ELSEVIER

Progress in Biophysics & Molecular Biology 85 (2004) 217–234

Progress in Biophysics & Molecular Biology

www.elsevier.com/locate/pbiomolbio

A multi-scaled approach for simulating chemical reaction systems

Kevin Burrage*, Tianhai Tian, Pamela Burrage

Department of Mathematics, Advanced Computational Modelling Centre, University of Queensland, Brisbane QLD4072, Australia

Abstract

In this paper we give an overview of some very recent work, as well as presenting a new approach, on the stochastic simulation of multi-scaled systems involving chemical reactions. In many biological systems (such as genetic regulation and cellular dynamics) there is a mix between small numbers of key regulatory proteins, and medium and large numbers of molecules. In addition, it is important to be able to follow the trajectories of individual molecules by taking proper account of the randomness inherent in such a system. We describe different types of simulation techniques (including the stochastic simulation algorithm, Poisson Runge–Kutta methods and the balanced Euler method) for treating simulations in the three different reaction regimes: slow, medium and fast. We then review some recent techniques on the treatment of coupled slow and fast reactions for stochastic chemical kinetics and present a new approach which couples the three regimes mentioned above. We then apply this approach to a biologically inspired problem involving the expression and activity of LacZ and LacY proteins in *E. coli*, and conclude with a discussion on the significance of this work.

MSC: 60C35; 65C30; 80A30

Keywords: Stochastic simulation methods; Poisson Runge–Kutta methods; Chemical reaction systems; Multi-scaled approaches; Biological applications

1. Introduction

There is now considerable evidence from both theoretical and experimental perspectives of the role of noise in genetic regulation. Federoff and Fontana (2002) remark that "stochasticity is

*Corresponding author. Tel.: +61-7-3365-3487; fax: +61-7-3365-6136.

E-mail addresses: kb@maths.uq.edu.au (K. Burrage), tian@maths.uq.edu.au (T. Tian), pmb@maths.uq.edu.au (P. Burrage).

0079-6107/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pbiomolbio.2004.01.014

evident in all biological processes. The proliferation of both noise and noise reduction is a hallmark of organismal evolution.'' However, a natural question to ask is what is the nature of this stochasticity? Hume (2000) notes that ''transcription in higher eukaryotes occurs with a relatively low frequency in biologic time and is regulated in a probabilistic manner.'' The comment about ''low frequency'' is significant here and we will return to this later.

Gene expression within a cell is a complex process involving such factors as chromatin remodelling, transcription, the export of RNA and the translation of mRNA into proteins. Physiological activity and cell differentiaton within a mammalian cell is controlled by perhaps more than 10,000 protein coding genes and thousands of genes are expressed at very low copy numbers. This means that new gene profiling techniques such as microarrays may not be able to reliably detect these numbers. Thus there is a great need for good models and effective simulations to guide the experimentalist and to provide additional insights into the nature of genetic regulation, see, for example, Kepler and Elston (2001) and McAdams and Arkin (1999).

Sano et al. (2001) remark that ''initiation of gene transcription is a discrete process in which individual protein-coding genes in an off state can be stochastically switched on, resulting in sporadic pulses of mRNA production.'' This is the dichotomy that we must resolve—proteins are discrete objects, yet their effects are often modelled (as ordinary differential equations) in terms of concentrations.

This leads us to the modelling process of how to represent genetic regulation mathematically. There are many approaches. These include:

- Directed graphs in which the genes are vertices and the gene interactions are the edges;
- Bayesian networks in which the vertices correspond to random variables that describe an expression while the network defines a joint probability density function;
- Boolean networks in which a gene is either in an on or off state;
- Ordinary differential equations in which chemical kinetics rate equations are used to represent protein concentrations;
- Partial differential equations in which the spatial structure of cells are taken into account; and finally
- Stochastic differential equations in which we have to resolve the issue of whether we work with concentrations or with individual molecules.

As the previous discussion would suggest we can consider three different types of modelling regimes for understanding genetic regulation. These include the discrete and stochastic, the continuous and stochastic and the continuous and deterministic.

Essentially, the characterisations of these regimes depend on the nature of the reactions and the number of molecules in the system being studied. In this paper the focus will be on mixed systems with small numbers of key regulatory proteins and a mix of medium and large numbers of other types of molecules. The basis of our work is the stochastic simulation approach to biochemical reactions which was developed by Gillespie (1977) through the stochastic simulation algorithm (SSA). This is an essentially exact procedure for numerically simulating the time evolution of a well-stirred chemically reacting system by taking proper account of the randomness inherent in such a system. It is rigorously based on the same microphysical premise that underlies the chemical master equation (Gillespie, 1992b) and gives a more realistic representation of a system's evolution than the deterministic reaction rate equation (RRE). In particular, the RRE is entirely

inappropriate if the molecular population of some critical reactant species is so small that microscopic fluctuations can produce macroscopic effects. This is especially true for the genetic/enzymatic reactions in living cells. As with the chemical master equation, the SSA converges, in the limit of large numbers of reactants, to the same solution as the law of mass action.

Despite continued refinements to the numerical methods used in the SSA, it remains a computationally demanding approach limiting its applicability, especially for large reaction networks required for modelling most realistic gene networks. The algorithm takes time steps of variable length, based on the rate constants and population size of each chemical species. The probability of one reaction occurring relative to another is obtained by multiplying the rate constant of each reaction with the numbers of its substrate molecules. According to the correct probability distribution derived from the statistical thermodynamics theory, a random variable is then used to choose which reaction will occur, and another random variable determines how long the step will last. The chemical populations are altered according to the stoichiometry of the reaction and the process is repeated. The cost of this detailed SSA is the large amount of computing time. The key issue is that the time step for the next reaction can be very small indeed if we are to guarantee that only one reaction can take place in that time interval.

In recent years, the SSA has been successfully applied for simulating genetic/enzymatic reactions in which the molecular population of some critical reactant species is relatively small, for example, lambda phage, Arkin et al. (1998); and circadian rhythms, Elowitz and Leibler (2000), Gonze et al. (2002). It has also been applied to much larger systems than originally designed for. For example, Arkin et al. (1998) used the SSA to simulate a model of lambda phage containing 75 equations in 57 chemical species.

