BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 15 2007, pages 1969–1977
doi:10.1093/bioinformatics/btm278

Systems biology

Computational methods for diffusion-influenced biochemical reactions

Maciej Dobrzyński[1,*], Jordi Vidal Rodríguez[2], Jaap A. Kaandorp[2] and Joke G. Blom[1]

[1]CWI (Center for Mathematics and Computer Science), Kruislaan 413 and [2]Section Computational Science, Faculty of Science, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands

Received on January 31, 2007; revised on April 20, 2007; accepted on May 17, 2007

Advance Access publication May 30, 2007

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: We compare stochastic computational methods accounting for space and discrete nature of reactants in biochemical systems. Implementations based on Brownian dynamics (BD) and the reaction-diffusion master equation are applied to a simplified gene expression model and to a signal transduction pathway in *Escherichia coli.*

Results: In the regime where the number of molecules is small and reactions are diffusion-limited predicted fluctuations in the product number vary between the methods, while the average is the same. Computational approaches at the level of the reaction-diffusion master equation compute the same fluctuations as the reference result obtained from the particle-based method if the size of the sub-volumes is comparable to the diameter of reactants. Using numerical simulations of reversible binding of a pair of molecules we argue that the disagreement in predicted fluctuations is due to different modeling of inter-arrival times between reaction events. Simulations for a more complex biological study show that the different approaches lead to different results due to modeling issues. Finally, we present the physical assumptions behind the mesoscopic models for the reaction-diffusion systems.

Availability: Input files for the simulations and the source code of GMP can be found under the following address: http://www.cwi.nl/projects/sic/bioinformatics2007/

Contact: m.dobrzynski@cwi.nl

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

There are many examples of biochemical reactions where spatial effects play an important role. In case of gene expression, transcription of a gene involves an encounter of RNA polymerase and transcription factors with a specific place on a DNA strand. The inclusion of diffusive effects is also important in the description of signaling pathways where additional noise due to sub-cellular compartmentalization can cause the signal weakening (Bhalla, 2004). Especially if the reactions are fast, diffusion can be a limiting factor in these processes since the environment is crowded and the dimensions of a cell are large compared to the size of the molecules. Besides, the number of molecules involved can be low which is an additional source of stochasticity. The presence of the stochastic effects in biological systems has numerous consequences. One of them is the appearance of redundancy in regulatory pathways in order to obtain deterministic behavior (McAdams and Arkin, 1999). Fluctuations may also increase the phenotypic heterogeneity which in turn improves the organism's environmental adaptation (Kærn *et al*., 2005).

The need for discrete-spatial-stochastic computational methods is apparent when confronted with theoretical studies of biochemical networks. Spatial coupling between chemically reacting systems is known to stabilize the autocatalytic reaction kinetics (Marion *et al*., 2002). Numerical analyses of a population model (Shnerb *et al*., 2000), a 4-component autocatalytic loop (Togashi and Kaneko, 2001) or a simple reaction-diffusion system (Togashi and Kaneko, 2004, 2005) show the emergence of a new behavior induced by the discrete nature of reactants. A behavior that could not be captured by the continuum approaches, let alone methods without space. The models of calcium wave propagation (Stundzia and Lumsden, 1996) and intracellular Ca^{+2} oscillations (Zhdanov, 2002), the study of Soj protein relocation in *Bacillus subtilis* (Doubrovinski and Howard, 2005) or MinD/MinE protein oscillations in *Escherichia coli* (Fange and Elf, 2006) are another illustration where the stochastic effects due to space and discreteness need to be accounted for to explain the experimental results.

In this article, we focus on some computational approaches that have been published recently and applied to biological systems. Two Brownian dynamics-based methods, Green's Function Reaction Dynamics and Smoldyn, have been respectively used to study fluctuations in gene expression (van Zon and ten Wolde, 2005a; van Zon *et al*., 2006) and to model signal transduction in *E.coli* chemotaxis (Lipkow *et al*., 2005; Lipkow, 2006). MesoRD and Gillespie Multi-Particle, two implementations of the reaction-diffusion master equation, allowed to study spatio-temporal dynamics of the cellular processes (Elf and Ehrenberg, 2004; Fange and Elf, 2006; Rodríguez *et al*., 2006). Finally, the Stochastic Simulation Algorithm by Gillespie has been frequently applied to investigate the influence of noise on biochemical networks

[*]To whom correspondence should be addressed.

