

## Systems biology

## Simulated maximum likelihood method for estimating kinetic rates in gene expression

Tianhai Tian<sup>1,\*</sup>, Songlin Xu<sup>1,2</sup>, Junbin Gao<sup>3</sup> and Kevin Burrage<sup>1</sup><sup>1</sup>Advanced Computational Modelling Centre, University of Queensland, Brisbane, QLD 4072, Australia, <sup>2</sup>Department of Mathematics, Hubei University of Technology, Wuhan, Hubei 430068, P. R. China and <sup>3</sup>School of Information Technology, Charles Sturt University, Bathurst, NSW 2795, Australia

Received on June 23, 2006; revised on October 5, 2006; accepted on October 23, 2006

Advance Access publication October 26, 2006

Associate Editor: Jonathan Wren

## ABSTRACT

**Motivation:** Kinetic rate in gene expression is a key measurement of the stability of gene products and gives important information for the reconstruction of genetic regulatory networks. Recent developments in experimental technologies have made it possible to measure the numbers of transcripts and protein molecules in single cells. Although estimation methods based on deterministic models have been proposed aimed at evaluating kinetic rates from experimental observations, these methods cannot tackle noise in gene expression that may arise from discrete processes of gene expression, small numbers of mRNA transcript, fluctuations in the activity of transcriptional factors and variability in the experimental environment.

**Results:** In this paper, we develop effective methods for estimating kinetic rates in genetic regulatory networks. The simulated maximum likelihood method is used to evaluate parameters in stochastic models described by either stochastic differential equations or discrete biochemical reactions. Different types of non-parametric density functions are used to measure the transitional probability of experimental observations. For stochastic models described by biochemical reactions, we propose to use the simulated frequency distribution to evaluate the transitional density based on the discrete nature of stochastic simulations. The genetic optimization algorithm is used as an efficient tool to search for optimal reaction rates. Numerical results indicate that the proposed methods can give robust estimations of kinetic rates with good accuracy.

**Contact:** tian@maths.uq.edu.au

## 1 INTRODUCTION

Gene expression is the process by which a gene's DNA sequence is converted into the structures and functions of a cell. This process occurs in two important steps: transcription for making mRNA from protein encoding genes and translation for the biosynthesis of protein from mRNA. Gene expression processes have been increasingly studied in recent years from global perspectives in order to understand their pathways, properties and behaviours as a system. Progress in this research area will also be a key step towards the inference of genetic regulatory networks, which is one of the major challenges in systems biology during the postgenomic era (Akutsu, *et al.*, 2000; Crampin, *et al.*, 2004; Blais and Dynlacht, 2005; Joyce and Palsson, 2006). However, the bottleneck in the inference of

regulatory networks is the lack of synthesis and decay rates in gene expression that are very expensive to be determined by experiments (Yang *et al.*, 2003). In recent years, there have been significant advances in high-throughput technologies to monitor the various components of the mRNA and protein synthesis machineries. In addition, the combination of specific probes and advanced optical microscopy now allows observations of real-time production of single transcripts and protein molecules in individual cells (Golding *et al.*, 2005; Yu *et al.*, 2006). The availability of both massive 'omics' datasets and real-time molecular numbers has made it possible to study the function and stability of gene products and to reconstruct genetic regulatory networks at the genome scale (Joyce and Palsson, 2006).

It has been widely accepted that gene expression is a noisy business (McAdams and Arkin, 1999). Biological experiments and theoretical analysis have indicated that noise plays a very important role in gene expression, and different approaches have been proposed to investigate the impact of noise on the dynamics of regulatory networks (Arkin *et al.*, 1998; Rao *et al.*, 2002; Hasty *et al.*, 2000; Tian and Burrage, 2004a; Puchalka and Kierzek, 2004; Mao and Resat, 2004; Kaern, *et al.*, 2005; Tian and Burrage, 2006). Stochasticity in gene expression may result from small numbers of gene products, intermittent gene activity, and variability of transcriptional factor activities. With a limited number of promoter sites, the activation of gene expression is a discrete process, that switches randomly between the OFF states to the ON states. The copy numbers of mRNA are usually less than 10 per cell and these small numbers can lead to fluctuations in protein concentrations because of the unfrequent events in gene translation (Hasty *et al.*, 2002; Golding *et al.*, 2005). In addition to intrinsic noise, which is derived from uncertainty in biochemical reactions, extrinsic noise arising from environmental variability also has significant influence on the dynamics of the whole system.

A number of mathematical models have been proposed for studying gene expression processes and the stability of gene products (Hargrove and Schmidt, 1989; Yang *et al.*, 2003; Cao and Parker, 2003). Recently, Bhasi *et al.* (2005) have developed a method aimed at estimating the synthesis and degradation rates of gene products based on high-throughput 'omics' datasets. Based on deterministic models described by ordinary differential equations (ODEs), this method can be used to analyze gene expression data averaged from a population of cells. In fact this method is one of the approaches to estimating parameters in mathematical

\*To whom correspondence should be addressed.

models of biological pathways (Moles *et al.*, 2003; Gadkar *et al.*, 2005; Sugimoto *et al.*, 2005; Kell, 2006), which remains a challenging problem and a bottleneck in the development of mathematical models (Gadkar *et al.*, 2005). Furthermore, it is more challenging to evaluate kinetic rates in stochastic models that can generate different trajectories from the same parameters. Although the simulated maximum likelihood (SML) methods have been used to estimate parameters in stochastic differential equations (SDEs) for financial market models for which a large amount of information can be collected and very small time intervals can be used in parameter estimation (Hurn and Lindsay, 1999; Alcock and Burrage, 2004), the performance of these methods when they are applied to biological systems with sparse quantitative information, and especially when they are applied to systems described by discrete molecular numbers rather than continuous protein concentrations, is open to debate. Although an approach has been proposed most recently for estimating parameters in stochastic models of biological systems (Reinker *et al.*, 2006), this method is based on the analytic evaluation of transitional probabilities and thus it may not be appropriate to apply this method to biological systems in which the time intervals of the time series datasets are not small.

In this paper, we propose to use the SML method to estimate kinetic rates in gene expression processes that are described by either SDEs or discrete biochemical reactions. The joint transitional density is used to measure the fitness of stochastic simulations to gene expression profiles. For stochastic models with small numbers of molecular species, we propose to use the simulated frequency distribution to evaluate the transitional density based on the discrete nature of stochastic simulations.