An alternative approach to the SSA is via the StochSim package developed initially by Morton-Firth (1998) as part of a study of bacterial chemotaxis. The aim was to develop a realistic way of representing the stochastic features of this signalling pathway and to handle the large numbers of individual reactions encountered (Firth and Bray, 2000). Molecules or molecular complexes are represented as individual software objects. Reactions between molecules occur stochastically, according to probabilities derived from known rate constants.

StochSim works by quantising time into a series of discrete, independent time intervals, the sizes of which are determined by the most rapid reaction in the system. At the start of the simulation, the user assigns the maximum number of molecules in the system. In each time interval, a molecule is selected at random and then another object (either a molecule or a pseudo-molecule) is again selected at random. If two molecules are selected, any reaction that occurs is bimolecular, whereas if one molecule and a pseudo-molecule are selected, it is unimolecular. Another random number is then generated to determine if a reaction will occur. The probability of a reaction is retrieved from a look-up table and if this exceeds the random number, the particles do not react. On the other hand, if the probability is less than the random number, the particles react, and the system is updated.

StochSim is likely to be slower than the Gillespie algorithm in calculating the eventual outcome of a small set of simple biochemical reactions, especially when the number of molecules is large. However, if the system contains molecules that can exist in multiple states, then StochSim may not only be faster but also closer to physical reality. StochSim has been extended to incorporate explicit spatial representation in which nearest-neighbour interactions of molecules (such as clustered receptors on a membrane) can be simulated. Three geometries in two spatial dimensions,

squares, triangles and hexagons, are supported and this work has been used in the prediction of the structural arrangement of the chemotaxis receptor complex (Shimizu et al., 2000).

One of the great challenges in the efficient simulations of chemical kinetic systems is how we deal with mixed systems in which some key species have low abundances (as is the case of some molecules in genetic regulation) while other molecules have large abundances and can be modelled via continuous SDEs. Thus a vital question to address is how we can link discrete and continuous models and simulation algorithms in a sensible and efficient manner when treating chemical kinetic systems?

Before we turn our attention to the main focus of this paper we feel that it is timely to make a remark about rate constants in chemical reactions. As Schnell and Maini (2003) remark, "many biochemists have devoted their attention to the art of accurately determining the kinetic parameters by employing the velocity expression rather than studying the conditions under which the velocity expression can be used. As a consequence, the velocity equations of the catalytic reaction have been employed on a number of occasions outside of the conditions for which they are valid."

Furthermore, experimental measurements are inherently inaccurate. Thus, Schnell and Maini (2003) argue for the use of numerically fitting procedures for calculating kinetic parameters from progress curves. In a subsequent paper, Schnell and Turner (2004) discuss how the conventional equations based on rate constants fail to describe the reactions in vivo conditions. When minimal obstructions to diffusion are present, the rate constant approach is reasonable but in the presence of significant obstructions to diffusion, $\log(k)$ decays linearly on a logarithmic time scale and so k is time-dependent. Thus Schnell and Turner present a modification to fractal-like kinetics for biochemical reactions occurring in crowded intracellular, non-homogeneous, environments. Spatial simulations of this approach give excellent agreement with lattice gas data.

Thus this paper is organised in the following manner. In Section 2, we give a brief overview of the SSA approach and discuss some new simulation techniques that have been developed to overcome the inherent limitations of the SSA. In Section 3, we consider different ways of treating the question we just raised including stochastic partitioning and our new approach for treating fast, medium and slow reactions. In Section 4, we will investigate the performance of these new ideas on a biologically significant problem and the paper will conclude in Section 5 with some general remarks and discussion for future work.

2. Stochastic simulation methods for chemical reaction systems

In this section we will give a brief introduction to stochastic simulation methods for chemical reaction systems. The first part of this section gives a brief overview of the three modelling regimes described in the introduction. In doing so, we will make some assumptions that while restrictive will allow us to make progress mathematically.

In particular, we will assume that we have a well-stirred mixture at constant temperature in a fixed volume Ω. This mixture consists of $N \geqslant 1$ molecular species $\{S_1, \ldots, S_N\}$ that chemically interact through $M \geqslant 1$ reaction channels $\{R_1, \ldots, R_M\}$. The restriction that Ω is fixed can be relaxed but we will not do that here.

The dynamic state of this system is denoted as $X(t) \equiv (X_1(t), \ldots, X_N(t))^\top$, where $X_i(t)$ is the number of S_i molecules in the system at time t. The initial state is given by $X(t_0) = X_0$. For each $j, j = 1, \ldots, M$, we will define the propensity function $a_j(X)$ such that $a_j(X(t))\,\mathrm{d}t$ is the probability that given $X(t) = X$, one reaction R_j will occur inside Ω in the next infinitesimal time interval $[t, t + \mathrm{d}t)$.

When that reaction occurs, $X(t)$ changes its state. The amount by which X_i changes is given by v_{ji}, which represents the change in the number of S_i molecules produced by one R_j reaction. The $N \times M$ matrix v with elements v_{ji} is called the stoichiometric matrix. In particular, if just the jth reaction occurs in the time interval $[t, t + \mathrm{d}t)$, the jth vector v_j of the stoichiometric matrix is used to update the state of the system by

$$X(t + \mathrm{d}t) = X(t) + v_j.$$

We see that the propensity functions and state-change vectors completely characterize the chemical reaction system.

In the discrete and stochastic case the $X_i(t)$ represent the number of S_i molecules at time t and thus $X(t)$ takes on integer values in a non-negative integer lattice of dimension N. In fact $X(t)$ is a discrete (jump) Markov process. As such it has a time evolution equation associated with it which describes the probability $P(x, t|x_0, t_0)$ that $X(t) = x$ given $X(t_0) = x_0$. This equation is called the chemical master equation (CME) and it can be written as

$$\frac{\partial}{\partial t} P(x, t|x_0, t_0) = \sum_{j=1}^{M} (a_j(x - v_j)P(x - v_j, t|x_0, t_0) - a_j(x)P(x, t|x_0, t_0)).$$

In general, this discrete parabolic partial differential equation is too difficult to solve (either analytically or numerically) and other techniques are needed to simulate the $X(t)$.

This leads to the so-called SSA of Gillespie (1977), which is an exact and direct representation of the evolution of $X(t)$. There are several forms of this algorithm. The direct method works in the following manner.