Table 1. Models, their main features and assumptions with respect to BD for modeling biochemical systems

Model	Space	Discrete	Extra assumptions
BD	Yes	Yes	–
RDME	Yes	Yes	WM locally
CME	No	Yes	WM
PDE	Yes	No	C, LMA
ODE	No	No	WM, C, LMA

Abbreviations: WM—well-mixed system, C—continuum hypothesis for reactants, LMA—law of mass action.

(Arkin *et al*., 1998; Kierzek *et al*., 2001; Krishna *et al*., 2005). A general overview of the main features of the methods can be found in Takahashi *et al*. (Takahashi *et al*., 2005). Here we compare the various assumptions in the different models (Table 1) and the computational results. For clarity, we choose a simplified model of gene regulation as a case study. Using this model we make a detailed comparison between the methods. In particular we are looking at fluctuations of the product protein in gene expression. For a more realistic and complex biological system we discuss the influence of the necessary modeling choices.

2 REGIMES AND MODELS IN BIOCHEMISTRY

Biological phenomena in a single living cell span over a wide range of spatial and temporal scales. Also the number of molecular species involved can vary significantly. Concentrations of agents in reactions involved in gene expression reach nanomoles, while molecules are highly abundant in metabolic pathways (Fig. 1).

Current silicon cell platforms can often make reliable predictions for metabolic networks based on ordinary differential equations (ODEs) using the assumption that concentrations are high and space is not important. Only the rates of the processes determine changes in concentration of the metabolites. When spatial effects come into play, and the correlation length (CL)[1] decreases, indicating that the volume can no longer be treated homogeneous, methods based on partial differential equations (PDEs) are an appropriate approach. PDEs describe the change of continuous concentrations in time and also in the spatial dimension. This can be a good model for biochemical networks where some of the biomolecules are bound to the membrane like in signaling pathways or in eukaryotic cells in general because molecules are localized in compartments such as nucleus, mitochondria, endoplasmic reticulum, etc. In all of these instances possibly significant concentration gradients appear, and a simulation may require spatial methods (Francke *et al*., 2003).

It is known that the process at the very origin of the whole cellular machinery, gene expression, gives rise to fluctuations in the concentration of the final protein products. One of the sources of stochasticity in gene regulation is the low number of DNA-binding proteins which have to find their specific target in order to initiate translation (Halford and Marko, 2004). A low copy number of regulators and the positioning of genes on the chromosomes result in decreasing frequency of gene activation, thus increasing the fluctuations of mRNA (Becskei *et al*., 2005).

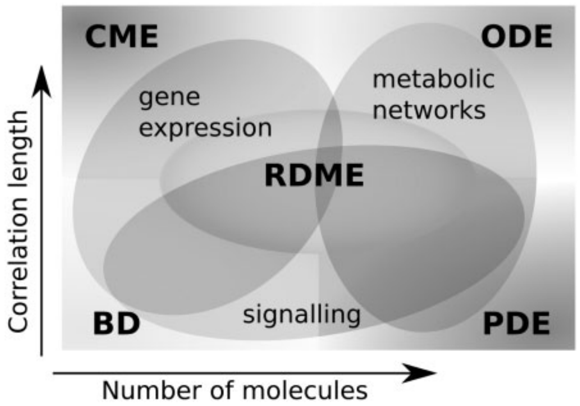

Fig. 1. Biological problems and relevant models placed in correlation-length versus number-of-molecules space. Abbreviations for (1) models with space: BD—Brownian dynamics, PDE—partial differential equation, RDME—reaction-diffusion master equation and (2) models without spatial detail: CME—chemical master equation, ODE—ordinary differential equation. ODE and PDE are deterministic models; CME, RDME and BD are stochastic. Colour version of this figure is available as Supplementary material online.