## 2 METHODS

The start point of our discussion is the widely used deterministic model represented by ODEs

$$\frac{d\bar{x}_i}{dt} = f_i(\bar{x}) - g_i(\bar{x}), \quad i = 1, \dots, N, \quad (1)$$

where  $\bar{x} = (\bar{x}_1, \dots, \bar{x}_N)$  and the  $\bar{x}_i$  represent the concentrations of gene products and regulatory proteins in the regulatory network. Functions  $f_i(\bar{x})$  and  $g_i(\bar{x})$  represent the increase and decrease processes of molecule  $\bar{x}_i$  in gene expression. This deterministic model is valid if molecular numbers in the system are large. When molecular numbers of gene products are not large, stochastic models based on biochemical reactions have been used to describe the processes in gene expression (Thattai and van Oudenaarden, 2001; Swain *et al.*, 2002). Recently, a general modelling approach has been proposed for the development of stochastic models based on macroscopic reactions (Tian and Burrage, 2006). In this approach the increase and decrease processes in model (1) are replaced by Poisson random variables, given by

$$x_i(t + \tau) = x_i(t) + P[f_i(x)\tau] - P[g_i(x)\tau], \quad (2)$$

where  $x = (x_1, \dots, x_N)$  and the  $x_i$  are molecular numbers. If the processes in Equation (2) contain a number of macroscopic reactions, e.g.  $f_i(x) = f_{i1}(x) + \dots + f_{ik}(x)$  with  $f_{ij}(x) \geq 0$ , the Poisson random variable  $P[f_i(x)\tau]$  can be replaced by  $P[f_{i1}(x)\tau] + \dots + P[f_{ik}(x)\tau]$ .

If molecular numbers in the system are relatively large, stochastic models in the form of SDEs can be developed by means of the Langevin approach (Gillespie, 2001), given by

$$dx_i = [f_i(x) - g_i(x)]dt + \sqrt{f_i(x)}dW_{i1}(t) + \sqrt{g_i(x)}dW_{i2}(t), \quad (3)$$

where the  $W_{ij}(t)$  are the Wiener process. In this paper, SDEs (3) are simulated by the Euler method with a small stepsize. High order and implicit methods can also be used to improve the accuracy and stability properties of numerical simulations (Burrage *et al.*, 2004a).

It should be noticed that stochastic models can give better descriptions of gene expression than deterministic models if certain species in the system have small molecular numbers. Especially, the stochastic simulation algorithm (SSA) is a statistically exact method for simulating biochemical reaction systems (Gillespie, 1977). Compared with the SSA, the simulation time of SDE models is smaller but SDE models can give good approximation of the system dynamics only when molecular numbers in the system are relatively large. Furthermore, stochastic components in the SDE model (3) are negligible if all molecular numbers in the system are large, then the solution of deterministic model (1) gives the averaged behaviour of stochastic simulations with good accuracy. In this case, the advantage of deterministic models is the computational efficiency because a large computational time is usually required for stochastic simulation.

Parameter estimation in deterministic models can be achieved by the best fit of numerical simulations to experimental observations. Recently, Bhasi *et al.* (2005) have developed a program SPLINDID for estimating transcriptional rates and gene regulatory parameters. Based on the deterministic model with functional transcriptional rates described by spline functions, this method can estimate degradation rates with very good accuracy when the half-life of gene products is of the order of hours. However, it is not appropriate to use this program if the molecular numbers of gene products are not large. Instead, we should use methods based on stochastic models, such as the SML method (Hurn and Lindsay, 1999). Based on a sequence of  $N + 1$  observations  $\{x_0, x_1, \dots, x_N\}$  at time points  $\{t_0, t_1, \dots, t_N\}$ , we define the joint transitional density or likelihood function of these observations as

$$f_0[(t_0, x_0) | \theta] \prod_{i=1}^N f[(t_i, x_i) | (t_{i-1}, x_{i-1}), \dots, (t_0, x_0); \theta], \quad (4)$$

where  $\theta = (\theta_1, \dots, \theta_s)$  are the undetermined parameters in model (2) or (3),  $f_0[\cdot]$  is the density of the initial state, and  $f[(t_i, x_i) | (t_{i-1}, x_{i-1}), \dots, (t_0, x_0); \theta]$  is the transitional density starting from  $(t_{i-1}, x_{i-1})$  and evolving to  $(t_i, x_i)$ . When gene expression is described by the stochastic model (2) or (3), the stochastic process  $x$  is Markov (Gillespie, 2001), and the transitional density can be simplified as

$$f[(t_i, x_i) | (t_{i-1}, x_{i-1}), \dots, (t_0, x_0); \theta] = f[(t_i, x_i) | (t_{i-1}, x_{i-1}); \theta]. \quad (5)$$

An equivalent form of the maximum of the joint transitional density (4) is the minimum of the negative log-likelihood function, given by

$$L(\theta) = -\log(f_0[(t_0, x_0) | \theta]) - \sum_{i=1}^N \log f[(t_i, x_i) | (t_{i-1}, x_{i-1}); \theta]. \quad (6)$$

Because the closed-form expression of the transitional density (5) is usually unavailable, we use a non-parametric kernel density function

$$\bar{f}_M[(t, x) | (t_{i-1}, x_{i-1}); \theta] = \frac{1}{MB} \sum_{j=1}^M K\left(\frac{x - y_j}{B}\right) \quad (7)$$

to evaluate the transitional density based on the  $M$  realizations  $y_1, \dots, y_M$  of  $x_i$  at  $t_i$  given the initial condition  $(t_{i-1}, x_{i-1})$ . Here  $B$  is the kernel bandwidth and  $K(\cdot)$  is a non-negative kernel function enclosing unit probability mass. In the case of SDE models with a single variable, the normal kernel is widely used and the bandwidth can be chosen as  $B = 0.9 \sigma M^{-1/5}$ , where  $\sigma$  is the sample standard deviation of the  $M$  realizations (Hurn and Lindsay, 1999). For SDE models with multiple variables, we can either assume the independence of random variables or use the theory of multivariate density estimation (Scott, 1992). Note that the same increments of the Wiener process should be used in numerical simulations with different values of parameter  $\theta$ . Finally the optimal value of parameter  $\theta$  can be estimated by minimizing the log-likelihood function (6) over  $\theta$ . Thus we can derive the first SML method for estimating kinetic rates when gene expression is modelled by SDEs.