Method 1 (*The direct method*). With two independent samples r_1 and r_2 of the uniformly distributed random variable $\mathbf{U}(0, 1)$, the length of the time interval $[t, t + \mathrm{d}t)$ is determined by

$$\mathrm{d}t = \frac{1}{a_0(X)} \ln\left(\frac{1}{r_1}\right),$$

where $a_0(X(t))$ is the sum of all the propensity functions

$$a_0(X) = \sum_{k=1}^{M} a_k(X).$$

The determination of the specific reaction occurring in $[t, t + \mathrm{d}t)$ is given by the index j satisfying

$$\sum_{k=1}^{j-1} a_k(X) < r_2 a_0(X) \leqslant \sum_{k=1}^{j} a_k(X).$$

The update of the system is then given by

$$X(t + \mathrm{d}t) = X(t) + v_j.$$

The point about the SSA is that the time step τ is taken small enough to guarantee that only one reaction occurs in that time interval. Clearly the SSA can be very computationally inefficient especially when there are large numbers of molecules or the propensity functions are large.

Now if the system possesses a macroscopically infinitesimal time scale so that during any dt all of the reaction channels can fire many times, yet none of the propensity functions change appreciably, then the jump Markov process can be approximated by a continuous Markov process. This Markov process is described by the chemical langevin equation (CLE), which is a stochastic ordinary differential equation (SDE)—see Gillespie (1992a). It takes the Itô form

$$dX = \sum_{j=1}^{M} v_j a_j(X)\,dt + \sum_{j=1}^{M} v_j \sqrt{a_j(X)}\,dW_j(t), \tag{1}$$

where the $W_j(t)$ are independent Wiener processes.

The CLE represents processes in the intermediate regime, that is those processes that are stochastic and continuous. A Wiener process is a stochastic process satisfying

$$E(W(t)) = 0, \quad E(W(t)W(s)) = \min\{t, s\}.$$

It is known that the Wiener increments are independent Gaussian processes with mean 0 and variance $|t-s|$ (that is, $N(0,|t-s|)$). Thus the Wiener increment $\Delta W(t) \equiv W(t+\Delta t) - W(t)$ is a Gaussian random variable $N(0,\Delta t) = \sqrt{\Delta t}N(0,1)$.

The CLE is an example of the more general class of Itô stochastic differential equations given by

$$dy(t) = g_0(y(t))\,dt + \sum_{j=1}^{d} g_j(y(t))\,dW_j(t), \quad y(t_0) = y_0, \quad y \in \mathbb{R}^m. \tag{2}$$

Thus, general classes of methods that can be used to solve (2) can also be used to simulate solutions of (1), (see Kloeden and Platen 1992, for example).

Finally, the third regime occurs when the noise terms are negligible compared with the deterministic term. This leads to the standard chemical kinetic approach that is described by the reaction rate equations

$$X'(t) = \sum_{j=1}^{M} v_j a_j(X(t)).$$

Recently, considerable attention has been paid to reducing the computational time of simulation algorithms for stochastic chemical kinetics. Gibson and Bruck (2000) refined the first reaction SSA of Gillespie by reducing the number of random variables that need to be simulated. This can be effective for systems in which some reactions occur much more frequently than others. A different approach is adopted by Rao and Arkin (2003) who simulate systems that have been simplified by quasi-steady state assumptions. Resat et al. (2001) treat systems which have widely varying rate constants by applying a weighted Monte Carlo approach.

Gillespie (2001) proposed two new methods, namely the τ-leap method and the midpoint τ-leap method in order to improve the efficiency of the SSA while maintaining acceptable losses in accuracy. The key idea here is to take a larger time step and allow for more reactions to take place in that step, but under the proviso that the propensity functions do not change too much in that

interval. Thus in the time interval $[t, t+\tau)$ and with the present state $X(t)$ at time t, then the number of times that the reaction channel R_j will fire is a Poisson random variable

$$K_j(\tau; X, t) = P(a_j(X), \tau), \quad j = 1, \ldots, M.$$

Here, the notation $P(\lambda, t)$ denotes a stochastic Poisson process with mean λt and variance λt and where

$$\Pr(P(\lambda, t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}.$$

These considerations lead to the τ-leap method.

Method 2 (*The τ-leap method*). Choose a value for τ that satisfies the Leap Condition: i.e., a temporal leap by τ will result in a state change λ such that for every reaction channel $R_j, |a_j(X+\lambda) - a_j(X)|$ is "effectively infinitesimal." Generate for each $j = 1, \ldots, M$ a sample value k_j of the Poisson random variable $P(a_j(X), \tau)$, and compute $\lambda = \sum_{j=1}^{M} k_j v_j$. Finally, perform the updates by replacing t by $t+\tau$ and X by $X+\lambda$.

Since the τ-leap method uses the initial state X to approximate the states in the time interval $[t, t+\tau)$, its efficiency can be improved by computing a better approximation to the states in the given time interval—for example, by an estimation at the midpoint $t+\tau/2$. This leads to the midpoint τ-leap method.

Method 3(*The midpoint τ-leap method*). For the selected leaping time τ (which satisfies the Leap Condition), compute the expected state change $\bar{\lambda} = \tau/2 \sum_{j=1}^{M} a_j(X) v_j$ during the time period $[t, t+\tau/2)$. Then use the estimated state $X' \equiv X + [\bar{\lambda}]$ to generate for each $j = 1, \ldots, M$ a sample value k_j of the Poisson random variable $P(a_j(X'), \tau)$. Compute the actual state change, $\lambda = \sum_{j=1}^{M} k_j v_j$, and perform the updates by replacing t by $t+\tau$ and X by $X+\lambda$. Here [ ] denotes the integer part.

Burrage and Tian (2003) introduced the framework of Poission Runge–Kutta (PRK) methods for simulating chemical reaction systems. These PRK methods are related to the class of stochastic Runge–Kutta methods for solving stochastic differential equations driven by Wiener noise.