The discrete nature of matter as expressed in low-molecule-number conditions violates the continuum hypothesis used in ODEs and PDEs. A model accounting for this is based on the chemical master equation (CME), a deterministic linear ODE for the evolution of the probability density function for a Markov process (van Kampen, 1997). The Markov process models the stochastic transitions between discrete states of the system. In this case stochasticity reflects the fluctuation in the number of reactants' collisions, and hence the fluctuation in the number of molecules participating in a chemical reaction. The CME approach remains valid as long as the system is well mixed or, equivalently, has a large correlation length.

The question is whether this is a correct assumption when dealing e.g. with gene expression. Since there is a specific binding site which needs to be found by a relatively small number of competing transcription factors, diffusion might limit the process thus giving rise to larger fluctuations (Metzler, 2001). The probability of a reaction becomes inherently dependent on the distance from the target site. As a result the frequency of diffusion-limited binding events, for times smaller than the typical time needed to cross the volume, has a power-law distribution (Redner, 2001) instead of the exponential one used in mean-field approaches as CME. In order to resolve single diffusive encounters between biomolecules a more detailed approach such as Brownian dynamics (BD) is needed. In this approach, the solvent is treated as a continuum medium while solute molecules are modeled explicitly in space (Allen and Tildesley, 2002). Their trajectory is described by a

[1] The correlation length (CL) is a measure of the typical length scale at which a system retains its spatial homogeneity. Estimating CL can be a difficult task, since the different sub-processes involved can all have a different CL.

random walk due to collisions with the much smaller solvent molecules. Since the majority of degrees of freedom is characterized by the fluctuating force, the computational cost is much smaller than that of molecular dynamics (MD) where the positions and velocities of all atoms or groups of atoms are traced.

Unfortunately brute-force BD is too expensive for whole-cell simulations. Much more promising candidates for a versatile multi-scale framework are methods based on the reaction-diffusion master equation (RDME)—an extension of CME for spatially distributed systems. Space is incorporated by dividing the volume into smaller sub-volumes, which allows to tackle inhomogeneities due to diffusion (Gardiner, 1983). Tracking a single molecule is not possible in this model; unlike in BD, apart from the occupancy of the sub-volumes no exact positions of molecules are stored. Diffusive effects are treated correctly with RDME if the size of a sub-volume is of the order of the correlation length. Small sub-volumes are important if we want to account for fluctuations not only due to the probabilistic nature of chemical reactions but also resulting from rare binding events in diffusion-limited sparse (i.e. low concentration) systems. Obviously such detailed simulations are computationally expensive. Faster computations with large sub-volumes will give only a crude estimation of higher moments, but also the average will not be correct if the sub-volumes' size is larger than the CL and if the reactions are non-linear.

3 TEST CASES

3.1 Gene expression

In order to study fluctuations due to low number of molecules and spatial effects van Zon and ten Wolde (2005b) used a very simplified model to focus on the first step of gene regulation, reversible binding of polymerase to the operator site. Only this step is modeled explicitly in space.

The system under consideration is a closed volume V with a DNA binding site fixed in the center surrounded by molecules A diffusing freely with diffusion coefficient D. Once the DNA·A complex is formed with association rate k_a it can either dissociate back to separate DNA and A (with rate k_d) or a protein P can be produced with a production rate k_{prod} with subsequent complex dissociation. In both cases dissociation of DNA·A results in two separate molecules, DNA and A, at contact. The protein further decays at rate k_{dec}. Obviously the single protein production step in this model encompasses both transcription and translation which, as a matter of fact, consist of many biochemical reactions. Protein degradation is also simplified and treated as a first-order reaction. Table 2 includes the chemical reactions in the model.

The assumption that after protein production, molecules are placed at contact is not fully correct if we treat A as a RNA polymerase like in the original study. In fact polymerase travels a certain distance along DNA and unbinds at a position further than the initial one. Hence, we would like to remark that the freely diffusing agent A could be a transcription factor or an activator, which are also reported to occur in small quantities, instead of the RNA polymerase. Protein P could be seen as mRNA. In fact the idea of this model is to demonstrate product fluctuations due to rare events where the frequency is diffusion limited.

