computational Biology and Chemistry*, 30(1):39 – 49, 2006.

[76] Hong Li and Linda Petzold. Logarithmic direct method for discrete stochastic simulation of chemically reacting systems. 2006.

[77] Larry Lok and Roger Brent. Automatic generation of cellular reaction networks with moleculizer 1.0. *Nature Biotechnology*, 23:131–136, 2005.

[78] Alexander Slepoy, Aidan P. Thompson, and Steven J. Plimpton. A constant-time kinetic monte carlo algorithm for simulation of large biochemical reaction networks. *The Journal of Chemical Physics*, 128(20):205101, 2008.

[79] David F Anderson. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of chemical physics*, 127(21):214107, 2007.

[80] Eric Mjolsness, David Orendorff, Philippe Chatelain, and Petros Koumoutsakos. An exact accelerated stochastic simulation algorithm. *The Journal of Chemical Physics*, 130(14):144110, 2009.

[81] Manuel Barrio, Kevin Burrage, Andrè Leier, and Tianhai Tian. Oscillatory regulation of hes1: Discrete stochastic delay modelling and simulation. *PLoS Comput Biol*, 2(9):e117, 09 2006.

[82] Christian A. Yates and Guido Klingbeil. Recycling random numbers in the stochastic simulation algorithm. *The Journal of Chemical Physics*, 138(9):094103, 2013.

[83] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of Chemical Physics*, 124(4):044109, 2006.

[84] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics*, 119(24):12784–12794, 2003.

[85] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1):014116, 2005.

[86] Yang Cao, Dan Gillespie, and Linda Petzold. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *Journal of Computational Physics*, 206(2):395 – 411, 2005.

[87] Martin Rein. The partialequilibrium approximation in reacting flows. *Physics of Fluids A: Fluid Dynamics (1989-1993)*, 4(5):873–886, 1992.

[88] John D. Ramshaw. Partial chemical equilibrium in fluid dynamics. *Physics of Fluids (1958-1988)*, 23(4):675–680, 1980.

[89] Christopher V. Rao and Adam P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm. *The Journal of Chemical Physics*, 118(11):4999–5010, 2003.

[90] Doraiswami Ramkrishna, Che-Chi Shu, and Vu Tran. New "tau-leap" strategy for accelerated stochastic simulation. *Industrial & Engineering Chemistry Research*, 53(49):18975–18981, 2014.

[91] Tianhai Tian and Kevin Burrage. Binomial leap methods for simulating stochastic chemical kinetics. *The Journal of Chemical Physics*, 121(21):10356–10364, 2004.

[92] Abhijit Chatterjee, Dionisios G. Vlachos, and Markos A. Katsoulakis. Binomial distribution based τ-leap accelerated stochastic simulation. *The Journal of Chemical Physics*, 122(2), 2005.

[93] Michel F. Pettigrew and Haluk Resat. Multinomial tau-leaping method for stochastic kinetic simulations. *The Journal of Chemical Physics*, 126(8), 2007.

[94] Xinjun Peng, Wen Zhou, and Yifei Wang. Efficient binomial leap method for simulating chemical kinetics. *The Journal of Chemical Physics*, 126(22), 2007.

[95] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Avoiding negative populations in explicit poisson tau-leaping. *The Journal of Chemical Physics*, 123(5):054104, 2005.

[96] Yuting Xu and Yueheng Lan. The n-leap method for stochastic simulation of coupled chemical reactions. *The Journal of Chemical Physics*, 137(20):204103, 2012.

[97] Xiaodong Cai and Zhouyi Xu. K-leap method for accelerating stochastic simulation of coupled chemical reactions. *The Journal of Chemical Physics*, 126(7):074102, 2007.

[98] Anne Auger, Philippe Chatelain, and Petros Koumoutsakos. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of Chemical Physics*, 125(8):084103, 2006.

[99] Zhen Liu, Yang Pu, Fei Li, Clifford A. Shaffer, Stefan Hoops, John J. Tyson, and Yang Cao. Hybrid modeling and simulation of stochastic effects on progression through the eukaryotic cell cycle. *The Journal of Chemical Physics*, 136(3), 2012.

[100] Thomas R. Kiehl, Robert M. Mattheyses, and Melvin K. Simmons. Hybrid simulation of cellular behavior. *Bioinformatics*, 20(3):316–322, 2004.

[101] Eric L. Haseltine and James B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of Chemical Physics*, 117(15):6959–6969, 2002.

[102] Andreas Hellander and Per Lötstedt. Hybrid method for the chemical master equation. *Journal of Computational Physics*, 227(1):100–122, 2007.

[103] Chetan Gadgil, ChangHyeong Lee, and HansG. Othmer. A stochastic analysis of first-order reaction networks. *Bulletin of Mathematical Biology*, 67(5):901–946, 2005.

[104] William J. Heuett and Hong Qian. Grand canonical markov model: A stochastic theory for open nonequilibrium biochemical networks. *The Journal of Chemical Physics*, 124(4), 2006.

[105] Ian J. Laurenzi. An analytical solution of the stochastic master equation for reversible bimolecular reaction kinetics. *The Journal of Chemical Physics*, 113(8):3315–3322, 2000.

[106] Chang Hyeong Lee, Kyeong-Hun Kim, and Pilwon Kim. A moment closure method for stochastic reaction networks. *The Journal of Chemical Physics*, 130(13), 2009.

[107] David Schnoerr, Guido Sanguinetti, and Ramon Grima. Validity conditions for moment closure approximations in stochastic chemical kinetics. *The Journal of chemical physics*, 141(8):084103, 2014.