## Method 1

- (1) Input the system states  $\{x_0, x_1, \dots, x_N\}$  and time points  $\{t_0, t_1, \dots, t_N\}$ .
- (2) Take  $x_{i-1}$  at time  $t_{i-1}$  ( $i = 1, \dots, N$ ) as the starting value and use a numerical method to generate  $M$  realizations  $y_1, \dots, y_M$  of  $x$  at  $t_i$ . A random seed is specified for generating samples of the Gaussian random variables.
- (3) Use the non-parametric density (7) with the normal kernel or multivariate density functions to evaluate the transitional density (5).
- (4) Steps 2 and 3 are repeated for each time point  $t_0, \dots, t_{N-1}$ , and results are used to construct the log-likelihood function (6).
- (5) Search the optimal kinetic rates by a genetic optimization algorithm based on the minimum of  $L(\theta)$  in (6).

Although the normal kernel has been used successfully in estimating parameters in SDEs that have continuous solutions, it is not appropriate to use this kernel function to measure the transitional probability for systems described by discrete biochemical reactions. Consider the following example for the transcription of a single gene

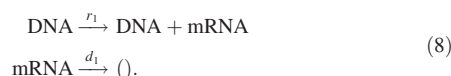


Figure 1 gives the evaluated density functions (7) with the normal kernel based on 1000 and 5000 realizations at  $t_1 = 3$ , respectively. Because of small molecular numbers, there is a significant difference between the values of density functions at integer points and those at other points in Figure 1A and B, and these values also depend on the numbers of realizations. Because integer molecular numbers are observed in discrete stochastic simulations, we are interested in the values of density functions only at integer points that do not satisfy the property of the unit probability mass, namely

$$\sum_{j=0}^{\infty} \tilde{f}_M[(t_1, j) | (t_0, x_0); \theta] > 1. \quad (9)$$

When molecular numbers are relatively large, the simulated density function in Figure 1C obtained from 1000 realizations is close to that derived from 5000 realizations in Figure 1D. However, we can still observe fluctuations in the density function values in Figure 1D.

Based on the discrete nature of biochemical reactions and low molecular numbers in gene transcription, we propose to use the frequency distribution of simulated molecular numbers to evaluate the transitional density (5). For systems with one single variable, after generating  $M$  realizations  $y_1, \dots, y_M$  of  $x_i$  at  $t_i$ , the frequency distribution is evaluated by

$$F[x(t_i) = m] = \frac{1}{M} \sum_{j=1}^M [1 - \delta(m - y_j)] \quad (10)$$

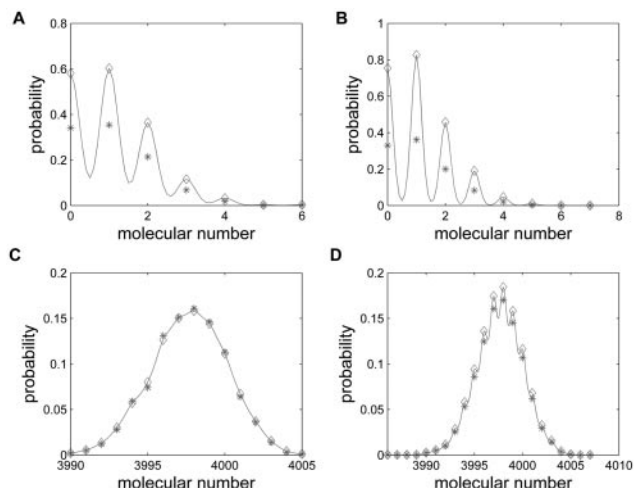
with  $m = 0, 1, \dots$ . Here

$$\delta(m - y_j) = \begin{cases} 0 & m = y_j \\ 1 & \text{else} \end{cases}. \quad (11)$$

This frequency distribution satisfies the property of the unit probability mass at integer points, namely

$$\sum_{m=0}^{\infty} F[x(t_i) = m] = 1. \quad (12)$$

For systems with species of both small and large molecular numbers, it may be difficult for simulations to match the experimental data with large molecular numbers in a finite number of simulations. Similar to the weighted distance measure in deterministic models (Moles, *et al.*, 2003; Sugimoto *et al.*, 2005), we can define the weighted frequency distribution in which the



**Fig. 1.** Values of the non-parametric density functions of system (8) calculated from the normal kernel (solid-line and diamond) and simulated frequency distribution (star). (A and B) function values are based on 1000 and 5000 simulations, respectively, with kinetic rates  $r_1 = 0.6$ ,  $k_1 = 0.3466$  and initial mRNA number  $x_0 = 0$  at  $t_0 = 0$ ; (C and D) function values are based on 1000 and 5000 simulations, respectively, with kinetic rates  $r_1 = 0.6$ ,  $k_1 = 3.466 \times 10^{-4}$  and initial mRNA number  $x_0 = 4000$  at  $t_0 = 0$ .

function  $\delta(x)$  is defined by

$$\delta_w(m - y_j) = \begin{cases} 0 & w_{ij} |m - y_j| < 1 \\ 1 & \text{else} \end{cases}, \quad (13)$$

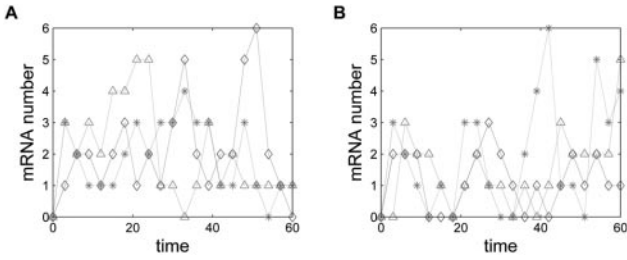
where  $w_{ij} = 1/(x_i \varepsilon)$ . This weighted frequency distribution is the frequency distribution (11) if  $w_{ij} = 1$ . In addition, these two types of frequency distributions are consistent for variables with small molecular numbers. For example, if  $\varepsilon = 0.05$ , the two frequency distributions (11) and (13) are the same if  $x_i < 20$ . Thus, the weighted frequency distribution is a good approximation of the transitional density for systems with both small and large molecular numbers.