The reason for adopting this framework is as follows. A Poisson random variable $P(a_j(X), \tau)$ with a large mean $a_j(X)\tau$ can be approximated by a Gaussian random variable $N(a_j(X)\tau, a_j(X)\tau)$, since

$$P(a_j(X), \tau) \approx N(a_j(X)\tau, a_j(X)\tau) = a_j(X)\tau + \sqrt{a_j(X)\tau}\, N(0, 1),$$

where $N(\mu, \sigma^2)$ is a Gaussian random variable with mean μ and variance σ^2. This can be viewed as

$$P(a_j(X), \tau) \approx a_j(X)\tau + \sqrt{a_j(X)}\, \Delta W(t). \tag{3}$$

Now, the simplest numerical method for solving (2) is the Euler–Maruyama method. It takes the form

$$y_{n+1} = y_n + h g_0(y_n) + \sum_{j=1}^{d} \Delta W_j^{(n)} g_j(y_n), \quad t_{n+1} = t_n + h,$$

where $\Delta W_j^{(n)} \equiv W_j(t_n + h) - W_j(t_n)$ is a Gaussian random variable $N(0, h)$.

The Euler–Maruyama method converges with strong order 0.5 and weak order 1 to the Itô form of the SDE. If it is applied to (1) it takes the form

$$X_{n+1} = X_n + \tau \sum_{j=1}^{M} v_j a_j(X_n) + \sum_{j=1}^{M} \Delta W_j^{(n)} v_j \sqrt{a_j(X_n)}.$$

Now using the approximation in (3) we can write this as

$$X_{n+1} = X_n + \sum_{j=1}^{M} v_j P_j(a_j(X_n), \tau).$$

This method is nothing but the τ-leap method of Gillespie. Thus the τ-leap method is the Euler-Maruyama method applied in the discrete setting when there are small numbers of molecules.

This has led Burrage and Tian (2003) to consider a general class of explicit PRK methods in which s intermediate approximations are simulated within a given step. This class of method takes the form

$$Y_i = X_n + \sum_{k=1}^{M} v_k P_k(\sum_{j=1}^{s} W_{ij} a_k(Y_j), \tau), \quad i = 1, \ldots, s,$$

$$X_{n+1} = X_n + \sum_{k=1}^{M} v_k P_k(\sum_{j=1}^{s} \beta_j a_k(Y_j), \tau).$$

In general, it is sufficient to consider simulation methods in which s is 1 or 2, and this gives rise to a general class of two-stage methods of the form

$$Y = X_n + \sum_{k=1}^{M} v_k P_k(\theta a_k(X_n), \tau),$$

$$X_{n+1} = X_n + \sum_{k=1}^{M} v_k P_k((1-\beta) a_k(X_n) + \beta_k(Y), \tau).$$

This method can be viewed as the application of a two-stage Runge–Kutta method, which in tableau form is given by

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \theta & 0 & 0 \\ \hline & 1-\beta & \beta. \end{array} \tag{4}$$

Runge–Kutta methods represent a very important class of methods for solving ordinary differential equations (see Butcher, 1987). Note that if $\beta = 1/2\theta$, (4) is of order two when applied to ordinary differential equations of initial value type.

Burrage and Tian (2003) consider two new stochastic simulation methods with $\beta = 1/2\theta$: the Heun PRK method ($\theta = 1$) and the R2PRK method ($\theta = \frac{2}{3}$). The latter is so-called because it is directly related to the R2 method for solving Stratonovich SDEs (Burrage, 1999).

Rathinam et al. (2003) consider how stiffness manifests itself at both the continuous deterministic and discrete stochastic levels. In this case explicit methods become impractical. The authors construct two implicit versions of the explicit τ-leap method known as the rounded and

unrounded implicit τ-leap method which have better stability properties than the explicit τ-leap method and are suitable for solving stiff chemical systems. The unrounded method has the form

$$X_{n+1} = X_n + \tau \sum_{j=1}^{M} v_j(a_j(X_{n+1}) - a_j(X_n)) + \sum_{j=1}^{M} v_j P_j(a_j(X_n), \tau),$$

but suffers from the drawback that $X_{n+1} - X_n$ is typically not an integer vector. Rathinam et al. (2003) overcome this difficulty by a two-stage process which is similar to a prediction–correction process given by

$$X = X_n + \tau \sum_{j=1}^{M} v_j(a_j(X) - a_j(X_n)) + \sum_{j=1}^{M} v_j P_j(a_j(X_n), \tau),$$

$$X_{n+1} = X_n + \sum_{j=1}^{M} v_j[\tau(a_j(X) - a_j(X_n))] + \sum_{j=1}^{M} v_j P_j(a_j(X_n), \tau),$$

where again [ ] denotes the nearest nonnegative integer.

Burrage and Tian (2003) have proposed a general class of Poisson Runge–Kutta methods that allows for implicitness (as considered by Rathinam, for example) and takes the form in the unrounded setting

$$Y_i = X_n + \tau \sum_{k=1}^{M} v_k \left(\sum_{j=1}^{s} U_{ij} a_k(Y_j) \right) + \sum_{k=1}^{M} v_k P_k \left(\sum_{j=1}^{s} W_{ij} a_k(Y_j), \tau \right), \quad i = 1, \ldots, s,$$

$$X_{n+1} = X_n + \tau \sum_{k=1}^{M} v_k \left(\sum_{j=1}^{s} \alpha_j a_k(Y_j) \right) + \sum_{k=1}^{M} v_k P_k \left(\sum_{j=1}^{s} \beta_j a_k(Y_j), \tau \right),$$

and which is presented in tableau form as

	U	W
	α	β.

The rounded setting formulation is obvious and not given here.

We now present two particular methods in this tableau form

- the general class of two stage explicit PRK methods is

	0	0	0	0
	0	0	θ	0
	0	0	$1 - 1/2\theta$	$1/2\theta$

- while the implicit τ-leap method is

	0	0	0	0
	-1	1	1	0
	-1	1	1	0.

The astute reader will note that in these cases

$$\alpha e = 0, \quad \beta e = 1, \quad e = (1, \ldots, 1)^{\top}.$$

We do not have the space in this paper for a detailed analysis of the order conditions of PRK methods, but merely note that these are consistency conditions so that the mean and covariance matrix of the PRK methods have the same formal Taylor-like expansion as the SSA method for the $O(\tau)$ condition.