[108] Ramon Grima. A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. *Journal of Chemical Physics*, 136:154105, 2012.

[109] ChangHyeong Lee, Jaemin Shin, and Junseok Kim. A numerical characteristic method for probability generating functions on stochastic first-order reaction networks. *Journal of Mathematical Chemistry*, 51(1):316–337, 2013.

[110] Pilwon Kim and Chang Hyeong Lee. A probability generating function method for stochastic reaction networks. *The Journal of Chemical Physics*, 136(23), 2012.

[111] Kevin R. Sanft, Sheng Wu, Min Roh, Jin Fu, Rone Kwei Lim, and Linda R. Petzold. Stochkit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458, 2011.

[112] Karan Vasudeva and Upinder S. Bhalla. Adaptive stochastic-deterministic chemical kinetic simulations. *Bioinformatics*, 20(1):78–84, 2004.

[113] David Adalsteinsson, David McMillen, and Timothy Elston. Biochemical network stochastic simulator (bionets): software for stochastic modeling of biochemical networks. *BMC Bioinformatics*, 5(1):24, 2004.

[114] Andrzej M. Kierzek. Stocks: Stochastic kinetic simulations of biochemical systems with gillespie algorithm. *Bioinformatics*, 18(3):470–481, 2002.

[115] S. Engblom, L. Ferm, A. Hellander, and P. Lötstedt. Simulation of stochastic reaction-diffusion processes on unstructured meshes. *SIAM Journal on Scientific Computing*, 31(3):1774–1797, 2009.

[116] T. C. Gard. Introduction to stochastic differential equations (marcel dekker: New york). 1988.

[117] Joseph L Doob. Topics in the theory of markoff chains. *Transactions of the American Mathematical Society*, 52:37–64, 1942.

[118] Joseph L Doob. Markoff chains - denumerable case. *Transactions of the American Mathematical Society*, 58:455–473, 1945.

[119] Didier Gonze, Jos Halloy, and Albert Goldbeter. Robustness of circadian rhythms with respect to molecular noise. *Proceedings of the National Academy of Sciences*, 99(2):673–678, 2002.

[120] Adam Arkin, John Ross, and Harley H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage -infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[121] Drew Endy and Roger Brent. Modelling cellular behaviour. *Nature*, 409(6818):391–395, Jan 2001.

[122] Jeff Hasty, David McMillen, and J. J. Collins. Engineered gene circuits. *Nature*, 420(6912):224–230, Nov 2002.

[123] Kevin Burrage and Tianhai Tian. Advances in scientific computings and applications (science press: Beijing/new york). 2004.

[124] Junpo He, Hongdong Zhang, Jingming Chen, and Yuliang Yang. Monte carlo simulation of kinetics and chain length distributions in living free-radical polymerization. *Macromolecules*, 30(25):8010–8018, 1997.

[125] Michel F. Pettigrew and Haluk Resat. Modeling signal transduction networks: A comparison of two stochastic kinetic simulation algorithms. *The Journal of Chemical Physics*, 123(11), 2005.

[126] Nina Fedoroff and Walter Fontana. Small numbers of big molecules. *Science*, 297(5584):1129–1131, 2002.

[127] Andr Leier, Tatiana T. Marquez-Lago, and Kevin Burrage. Generalized binomial τ-leap method for biochemical kinetics incorporating both delay and intrinsic noise. *The Journal of Chemical Physics*, 128(20):–, 2008.

[128] Tatiana T. Marquez-Lago and Kevin Burrage. Binomial τ-leap spatial stochastic simulation algorithm for applications in chemical kinetics. *The Journal of Chemical Physics*, 127(10):–, 2007.

[129] Jos M. G. Vilar, Hao Yuan Kueh, Naama Barkai, and Stanislas Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proceedings of the National Academy of Sciences*, 99(9):5988–5992, 2002.

[130] Howard Salis and Yiannis Kaznessis. Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *The Journal of Chemical Physics*, 122(5), 2005.

[131] Christian A. Yates and Kevin Burrage. Look before you leap: A confidence-based method for selecting species criticality while avoiding negative populations in τ-leaping. *The Journal of Chemical Physics*, 134(8), 2011.

[132] C. W. J. Beenakker and M. Büttiker. Suppression of shot noise in metallic diffusive conductors. *Phys. Rev. B*, 46:1889–1892, Jul 1992.

[133] Ya M Blanter and M Büttiker. Shot noise in mesoscopic conductors. *Physics Reports*, 336(1–2):1 – 166, 2000.

[134] D. G. Vlachos. Temporal coarse-graining of microscopic-lattice kinetic monte carlo simulations via τ leaping. *Phys. Rev. E*, 78:046713, Oct 2008.

[135] Tiejun Li. Analysis of explicit tau-leaping schemes for simulating chemically reacting systems. *Multiscale Modeling & Simulation*, 6(2):417–436, 2007.

[136] David F. Anderson, Arnab Ganguly, and Thomas G. Kurtz. Error analysis of tau-leap simulation methods. *The Annals of Applied Probability*, 21(6):pp. 2226–2262, 2011.

[137] Philip L. Boyland. Bifurcations of circle maps: Arnol'd tongues, bistability and rotation intervals. *Comm. Math. Phys.*, 106(3):353–381, 1986.

[138] Mitchell J. Feigenbaum, Leo P. Kadanoff, and Scott J. Shenker. Quasiperiodicity in dissipative systems: A renormalization group analysis. *Physica D: Nonlinear Phenomena*, 5(2–3):370 – 386, 1982.