Figure 1 also gives the frequency distributions of the mRNA number based on 1000 and 5000 realizations. Frequency distributions obtained by 1000 realizations in Figure 1A and C are very close to those from 5000 realizations in Figure 1B and D, respectively. The frequency distribution gives more stable estimations of the transitional density than the normal kernel density function. If molecular numbers are relatively large, frequency distributions in Figure 1C and D are close to the transitional density based on the normal kernel. Thus the frequency distribution gives better approximation of the transitional density (7) and we propose the second SML method for estimating kinetic rates in gene expression.

## Method 2

- (1) Input the system states  $\{x_0, x_1, \dots, x_N\}$  and time points  $\{t_0, t_1, \dots, t_N\}$ .
- (2) Take  $x_{i-1}$  at time  $t_{i-1}$  ( $i = 1, \dots, N$ ) as the starting value, and use a stochastic simulation method to generate  $M$  realizations  $y_1, \dots, y_M$  of  $x(t)$  at  $t_i$ . A random seed is specified for generating samples of the uniform random variable.
- (3) Use the frequency distribution (10) or (13) to estimate the transitional density (7).
- (4) Steps 2 and 3 are repeated for each time point  $t_0, \dots, t_{N-1}$ , and results are used to construct the log-likelihood function (6).
- (5) Search the optimal kinetic rates by a genetic optimisation algorithm based on the minimum of  $L(\theta)$  in (6).





**Fig. 2.** Generated mRNA numbers from the first test system (8) in 60 min. (A) Three sets of time-series data with less zero molecular numbers (diamond: set 1, star: set 2, triangle: set 3); (B) Three sets of time-series data with more zero molecular numbers (diamond: set 4, star: set 5, triangle: set 6).

Although the global search is a feasible approach for systems with a limited number of undetermined parameters, in general sophisticated searching methods should be used for estimating the optimal reaction rates. In this paper, a genetic algorithm is used as the search method that is especially helpful for finding kinetic rates when the search space is associated with a complex error landscape. We used a MATLAB toolbox developed by Chipperfield *et al.* (1994), and developed programs in C++ that are external programs of the MATLAB environment. For each set of time series data, the genetic algorithm was run over 300 generations, and we used a population of 100 individuals in each generation. The values of kinetic rates are taken initially to be uniformly distributed in the range  $[0, W_{\max}]$ , and the value of  $W_{\max}$  will be specified for each parameter based on the possible range of kinetic rates. The initial estimation of kinetic rates can be changed by using different random seeds in the genetic algorithm, and different initial rates will lead to slightly different final estimations. Similarly, different random seeds for generating samples in step 2 of Methods 1 and 2 will result in slightly different estimations because a fixed number of stochastic simulations are used in estimation.

### 3 RESULTS

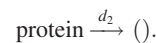
The first test system models the transcription of a single gene (8) with transcript initiation rate  $r_1 = 0.6 \text{ min}^{-1}$  and degradation rate  $d_1 = 0.3465$  (Thattai and van Oudenaarden, 2001). We used the SSA to generate six sets of mRNA numbers that were observed at every 3 min of 1 h. We present these gene expression profiles in two groups: the three sets with less zero molecular numbers (Fig. 2A) and the other three sets with more zero molecular numbers and more variability (Fig. 2B).

We used the genetic algorithm to search the optimal kinetic rates in the region  $(r_1, d_1) \in \{[0, 2] \times [0, 1]\}$ . The averaged relative error (ARE) is used to measure the accuracy of estimations. For Methods 1 and 2, we use 1000 simulations ( $M = 1000$ ) at each time point to evaluate the transitional density. Table 1 presents the estimated kinetic rates and corresponding ARE. Numerical results indicate that it is not appropriate to use the ODE model to estimate kinetic rates in biochemical reaction systems with small molecular numbers. The estimation of only one dataset has acceptable accuracy while others have large errors. Similarly, Method 1, the SML method based on the SDE model and normal kernel, cannot give robust estimations. Estimations have good accuracy for the first and sixth datasets, but have large errors for the other four datasets. On the contrary, Method 2, the SML method based on discrete biochemical reactions and frequency distribution, can give robust estimations with good accuracy. Among them, estimations for the first three datasets in Figure 2A have better accuracy than those from the last three datasets in Figure 2B. The reason may be

due to the larger fluctuations of mRNA numbers in the last three datasets. To demonstrate the effectiveness of frequency distribution, we use a modified Method 2 in which the normal kernel density function is used in Step 3 of Method 2. Simulation results, presented in Table 1 as Method 2 (normal kernel), indicate that the accuracy of this modified method is not as good as that of Method 2.

For this simple network with two parameters, we used the global search method to find the optimal kinetic rates in the region  $(r_1, d_1) \in \{[0, 2] \times [0, 1]\}$  with grid size  $h = 0.005$  (data not shown). For all methods in Table 1, estimations obtained from the global search method are consistent with those obtained from the genetic algorithm. In addition, we tested the influence of the number of realizations on the accuracy of estimations by evaluating kinetic rates based on 5000 realizations for the three methods based on stochastic models in Table 1 (data not shown). Although in most cases parameters estimated from 5000 simulations have better accuracy than those from 1000 simulations, the improvement in accuracy is not significant. This observation is consistent with the slow convergence property of the Monte-Carlo simulation methods. Thus 1000 simulations may be already large enough to achieve good accuracy, and a significant larger number of simulations should be needed if we hope to improve upon the accuracy of estimations derived from 1000 simulations.

The second test system describes the gene expression of a single gene with both transcription and translation



We use the simulated molecular numbers published in Figure 2 of Swain *et al.* (2002) as the gene expression profile. The lifetime of each cell cycle is 60 min and molecular numbers in 10 cell cycles were presented in Figure 2 of Swain *et al.* (2002). We estimated the numbers of mRNA transcript and protein at every 3 min from the gene expression profile of the first 4 cell cycles, which are reproduced in Figure 3.

We obtained 40 sets of estimations for each methods in Table 2 by using different random seeds in either the genetic algorithm or stochastic simulation. For the ODE model, we used 10 different random seeds in the genetic algorithm based on the gene expression profile in the 4 cell cycles. For Methods 1 and 2, we fixed the random seeds in stochastic simulation and used five different random seeds in the genetic algorithm to obtain 20 sets of estimations, and then fixed the random seed in the genetic algorithm and used five different random seeds in the stochastic simulations to obtain another 20 sets of estimations. The 40 sets of estimations obtained from Method 2 are presented in Figure 4.