Table 2. Reaction scheme and parameters associated with the gene expression model

Reaction			Rate
DNA+A	$\xrightarrow{k_a}$	DNA·A	$3{\cdot}10^9 M^{-1}\ s^{-1}$
DNA+A	$\xleftarrow{k_d}$	DNA·A	$21.5\ s^{-1}$
DNA·A	$\xrightarrow{k_{prod}}$	P+DNA+A	$89.55\ s^{-1}$
P	$\xrightarrow{k_{dec}}$	Ø	$0.04\ s^{-1}$

Initially there is one free DNA site fixed in the center and 18 molecules A (corresponds to a 30 nm concentration) diffusing with D=1 μm^{2}s^{-1}.

We also assume that the molecules perform a pure random walk where the mean square displacement of the molecules is linear with time. The diffusion coefficient and the reaction rates are taken constant. Therefore we do not consider anomalous diffusion due to molecular crowding or hydrodynamic effects.

3.2 Signal transduction

In our comparison, we include also a model for diffusion of phosphorylated CheY in the *E.coli* chemotaxis pathway as reported by Lipkow *et al.* (2005). The cell is modeled as a rectangular box of length 2.52 μm, and width and height 0.88 μm (see the detailed scheme of the geometry in the Supplementary Material). Chemotaxis receptors are positioned inside the cell at the anterior wall. They form an array of 35 by 36 CheA dimers, which amounts to a total size of the receptor of 510 by 520 nm. Four motors are placed on the long sidewalls of the cell at 500 nm distance from each other. Each motor consists of 34 FliM molecules positioned on the walls of a cube (empty inside) of 40 nm. The cytoplasm contains 8200 CheY signaling molecules (partially in phosphorylated form), and 1600 CheZ dimers, both diffusing freely.

The reaction network is schematically depicted in Figure 2. CheY monomers are phosphorylated at the receptors where the phosphotransfer from CheAp to CheY takes place. Active CheA dimers (CheA*) produced in this reaction are converted back to CheAp in an autophosphorylation reaction. Phosphorylated CheY (CheYp) diffuses in the cytoplasm and binds reversibly to the FliM motor protein. CheYp can be also dephosphorylated by CheZ scavengers diffusing in the cytoplasm or it can autodephosphorylate. Once dephosphorylated, CheY converts to CheYp in a relatively slow reaction or it diffuses back to the receptor to go through the CheA-mediated phosphorylation. Further, CheYp can diffuse and form again the FliM·CheYp complex.

4 METHODS

Here we describe shortly the main features of the algorithms used in the comparison. A more detailed explanation of the models and computational methods can be found in the Supplementary Material.

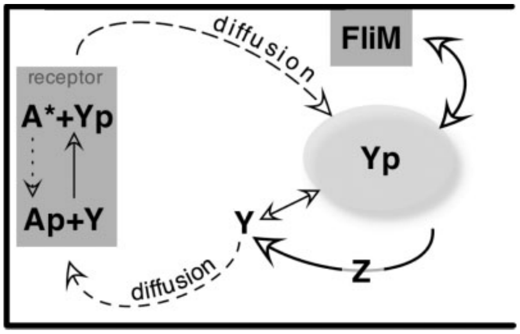

Fig. 2. Reaction scheme for the diffusion of CheY through the cytoplasm. Prefix *Che* is omitted in the names of the components. CheA* kinase dimers form an array of 1260 chemotaxis receptors inside the anterior cell wall. Dotted arrow in the receptor denotes an autophosphorylation of active dimers, CheA*→ CheAp. Solid arrow in the receptor is a phosphorylation of CheY. Only one flagellar motor (34 FliM proteins) is depicted here. A total of 8200 CheY signaling molecules (both non- and phosphorylated) and 1600 CheZ dimers diffuse freely in the cytoplasm. The biochemical network is described in the text, diffusion coefficients and reaction rate constants are listed in the Supplementary Material.

4.1 BD-level

Green's Function Reaction Dynamics (GFRD) developed by van Zon and ten Wolde (2005a, b) and Smoldyn (Smoluchowski Dynamics) by Andrews and Bray (2004), are two particle-based computational methods, which allow to explicitly model the gene expression problem described above. Reacting biomolecules are represented as spheres diffusing freely in a volume, and no excluded volume interactions are assumed.