The basic idea that we will introduce in Section 3 is that chemical reaction systems can be viewed as consisting of three different regimes and can be solved by coupling together three different simulation approaches applied to each of these regimes. Thus we intend to use the SSA when there are only a very few molecules; the explicit PRK approach (as typified by the τ -leap method) will be used for components of the system with moderate numbers of molecules. Finally, we will use a simple SDE method for solving the CLE (1) when there are very large numbers of molecules. Since the CLE is just an example of the general class of Itô SDEs (2) we conclude this section with a brief discussion on suitable classes of stochastic methods for solving stiff SDEs. We note that stiffness within an SDE is characterised by the problem having widely varying Lyapunov exponents (these are the stochastic counterparts of eigenvalues).

For solving a stiff SDE of the form (2) there are three approaches: explicit, semi-implicit and fully-implicit methods. In the first case, explicit methods can be suitable for stiff problems only if the stepsize is not too small or if the additional computation associated with implicit methods is prohibitive. Perhaps the simplest method in the middle class is the semi-implicit Euler method which takes the form

$$y_{n+1} = y_n + hg_0(y_{n+1}) + \sum_{j=1}^{d} \Delta W_j^{(n)} g_j(y_n).$$

This method works well if (2) is stiff only in the deterministic component but less well if there is also stiffness in the stochastic components. Milstein et al. (1998) introduced the balanced Euler method to overcome this limitation; it takes the form

$$y_{n+1} = y_n + (I + C_n)^{-1} \left(hg_0(y_n) + \sum_{j=1}^{d} \Delta W_j^{(n)} g_j(y_n) \right).$$

The matrix C_n is chosen to be of the form

$$c_0(y_n)h + \sum_{j=1}^{d} c_j(y_n)|\Delta W_j^{(n)}|,$$

where the $c_j(y_n)$ are matrix functions chosen to give appropriate damping and guarantee existence of solutions. Note that the fully-implicit Euler method

$$y_{n+1} = y_n + hg_0(y_{n+1}) + \sum_{j=1}^{d} \Delta W_j^{(n)} g_j(y_{n+1})$$

cannot guarantee convergence at any particular time step since the Wiener increments can take on positive or negative values with equal probability and in any case does not converge to the Itô

solution if convergence does take place—see Burrage and Tian (2001), for example. Alcock and Burrage (2003) have considered improvements over the Balanced Euler method in terms of better order and stability properties while Tian and Burrage (2001) have constructed high order implicit Taylor methods for stiff SDEs. Both the semi-implicit Euler method and the Balanced Euler method have strong order 0.5 and weak order 1.

3. Multi-scaled approaches to chemical reaction systems

Recently, two new approaches by Rao and Arkin (2003) and Haseltine and Rawlings (2002) have been considered in an attempt to speed up the performance of the SSA. Both of these ideas are based on partitioning of the system. In the case of Rao and Arkin, they consider a time scale separation in which a subset of the system is asymptotically at steady state. This is called the quasi-steady-state assumption (QSSA) and eliminates the fast dynamics that is responsible for the poor performance of the SSA. If the QSSA is applied in deterministic kinetics, the ODEs describing the intermediate species are set to 0. In the stochastic setting the system is split into primary (y) and ephemeral (z) subsystems.

Let $P(y,z;t)$ be the probability density function of the entire system so that

$$P(y,z;t) = P(z|y;t)P(y;t).$$

Then Rao and Arkin assume that z conditional on y is Markovian, so that for fixed y the conditional probability distribution $P(z|y;t)$ approximately satisfies a master equation. If, in addition,

$$\frac{\mathrm{d}P(z|y;t)}{\mathrm{d}t} \approx 0$$

so that

$$P(z|y;t) \approx P(z|y),$$

then a CME for describing the evolution of the probability density function can be obtained solely in terms of the primary species y. The SSA can then be applied to this subsystem in a transparent manner. As a particular case Rao and Arkin (2003) show how a simple enzymatic reaction involving an enzyme, substrate and enzyme-substrate complex in which the substrate concentration is much larger than the enzyme concentration leads, via QSSA arguments, to applying the SSA with a propensity function of the form $a(s) = \alpha s/(\beta + s)$, which is of course the Michaelis–Menten approximation. Finally, Rao and Arkin (2003) consider, as a specific example, the behaviour of the P_R promoter in conjunction with the Cro protein in λ bacteriophage. The P_R promoter plays an important regulatory component for determining the lysis or lysogenic pathways in the lambda infection of *E. coli*; see, for example, Shea and Ackers (1985), Arkin et al. (1998), Tian and Burrage (2004).

Using the ideas of Rao and Arkin (2003), Haseltine and Rawlings (2002) attempt to speed up the performance of the SSA by partitioning a chemical reaction system into slow and fast reaction subsets. The slow subsystem corresponds to extents with small propensity functions and few numbers of reactions, while the latter corresponds to large propensity functions and large numbers of reactions. This partitioning is achieved by exploiting the structure of the CME and

deriving master equations that describe the evolution of the probability density function for both the slow and fast subsystems. The slow system is treated by the SSA, while the fast system is treated either deterministically or by applying the explicit Euler–Maruyama method to the CLE. Thus at each time point t_n the CLE is repeatedly solved until $t_{n+1} = t_n + \tau$ is reached and then the SSA is applied to the slow subsystem with a stepsize of τ.

Some remarks can be made about this approach.

- In order to move from the continuous to the discrete stochastic regime a rounding process must be adopted. This causes negligible errors as the values for the molecular species in the continuous regime are large.
- In the Haseltine and Rawlings approach it is not clear what the specific details for partitioning into slow and fast reactions are but they recommend maintaining at least two orders of magnitude difference between the partitioned reaction probabilities. However it is important for the partitioning to be adaptive and to change throughout the interval of integration.
- Haseltine and Rawlings use an explicit method, namely the Euler–Maruyama method, for simulating the CLE. However, since the propensity functions in the CLE are large, the SDE is stiff (in the sense of widely varying Lyapunov exponents) and thus some consideration could be given to using semi-implicit or fully-implicit methods for this component. This could come at some cost if the dimension of the fast subsystem is at least moderately large.

In spite of these remarks, the papers by Rao and Arkin (2003) and Haseltine and Rawlings (2002) represent a significant attempt for developing simulation techniques that interface between microscopic and macroscopic regimes.