We calculated the mean and standard deviation of the 40 sets of estimations obtained from these three methods. The AREs in the top part of Table 2 again indicate that the ODE model is not appropriate for estimating kinetic rates in systems with small molecular numbers. Estimation errors and standard deviations are large for each parameter. Compared with the first test system, Method 1 can give estimations with better accuracy for the second

**Table 1.** Estimated kinetic rates of the first test system (8). Results are presented as  $(k_1, d_1, \text{ARE})$ . Exact values are  $(k_1, d_1) = (0.6, 0.3466)$ 

Dataset	ODE model	Method 1	Method 2 (normal kernel)	Method 2
1	(0.3222, 0.1324, 0.54)	(0.6592, 0.3249, 0.08)	(0.3877, 0.2058, 0.38)	(0.7616, 0.3649, 0.16)
2	(1.9522, 1.0000, 2.06)	(1.8586, 1.0000, 1.99)	(0.0931, 0.0603, 0.83)	(0.5475, 0.2876, 0.12)
3	(2.0000, 0.9546, 2.04)	(1.9040, 0.7517, 1.67)	(0.2047, 0.1543, 0.60)	(0.5278, 0.2565, 0.19)
4	(1.2009, 1.0000, 1.44)	(0.1999, 0.1798, 0.57)	(0.1164, 0.1540, 0.68)	(0.4363, 0.3584, 0.15)
5	(2.0000, 1.0000, 2.10)	(1.2284, 0.6393, 0.95)	(0.6410, 0.6020, 0.40)	(0.8777, 0.4111, 0.32)
6	(1.5001, 1.0000, 1.69)	(0.5687, 0.3604, 0.05)	(0.1209, 0.0674, 0.80)	(0.3521, 0.2224, 0.38)

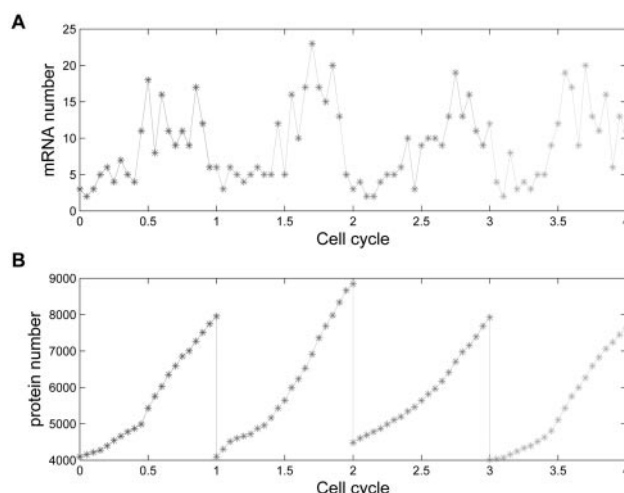
ARE represents averaged relative error.

system because of the relatively large protein numbers. Similar to the results of the first system, the estimation of Method 2 has better accuracy than that of Method 1.

We also tested the influence of the amount of data on the accuracy of estimations. For the second test system, we used the data in Figure 3 of the first cell cycle measured every 6, 12 or 30 min. Estimations obtained from Method 2 are presented in the bottom part of Table 2. If molecular numbers are measured every 6 min, there are molecular numbers at 11 time points and this amount of data can still ensure good accuracy of estimations although the accuracy is not as good as that of estimations from data measured every 3 min. However, if molecular numbers are measured every 12 min (6 time points) or every 30 min (only 3 time points), estimations in Table 2 have not only large AREs but also large standard derivations. Similar observations have also been found for the ODE model and Method 1.

The third test system is the genetic toggle switch interfaced with the SOS pathway (Gardner *et al.*, 2000; Kobayashi *et al.*, 2004). This network consists of two genes, *lacI* and  $\lambda$  *CI*, that encode the transcriptional regulator proteins LacR and  $\lambda$  CI, respectively. This system is regulated by a double-negative feedback loop and has two distinct bistable states. Transition between the two steady-states can be induced by a signal from the DNA damage that temporarily moves the system out of the bistable region. A deterministic model has been proposed for studying the existence of bistability properties (Gardner *et al.*, 2000; Kobayashi *et al.*, 2004), and a stochastic model based on the Poisson random variables has been used to realize the bimodal population distributions observed in experiments (Tian and Burrage, 2006). More detailed descriptions of this system and models can be found in (Gardner *et al.*, 2000; Kobayashi *et al.*, 2004; Tian and Burrage, 2006). Here we only present the corresponding model in terms of SDEs, given by

$$\begin{aligned}
 du &= \varepsilon \left( \alpha_1 + \frac{\beta_1 K_1^3}{K_1^3 + v^3} \right) dt + \sqrt{\varepsilon \alpha_1 + \frac{\varepsilon \beta_1 K_1^3}{K_1^3 + v^3}} dW_1(t) \\
 &\quad - d_1 [1 + \phi(s)] u dt + \sqrt{d_1 (1 + \phi(s))} u dW_2(t) \\
 dv &= \varepsilon \left( \alpha_2 + \frac{\beta_2 K_2^3}{K_2^3 + u^3} \right) dt + \sqrt{\varepsilon \alpha_2 + \frac{\varepsilon \beta_2 K_2^3}{K_2^3 + u^3}} dW_3(t) \\
 &\quad - d_2 v dt + \sqrt{d_2} v dW_4(t).
 \end{aligned} \quad (15)$$



**Fig. 3.** Molecular numbers of gene products in every 3 min estimated from the generated gene expression profile of the first 4 cell cycles in Figure 2 of Swain *et al.* (2002). (A) mRNA numbers; (B) protein numbers.