GFRD uses the analytical solution of the Smoluchowski diffusion equation to resolve the reactive collisions. This allows to increase the simulation time step as compared to the traditional BD approach. This varying time step depending on the nearest neighbor distance is particularly efficient for systems with a low number of molecules. The method is not available as a general tool, and the code has been obtained from the authors.

Smoldyn, on the other hand is a convenient package. It is a more coarse-grained approach to BD simulations of biochemical reactions. The simulation time step is set by the user such that the probability of any reaction event per time step is small. Also the mean square displacement of diffusing molecules must allow for correct treatment of collisions. Here every collision leads to a reaction and the length of the binding and the unbinding radius (larger than the binding distance) for every reaction reproduces the macroscopic reaction rate.

4.2 RDME-level

MesoRD (Hattne *et al.*, 2005) simulates trajectories of discrete, stochastic systems with space described by the RDME. Gillespie Multi-Particle (GMP) (Rodríguez *et al.*, 2006) approximates this trajectory by splitting diffusion and reactions into two separate processes. In both cases the simulation volume is divided into sub-volumes, and the number of reactants inside them is recorded. Thus the knowledge of the position of the reacting molecules is limited by the resolution of the space discretization.

MesoRD employs the *next sub-volume method* (Elf and Ehrenberg, 2004) in order to identify the region of the domain where the next event triggers. The event can be either a transfer of particles between neighboring cells due to diffusion or a (bio)-chemical reaction.

GMP is based on the Lattice Gas Automata algorithm (Chopard *et al.*, 1994) for the diffusion process. The time step in GMP is fixed for every diffusing species and prescribed by the size of the lattice and respective diffusion coefficient. Reactions are executed in every sub-volume between the diffusion steps (different for every species) using Gillespie's SSA algorithm. Note that the fixed diffusion time step is in fact the average time between diffusion events in the RDME. This fact assures that the macroscopic diffusion in GMP is the same as obtained from MesoRD.

4.3 CME-based

The widely used stochastic simulation algorithm (SSA) developed by Gillespie (1976, 1977) generates realizations of the Markov process whose probability density function is described by the chemical master equation (Gillespie, 1992). During every step of the simulation two random numbers are drawn from appropriate distributions, and provide time and type of the next chemical reaction. Time assumes continuous values and the state of the system is a discrete number of all components present. Space is not included in the CME model.

5 RESULTS

5.1 Gene expression

The simplicity of the gene expression model discussed in Section 3.1 allows to illustrate how the implementations perform in the regime where spatial fluctuations are important. Additionally it is possible to address the issue of choosing the proper lattice size in RDME methods. The results of the RDME methods are compared to the solution obtained with CME and with the two BD-level simulations, where single molecules are modeled explicitly in space.

We analyze the average of the protein level and its noise η quantified as the ratio of standard deviation over average. Values of the parameters for the simulation are given in Table 2. Note that protein fluctuations depend on the frequency of the encounters of A and DNA. The influence of space is omitted in computational schemes based on CME such as the SSA algorithm by Gillespie. In that case the distribution of times between successive bindings of A to the promoter site on DNA is exponential because the process is assumed to be dependent on the reactants' concentration and not on their position. For spatially resolved methods the distribution of arrival times changes due to diffusion and results in burst-like behavior of the protein production. Therefore, we consider this problem, despite the great simplification of the gene expression, to be a good setting for the analysis of noise influenced also by spatial effects.

5.2 Dynamics of gene expression

The parameters we use in the simulations are such that the production process is limited by diffusion and that new proteins appear in bursts. We anticipate that stochastic effects will be significant under such circumstances. This idea is supported in Figure 3, where the protein level behavior in time is shown. Fluctuations are higher for methods explicitly accounting for space (GFRD and Smoldyn) comparing to the widely used SSA while the averages are the same. The reader will also notice that GFRD yields significantly larger fluctuations than Smoldyn—in principle a method at the same level of detail.