We note here that when discussing slow, intermediate and fast sub-reactions we are in reality classifying reactions into slow, intermediate and fast regimes. These regimes are characterised by the presence of one or more slow, intermediate and fast reacting species. In some cases it is possible to scale systems such that each term in the governing equations is composed of an expression of order of magnitude unity, multiplied by a dimensionless parameter, and this can lead to semi-autonomous simplification procedures. However, for the complex systems that we will study, we will assume that these procedures are inappropriate. We emphasise that we are trying to get a completely general, adaptive, partitioning approach for simulating chemical reaction systems.

3.1. The implementation

As remarked previously, we intend to use the SSA, the τ-leap method, and the Euler-Maruyama method in the slow, intermediate and fast regimes, respectively. The main issue that we must first address is how to classify these regimes. In order to attempt this classification we need to analyse the τ-leap method in more detail.

Recall that the τ-leap method takes the form

$$X(t+\tau) = X(t) + \sum_{j=1}^{M} v_j P_j(a_j(X(t)), \tau). \tag{5}$$

Now if at some time point t, suppose $X(t)$ consists of only 0's or 1's. Then given that for a second order reaction of the form $[A]+[B]\rightarrow[C]$, the stoichiometric vector has entries 0, 1 or -1, it is clear that if $P(a(X(t)),\tau)\geqslant 2$ then we can obtain negative entries in $X(t+\tau)$. This is clearly unacceptable as the entries of $X(t)$ represent the numbers of molecules in the system.

Thus in the case of a very small number of molecules we have, for the τ-leap method, the constraint

$$a_j(X(t))\tau\leqslant 1,\quad j=1,\ldots,M.$$

Note that this constraint is very much like a stability constraint although it is not one.

This brief analysis shows that not only must we classify in terms of the size of the propensity functions but also in terms of the number of molecules in the system. Thus at every time step we will classify the system as slow, intermediate or fast (see Table 1). We then form three vectors corresponding to the slow, intermediate and moderate regimes and place in those vectors the corresponding reaction number. If there are no reactions in say the intermediate vector for a given time step then that means there are no intermediate reactions for that step and the simulation regime changes accordingly. Note that we do not make the somewhat arbitrary assumption of Haseltine and Rawlings of maintaining at least two orders of magnitude between the different regimes.

We now discuss our implementation. We will denote by $a^{(0)}_{\text{slow}}(X(t))$ and $a^{(0)}_{\text{mod}}(X(t))$, the sum over all the propensity functions of small and moderate sizes, respectively. We first note that in this paper we have not presented a theoretical basis for a classification into the three regimes: fast, intermediate and slow. We will do this in a later work. Instead, we use as our theoretical basis the work of Haseltine and Rawlings for a classification into fast and slow and then note that we will use two types of simulations in the latter regime namely SSA and τ-leap.

For the slow regime we first determine a stepsize for the slow reaction as

$$\tau_{\text{S}}:=\frac{1}{a^{(0)}_{\text{slow}}(X(t))}\ln\left(\frac{1}{r}\right),\quad r\sim U(0,1).$$

Then for given ε that allows us to control the relative changes in the propensity function and takes on a value typically between 0.01 and 0.1 (see Burrage and Tian, 2003 for more discussion on this), we determine the stepsize for the intermediate regime as

$$\tau_{\text{I}}:=\min_{j\in\text{Int}}\left\{\frac{a_j(X(t)+v)-a_j(X(t))}{a^{(0)}_{\text{mod}}(X(t))}\right\}<\varepsilon.$$

If $\tau_{\text{S}}<\tau_{\text{I}}$ we use τ_{S} as the stepsize for the τ-leap method in the intermediate regime, while if $\tau_{\text{I}}<\tau_{\text{S}}$ we integrate by the τ-leap method until we reach time point $t+\tau_{\text{S}}$. Note that we only update the intermediate reactions in this time interval.

Finally, we integrate the CLE by the explicit Euler–Maruyama method with a stepsize $\tau_{\text{F}}<\tau_{\text{S}}$, chosen appropriately to guarantee stability, until we reach the time point $t+\tau_{\text{S}}$. We should note in passing that when we apply the τ-leap repeatedly updating the intermediate steps from $[t,t+\tau_{\text{S}}]$ the number of molecules should not become negative in that step. By taking the mean of both sides in (5) we must have

$$X_I(t)\geqslant\tau_{\text{S}}a^{(0)}_{\text{mod}}(X_{\text{I}}(t)).$$

Table 1
Regime classification

# Molecules	Propensity function	Classification
Large	Large	Fast
Moderate	Large	Intermediate
Large	Moderate	Intermediate
Moderate	Moderate	Intermediate
Large	Small	Slow
Small	Large	Slow
Moderate	Small	Slow
Small	Moderate	Slow
Small	Small	Slow

This gives a mechanism for the classification of the slow and intermediate regimes. In all cases the classification and stepsize selection process is repeated from step to step.

4. Numerical simulations

The test problem that we will use for our multi-scaled simulation is one presented by Kierzek (2002). In this paper, Kierzek presents a quite sophisticated implementation of the SSA in a software package known as STOCKS. The implementation treats both the growing volume of a cell and the simulation of cell division. Since we will do the same with our implementation a brief discussion on how this is done is appropriate.

In a single generation it is assumed that the cell doubles its volume from 1 to 2. This is achieved by letting the volume grow as $V(t) = 1 + t/T$, where T is the cell generation time. Thus at each simulation time step, the rates of all the second order reactions are divided by the current volume. Secondly, when the system reaches the generation time, all of the reactants that model the DNA elements are doubled (implemented by a separate set of reactions from that being modelled). Then the numbers of all the molecules present in the system are divided by two, the volume of the cell is reset and the behaviour of a new cell is simulated for the next generation time.

The biological system that we will simulate is the expression and activity of LacZ and LacY proteins in *E. coli*. A detailed description of the biological significance of the model is given in Kierzek (2002) but we give the full list of reactions here in Table 2.