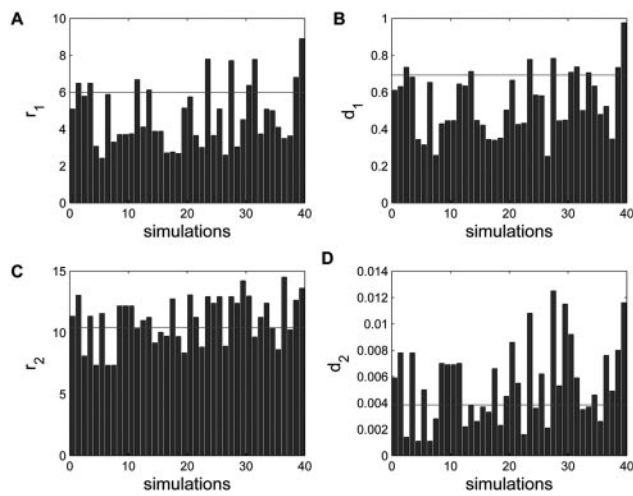
Instead of using  $\phi(s) = s/(1+s)$  which can lead to two different estimations, we use a linear function  $\phi(s) = s$  to represent the signal from DNA damage. Here  $\varepsilon = 1$  is a parameter associated with the number of the toggle switch plasmid (Kobayashi *et al.*, 2004), and constants in the Hill functions are  $K_1 = K_2 = 1 \mu M = 500$  molecule (Tian and Burrage, 2006). We are interested in the estimation of kinetic rates  $s$ ,  $\alpha_i$ ,  $\beta_i$  and  $d_i$  ( $i = 1, 2$ ). The unit of the rate constants  $\alpha_i$  and  $\beta_i$  in the stochastic model is molecule/min, and their values are obtained from  $\alpha_i = 500\alpha_i^{(0)}$  and  $\beta_i = 500\beta_i^{(0)}$ , where  $\alpha_i^{(0)}$  and  $\beta_i^{(0)}$  are parameters in the deterministic model (Tian and Burrage, 2006). In order to reduce the searching area, we use the genetic algorithm to find the optimal rates  $\alpha_i^{(0)}$  and  $\beta_i^{(0)}$ . Variables in all three types of models are molecular numbers in this paper.

Figure 5A gives a simulation of successful switching generated from the discrete stochastic model (Tian and Burrage, 2006) and molecular numbers are measured every 10 min. The signal from the DNA damage ( $s > 0$ ) is applied at  $t \in [10, 70]$  min. Then we estimated the values of the seven parameters based on deterministic model (Gardner *et al.*, 2000; Kobayashi *et al.*, 2004), Method 1 based on the SDE model (15), and Method 2 based on the discrete stochastic model (Tian and Burrage, 2006). Similar to the approach used for the second test system, we obtained 10 sets of estimations

**Table 2.** The mean, standard deviations and ARE of the estimations for the second test system

	$k_1$	$d_1$	$k_2$	$d_2$	ARE
ODE model	$3.442 \pm 1.852$	$0.093 \pm 0.144$	$3.383 \pm 1.995$	$7.61 \times 10^{-3} \pm 6.51 \times 10^{-3}$	0.736
Method 1	$3.822 \pm 0.926$	$0.434 \pm 0.088$	$10.671 \pm 1.932$	$5.12 \times 10^{-3} \pm 3.24 \times 10^{-3}$	0.273
Method 2	$4.735 \pm 1.695$	$0.542 \pm 0.168$	$11.056 \pm 1.963$	$5.47 \times 10^{-3} \pm 3.02 \times 10^{-3}$	0.228
Method 2 (6 min)	$4.595 \pm 2.115$	$0.538 \pm 0.285$	$10.564 \pm 2.647$	$6.20 \times 10^{-3} \pm 4.86 \times 10^{-3}$	0.271
Method 2 (15 min)	$3.205 \pm 1.284$	$0.345 \pm 1.33$	$11.597 \pm 2.774$	$7.82 \times 10^{-3} \pm 4.00 \times 10^{-3}$	0.528
Method 2 (30 min)	$0.784 \pm 0.137$	$0.698 \pm 0.317$	$7.820 \pm 5.331$	$0.175 \pm 0.160$	11.37
Exact	6	0.6931	10.3972	$3.852 \times 10^{-3}$	0

Upper part: estimations of the three methods based on the data measured every 3 min from the gene expression profile of the 4 cell cycles in Figure 3. Lower part: estimations of Method 2 based on the data measured every 6, 12 and 20 min, respectively, from the gene expression profile of the first cell cycle in Figure 3. ARE represents averaged relative error.

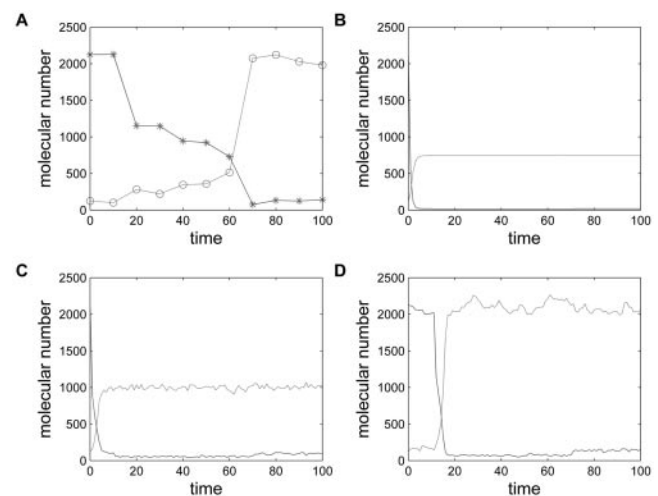


**Fig. 4.** Forty different estimations by using Method 2 for the kinetic rates of the second test system (14) based on molecular numbers in Figure 3. The exact values of these rates are  $(r_1, d_1, r_2, d_2) = (6, 0.6931, 10.3972, 3.852 \times 10^{-3})$  that are presented as a horizontal line in each figure. (A)  $r_1$ ; (B)  $d_1$ ; (C)  $r_2$ ; (D)  $d_2$ .

from each method and presented the mean, standard deviation and ARE of these estimations in Table 3. Compared with the first and second test systems, the ODE model can give relatively better estimations because molecular numbers in this system are relatively large. Estimations from stochastic models have better accuracy than that of the ODE model. The accuracy of Method 1 is even slightly better than that of Method 2. In addition, we simulated these three mathematical models by using the estimated parameters in Table 3. Simulations in Figure 5 indicate that only the estimation from Method 2 can keep the molecular numbers at the steady states of the original simulation in Figure 5A. Furthermore, simulations without a signal from DNA damage ( $s \equiv 0$ ) indicate that only the estimation from Method 2 can maintain the bistability property and genetic switching of the original system.

## 4 DISCUSSIONS

In this paper, we have developed the SML method for estimating kinetic rates in genetic regulation. Concentrating on gene expression processes with small molecular numbers, we used stochastic



**Fig. 5.** Simulations of the genetic toggle switch system. (A) A simulation of the discrete model (Tian and Burrage, 2006) with  $\alpha_1 = \alpha_2 = 0.2 \times 500$ ,  $\beta_1 = \beta_2 = 4 \times 500$ , and  $d_1 = d_2 = 1$ .  $s = 0.66$  ( $t \in [10, 70]$ ) and  $s = 0$  elsewhere. Data are measured at every 10 min. (B) A deterministic simulation of the ODE model by using the estimation of the ODE model in Table 3; (C) A stochastic simulation of model (15) by using the estimation of Method 1 in Table 3; (D) A stochastic simulation of the discrete model by using the estimation of Method 2 in Table 3. (Line:  $u$ ; dash-line:  $v$ )

models described by either SDEs or discrete biochemical reactions that give better descriptions of gene expression than deterministic models. Numerical results indicate that only the SML method based on discrete biochemical reactions can give robust estimations of kinetic rates with good accuracy when molecular numbers are small. If molecular numbers in the system are relatively large, the SML method based on either discrete biochemical reactions or SDEs can give robust estimations with good accuracy. Although we concentrated on estimating kinetic rates in genetic regulation, the proposed SML methods set up a general framework for estimating parameters in stochastic models of biochemical reaction systems and ecological systems where noise plays a very important role.

The transitional density function measures the transitional probability of the system states and is approximated by a non-parametric kernel function. When gene products are measured by molecular numbers, we have shown that the frequency distribution is a good

**Table 3.** The mean, standard deviations and ARE of the estimations for the third test system

	$\alpha_1$	$\beta_1$	$\alpha_2$	$\beta_2$	$d_1$	$d_2$	$s$	ARE
ODE model	0.028 ± 0.021	0.15 ± 0.15	0.34 ± 0.34	1.13 ± 0.98	1.60 ± 0.41	0.99 ± 0.76	0.92 ± 0.48	0.62
Method 1	0.11 ± 0.24	1.78 ± 3.37	0.25 ± 0.30	2.95 ± 1.49	1.59 ± 0.42	1.59 ± 0.55	0.83 ± 0.70	0.43
Method 2	0.34 ± 0.06	6.17 ± 1.28	0.31 ± 0.19	4.86 ± 1.47	1.54 ± 0.22	1.25 ± 0.37	0.94 ± 0.46	0.46
Exact	0.2	0.2	4	4	1	1	0.66	0

ARE represents averaged relative error.

non-parametric kernel. We may also use a specific discrete probability distribution to approximate the transitional density function, but the accuracy of estimations strongly depends on the assumption that the simulated molecular numbers follow this specific distribution. We have used the Poisson distribution to approximate the non-parametric density function. Although the estimated kinetic rates for the first system have good accuracy, simulated results for the second system have large errors because the simulated protein number does not follow a Poisson distribution (data not shown).

Another important issue in the application of the SML method is computational efficiency. When using the genetic algorithm to search for optimal kinetic rates, hundreds of generations should be simulated and there are tens to a hundred individuals in each generation. For each individual, a large number of trajectories are required to ensure the accuracy and robustness property of the SML method. Thus any slight improvement in the efficiency of each stochastic simulation means significant improvement on the efficiency of the SML method. In this approach the SSA and Poisson  $\tau$ -leap method have been used to simulate small-scale biochemical reaction systems. When the number of biochemical reactions or molecular numbers of certain species are relatively large, other simulation methods should be employed to accelerate stochastic simulations. Recently, there has been significant progress in the development of efficient methods for simulating biochemical reaction systems including the  $\tau$ -leap method (Gillespie, 2001; Tian and Burrage, 2004a; Chatterjee et al., 2005a; Cao et al., 2005) and the multi-scale simulation methods (Rao and Arkin, 2003; Haseltine and Rawlings, 2002; Burrage et al., 2004b; Puchalka and Kierzek, 2004; Salis and Kaznessis 2005; Weinan et al., 2005; Samant and Vlachos, 2005), and in the development of effective computer programs (Kierzek, 2002; Chatterjee et al., 2005b; Salis et al., 2006). More work is needed to develop sophisticated computer programs in order to increase the efficiency of the SML method for estimating parameters in complicated biochemical reaction systems.

Because of the possible local optima in the genetic algorithm and the finite number of simulations in the stochastic simulations, different estimations can be obtained by using different random seeds in either the genetic algorithm or stochastic simulation. It may be more reasonable to use the mean of these estimations obtained from different random seeds in the stochastic simulations rather than to use the mean of the estimations obtained from different random seeds in the genetic algorithm, although the latter is a normal approach in the machine learning algorithms. However, more information from biological systems, such as the bistability and genetic switching in the genetic toggle switch system, should be used as additional criteria to select an appropriate estimation.

Stochasticity in gene expression may result from small numbers of gene products, intermittent gene activity, fluctuations of the activity of transcriptional factors and environmental variability. In addition, recent studies in the gene expression of single cells indicate that gene expression should be represented by a number of ‘burst’ processes (Golding et al., 2005; Yu et al., 2006). More work is needed to develop stochastic models for regulatory networks by including both intrinsic and extrinsic noise, transcriptional regulation, ‘burst’ processes and time delay in transcription and translation. The next step should be based on the application of the SML method to more sophisticated stochastic models that can reflect the most recent progress in the study of gene expression.

## ACKNOWLEDGEMENTS

One of the authors (K.B.) would like to acknowledge support of the Australian Research Council for the award of a Federation Fellowship.

*Conflict of Interest:* none declared.