There are 22 reactions and 23 molecular species in this model. The initial state has PLac = 1, RNAP = 35, Ribosome = 350 and all other elements 0. Results from a simulation of the system on the interval $[0, 2000]$ are given in Figs. 1 and 2. In Fig. 1, we see that the steady-state numbers of lactose, lacZ and RbsLacY are approximately 30,000, 300 and 2; while from Fig. 2 we see that the steady-state values of the propensity functions for reactions 5, 17 and 20 are approximately 0.003, 0.1 and 1000, respectively. Clearly it makes a great deal of sense to classify this problem into three regimes both in terms of the numbers of molecules and the propensity functions. The

Table 2
A full list of reactions and rates

	Reaction	Rate constant
1	PLac + RNAP → PLacRNAP	0.17
2	PLacRNAP → PLac + RNAP	10
3	PLacRNAP → TrLacZ1	1
4	TrLacZ1 → RbsLacZ + PLac + TrLacZ2	1
5	TrLacZ2 → TrLacY1	0.015
6	TrLacY1 → RbsLacY + TrLacY2	1
7	TrLacY2 → RNAP	0.36
8	Ribosome + RbsLacZ → RbsRibosomeLacZ	0.17
9	Ribosome + RbsLacY → RbsRibosomeLacY	0.17
10	RbsRibosomeLacZ → Ribosome + RbsLacZ	0.45
11	RbsRibosomeLacY → Ribosome + RbsLacY	0.45
12	RbsRibosomeLacZ → TrRbsLacZ + RbsLacZ	0.4
13	RbsRibosomeLacY → TrRbsLacY + RbsLacY	0.4
14	TrRbsLacZ → LacZ	0.015
15	TrRbsLacY → LacY	0.036
16	LacZ → dgrLacZ	6.42E-5
17	LacY → dgrLacY	6.42E-5
18	RbsLacZ → dgrRbsLacZ	0.3
19	RbsLacY → dgrRbsLacY	0.3
20	LacZ + lactose → LacZlactose	9.52E-5
21	LacZlactose → product + LacZ	431
22	LacY → lactose + LacY	14

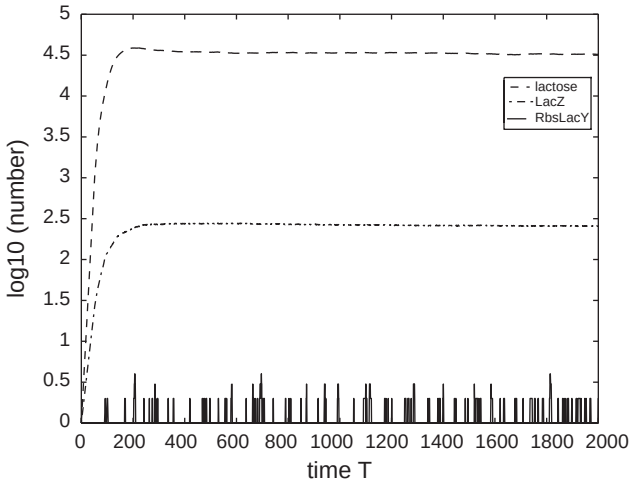

Fig. 1. Log of number of molecules vs. time.

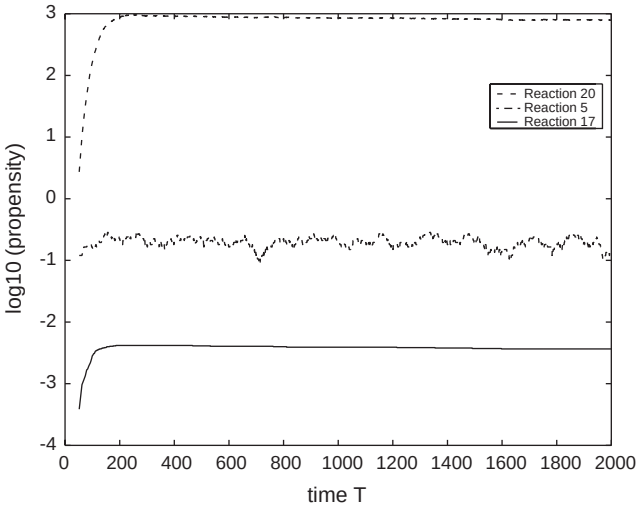

Fig. 2. Log of propensity functions vs. time.

Table 3
Comparison between SSA and Multi-scale approach

	SSA	Multi-scale(1)	Multi-scale(2)
Number of product molecules	1,596,036	1,623,072	1,723,546
Number of steps	4,833,348	3,294,700	3,497,500
Time taken	7 h 05 m 38 s	5 h 00 m29 s	5 h 21 m 10 s

particular classification that we use for this problem is

$$X_i \leqslant 100 \Rightarrow \text{small}, \qquad a_i(X) \leqslant 5 \Rightarrow \text{small},$$
$$X_i \in [101, 1000] \Rightarrow \text{moderate}, \qquad a_i(X) \in (5, 100] \Rightarrow \text{moderate},$$
$$X_i > 1000 \Rightarrow \text{large} \qquad a_i(X) > 100 \Rightarrow \text{large}.$$

The reason for this choice is that there is a trade-off between having a larger τ_S but having the τ-leap method giving negative numbers of molecules. In Table 3, we give some comparisons between SSA and the above classification. In this table, Multi-scale(1) and Multi-scale(2) refer to two different simulations of the multi-scale approach. The time taken was for MATLAB code executed on a Sun workstation, and is merely indicative of relative time taken; it is planned to develop this code in Fortran or C, and then to parallelise it to reduce the implementation time—see, for example, Burrage et al. (2003).

The improvements over the SSA implementation are substantial rather than dramatic. However, this work is just a first attempt to couple multi-scale simulations for a "real-life" challenging biological application. There are many opportunities for further work in the classification, the linking of the simulation techniques between the different regimes and reduction in the number of Poisson simulations, which we will consider in future work.

5. Concluding remarks

The dominating theme of the research described in this paper is the understanding of cellular dynamics in terms of interactions among the molecular components of a living cell. Of course we are a long way from this goal but new technologies such as the functional molecular cinematography unit offer a way of tracking the motion of individual molecules within a living cell. This offers a mechanism for the development and validation of more sophisticated models based on stochastic chemical reaction systems.