## REFERENCES

- Akutsu, T. et al. (2000) Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, **16**, 727–734.
- Alcock, J. and Burrage, K. (2004) A genetic estimation algorithm for parameters of stochastic ordinary differential equations. *Comput. Stat. Data An.*, **47**, 255–275.
- Arkin, A. et al. (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics*, **149**, 1633–1684.
- Blais, A. and Dynlacht, B.D. (2005) Constructing transcriptional regulatory networks. *Gene Dev.*, **19**, 1499–1511.
- Bhasi, K. et al. (2005) SPLINDID: a semi-parametric, model-based method for obtaining transcription rates and gene regulation parameters from genomic and proteomic expression profiles. *Bioinformatics*, **21**, 3873–3879.
- Burrage, K. et al. (2004a) Numerical methods for strong solutions of stochastic differential equations: an overview. *Proc. R. Soc. A-Math. Phys. Eng. Sci.*, **460**, 373–402.
- Burrage, K. et al. (2004b) A multi-scaled approach for simulating chemical reaction systems. *Prog. Biophys. Mol. Bio.*, **85**, 217–234.
- Cao, Y. et al. (2005) Avoiding negative populations in explicit Poisson tau-leaping. *J. Phys. Chem.*, **123**, 054104.
- Cao, D. and Parker, R. (2003) Computational Modeling and experimental analysis of nonsense-mediated delay in yeast. *Cell*, **113**, 533–545.
- Chatterjee, A. et al. (2005a) Binomial distribution based tau-leap accelerated stochastic simulation. *J. Phys. Chem.*, **122**, 024112.
- Chatterjee, A. et al. (2005b) Time accelerated Monte Carlo simulations of biological networks using the binomial tau-leap method. *Bioinformatics*, **21**, 2136–2137.
- Chipperfield, A. et al. (1994) A Genetic Algorithm Toolbox for MATLAB. *Proc. Int. Conf. Sys. Engineering*, 200–207.
- Crampin, E.J. et al. (2004) Mathematical and computational techniques to deduce complex biochemical reaction mechanisms. *Prog. Biophys. Mol. Bio.*, **86**, 77–112.
- Gadkar, K.G. et al. (2005) Iterative approach to model identification of biological system. *BMC Bioinformatics*, **6**, 155.



- Gardner,T.S. *et al.* (2000) Construction of a genetic toggle switch in *Escherichia coli*. *Nature*, **403**, 339–342.
- Gillespie,D.T. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.
- Gillespie,D.T. (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.*, **115**, 1716–1733.
- Golding,I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.
- Hargrove,J.L. and Schmidt,F.H. (1989) The role of mRNA and protein stability in gene expression. *FASEB. J.*, **3**, 2360–2370.
- Haseltine,E.L. and Rawlings,J.B. (2002) Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.*, **117**, 6959–6969.
- Hasty,J. *et al.* (2000) Noise-based switches and amplifiers for gene expression. *Proc. Natl. Acad. Sci. USA*, **97**, 2075–2080.
- Hasty,J. and Collins,J.J. (2002) Translating the noise. *Nat. Genet.*, **31**, 13–14.
- Hurn,A.S. and Lindsay,K.A. (1999) Estimating the parameters of stochastic differential equations. *Math. Comput. Simulat.*, **48**, 373–384.
- Joyce,A.R. and Palsson,B.Ø. (2006) The model organism as a system: integrating 'omics' data sets. *Nat. Rev. Mol. Cell Biol.*, **7**, 198–210.
- Kærn,M. *et al.* (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Rev. Genet.*, **6**, 451–464.
- Kobayashi,H. *et al.* (2004) Programmable cells: interfacing natural and engineered gene networks. *Proc. Natl. Acad. Sci. USA*, **101**, 8414–8419.
- Kell,D.B. (2006) Metabolomics, modelling and machine learning in systems biology—towards an understanding of the languages of cells. *FEBS J.*, **273**, 873–894.
- Kierzek,A.M. (2002) STOCKS: Stochastic kinetic simulations of biochemical systems with Gillespie algorithm. *Bioinformatics*, **18**, 470–481.
- Mao,L. and Resat,H. (2004) Probabilistic representation of gene regulatory networks. *Bioinformatics*, **20**, 2258–2269.
- McAdams,H.H. and Arkin,A. (1999) It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.*, **15**, 65–69.
- Moles,C.G. *et al.* (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.*, **13**, 2467–2474.
- Puchalka,J. and Kierzek,A.M. (2004) Bridging the gap between stochastic and deterministic regimes in the kinetic simulation of the biochemical reaction systems. *Biophys. J.*, **86**, 1357–1372.
- Rao,C.V. and Arkin,A.P. (2003) Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm. *J. Chem. Phys.*, **118**, 4999–5010.
- Rao,C.V. *et al.* (2002) Control, exploitation and tolerance of intracellular noise. *Nature*, **420**, 231–237.
- Reinker,S. *et al.* (2006) Parameter estimation in stochastic chemical reactions. *IEEE Proc. Sys. Biol.*, **153**, 168–178.
- Salis,H. and Kaznessis,Y. (2005) Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *J. Chem. Phys.*, **122**, 054103.
- Salis,H. *et al.* (2006) Multiscale Hy3S: hybrid stochastic simulation for supercomputers. *BMC Bioinformatics*, **7**, 93.
- Samant,A. and Vlachos,D.G. (2005) Overcoming stiffness in stochastic simulation stemming from partial equilibrium: a multiscale Monte Carlo algorithm. *J. Chem. Phys.*, **123**, 144114.
- Scott,D.W. (1992) Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, NY.
- Sugimoto,M. *et al.* (2005) Reverse engineering of biochemical equations from time-course data by means of genetic programming. *Biosystems*, **80**, 155–164.
- Swain,P.S. *et al.* (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. USA*, **99**, 12795–12800.
- Thattai,M. and van Oudenaarden,A. (2001) Intrinsic noise in gene regulatory networks. *Proc. Natl. Acad. Sci. USA*, **98**, 8614–8619.
- Tian,T. and Burrage,K. (2006) Stochastic models for regulatory networks of the genetic toggle switch. *Proc. Natl. Acad. Sci. USA*, **103**, 8372–8377.
- Tian,T. and Burrage,K. (2004a) Bistability and switching in the lysis/lysogeny genetic regulatory network of Bacteriophage lambda. *J. Theor. Biol.*, **227**, 229–237.
- Tian,T. and Burrage,K. (2004b) Binomial leap methods for simulating stochastic chemical kinetics. *J. Chem. Phys.*, **121**, 10356–10364.
- Weinan,E. *et al.* (2005) Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. *J. Chem. Phys.*, **123**, 194107.
- Yang,E. *et al.* (2003) Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.*, **13**, 1863–1872.
- Yu,J. *et al.* (2006) Probing gene expression in live cells, one protein molecule at a time. *Science*, **311**, 1600–1603.