In the meantime, Endy and Brent (2001) have observed that researchers investigating the cell doubling of relatively simple organisms such as *E. coli* require a single simulation of 10^{14}–10^{16} reactions. In addition, in order to collate meaningful statistics, hundreds, if not thousands, of these simulations are needed. The main focus of this paper has been on the development of a multi-scaled approach via the linking of appropriate simulation algorithms operating at the slow, intermediate and fast reaction regimes. Irrespective of these algorithmic advances there is also a need to couple these approaches to sophisticated implementations using, for example, parallel and grid computing. A number of groups are working on this—see, for example, Kierzek (2002), Burrage et al. (2003) and McCollum et al. (2002). If this is then coupled with sophisticated three-dimensional visualisation techniques then we can really start to approach the holy grail of genomics, namely the ability to predict the dynamic effects on an organism of gene expression.

Acknowledgements

The first named author would like to thank the Australian Research Council for funding under the Federation Fellowship scheme. The authors would also like to thank the referees for their insightful comments.

References

Alcock, J., Burrage, K., 2003. A note on the balanced method, BIT, in press.

Arkin, A., Ross, J., McAdams, H.H., 1998. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. Genetics 149, 1633–1648.

Burrage, K., Tian, T., 2003. Poisson Runge–Kutta methods for chemical reaction systems. Proceedings of the Hong Kong Conference on Scientific Computing, in press.

Burrage, K., Tian, T., 2001. The composite Euler method for solving stiff stochastic differential equations. J. Comput. Appl. Math. 131, 407–426.

Burrage, K., Burrage, P., Jeffrey, S., Pickett, T., Sidje, R., Tian, T., 2003. A grid implementation of chemical kinetic simulation methods in genetic regulation, Proceedings of APAC03 Conference on Advanced Computing, Grid Applications and eResearch, ISBN 0-9579303-1-3, Gold Coast, Australia.

Burrage, P.M., 1999. Runge–Kutta methods for stochastic differential equations, Ph.D. thesis, University of Queensland, Brisbane, Australia.

Butcher, J.C., 1987. The Numerical Analysis of Ordinary Differential Equations. Wiley, UK.

Elowitz, M.B., Leibler, S., 2000. A synthetic oscillatory network of transcriptional regulators. Nature 403, 335–338.

Endy, D., Brent, R., 2001. Modelling cellular behaviour. Nature 409, 391–395.

Federoff, N., Fontana, W., 2002. Small numbers of big molecules. Science 297, 1129–1131.

Firth, C.J., Bray, D., 2000. Stochastic simulation of cell signaling pathways. In: Bower, J.M., Bolouri, H. (Eds.), Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge, MA, pp. 263–286.

Gibson, M.A., Bruck, J., 2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. J. Phys. Chem. A 104, 1876–1889.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81, 2340–2361.

Gillespie, D.T., 1992a. A rigorous derivation of the chemical master equation. Physica A 188, 404–425.

Gillespie, D.T., 1992b. Markov Processes: An Introduction for physical scientists. Academic Press, New York.

Gillespie, D.T., 2001. Approximate accelerated stochastic simulation of chemical reaction systems. J. Chem. Phys. 115, 1716–1733.

Gonze, D., Halloy, J., Goldbeter, A., 2002. Robustness of circadian rhythms with respect to molecular noise. Proc. Natl. Acad. Sci. 99, 673–678.

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

Hume, D.A., 2000. Probability in transcriptional regulation and its implications for leukocyte differentiation and inducible gene expression. Blood 96, 2323–2328.

Kepler, T.B., Elston, T.C., 2001. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. Biophys. J. 81, 3116–3136.

Kierzek, A.M., 2002. STOCKS: stochastic kinetic simulations of biochemical systems with Gillespie algorithm. Bioinformatics 18 (3), 470–481.

Kloeden, P.E., Platen, E., 1992. Numerical Solution of Stochastic Differential Equations. Springer, Berlin.

McAdams, H.H., Arkin, A., 1999. It's a noisy business! Genetic regulation at the nanomolar scale. Trends Genet. 15, 65–69.

McCollum, J.M., Cox, C.D., Simpson, M.L., Peterson, G.D., 2002. Accelerating Gene Regulatory Network Modeling Using Grid-Based Simulation. To appear in Simulation Transaction: The Society for Modelling and Simulation.

Milstein, G., Platen, E., Schurz, H., 1998. Balanced implicit methods for stiff stochastic systems. SIAM J. Numer. Anal. 35, 1010–1019.

Morton-Firth, C.J., 1998. Stochastic simulation of cell signalling pathways, Ph.D. Thesis, Cambridge.

Rao, C., Arkin, A., 2003. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. J. Chem. Phys. 118 (11), 4999–5010.

Rathinam, M., Petzold, L.R., Gillespie, D.T., 2003. Stiffness in stochastic chemically reacting systems: The implicit tau-leap method. J. Chem. Phys. 119 (24), 12784–12794.

Resat, V.H., Wiley, H.S., Dixon, D.A., 2001. Probability-weighted dynamic Monte Carlo method for reaction kinetics simulations. J. Phys. Chem. B 105, 11026–11034.

Sano, Y., Shimada, T., Nakashima, H., Nicholson, R.H., Eliason, J.F., Kocarek, T.A., Ko, M.S.H., 2001. Random monoallelic expression of three genes clustered within 60 Kb of Mouse *t* Complex Genomic DNA. Genome Res. 11, 1833–1845.

Schnell, S., Maini, P.K., 2003. A century of enzyme kinetics. Reliability of the K_M and $v_{\max}$ estimates. Comm. Theoret. Biol. 8, 169–187.

Schnell, S., Turner, T.E., 2004. Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. Prog. Biophys. Mol. Bio., this issue.

Shea, M.A., Ackers, G.K., 1985. The O_R control system of bacteriophage lambda: a physical-chemical model for gene regulation. J. Mol. Biol. 181, 211.

Shimizu, T.S., Le Novere, N., Levin, M.D., Beavil, A.J., Sutton, B.J., Bray, D., 2000. Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis. Nat. Cell. Biol. 2, 792–796.

Tian, T., Burrage, K., 2001. Implicit Taylor methods for stiff stochastic differential equations. Appl. Numer. Math. 38, 167–185.

Tian, T., Burrage, K., 2004. Bistability and switching in the lysis/lysogeny genetic regulatory network of bacteriophage λ. J. Theor. Biol. 227, 229–237.