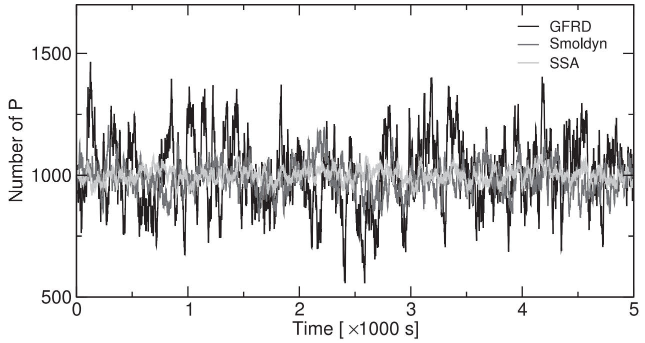

Fig. 3. Sample simulated time trajectories of the protein level for the gene expression problem. Note that the average is the same for all methods, while the fluctuations are higher for Smoldyn and GFRD, methods that include spatial effects. Parameters as in Table 2. Colour version of this figure is available as Supplementary material online.

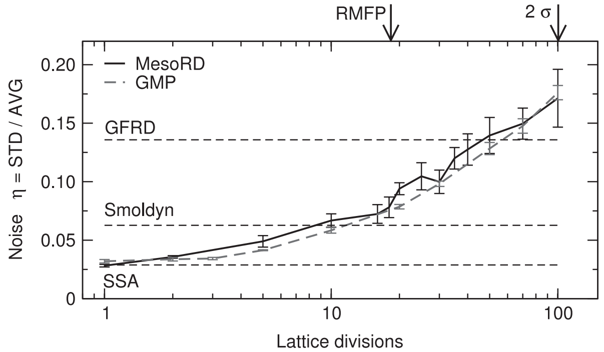

Fig. 4. Noise in the protein level as a function of n_{sv}—the number of sub-volumes per unit length, for RDME-level methods. The noise for SSA, Smoldyn and GFRD is computed from the experiment of Figure 3. RMFP—the reaction mean free path,[2] the estimate of the correlation length; 2σ—size of the sub-volume equal to two reaction distances which corresponds to two molecules fitting in one sub-volume. Noise is proportional to the square root of n_{sv}. Average protein level is the same for all experiments. Colour version of this figure is available as Supplementary material online.

The reason for the differences between the two BD methods is explained further in Section 5.3 and later in the Discussion Section. For now it is sufficient to say that GFRD produces more trustworthy results since it is an exact method to solve diffusion-limited reversible reactions.

The comparison for the RDME-class methods, MesoRD and Gillespie Multi-Particle, reveals the behavior for an increasing number of space divisions n_{sv} (the number of sub-volumes per unit length). We know that for n_{sv}=1, the well-mixed case, RDME methods are equivalent with the CME which does not include space. In Figure 4 we see that as the spatial detail is increased the predicted fluctuations reach the value given by GFRD, while the average number of proteins is the same, within statistical error, for all lattice sizes.

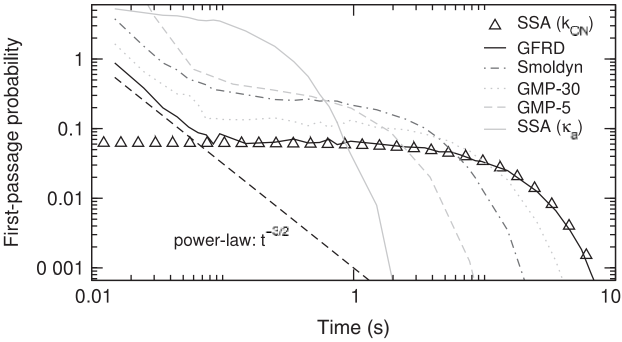

Fig. 5. Probability density function of time between subsequent bindings of an isolated pair of particles, also known as the first-passage probability. Methods including spatial effects, GFRD, Smoldyn and GMP, reproduce the power-law behavior for short times. For times larger than ~0.1 s, which is the average time needed to reach the boundary, the first-passage probability exhibits an exponential decay. GMP-5 and GMP-30 denote simulations where the whole volume's side is divided into 5 and 30 sub-volumes. The left- and rightmost curves are computed with SSA with the forward reaction rate equal to the *intrinsic* κ_a and to the *overall* k_{ON} reaction rate, respectively. The averages of all distributions are the same, equal $1/\kappa_a$, except for the SSA(k_{ON}), where the average is larger and amounts to $1/k_{ON}$. Note, that the position of the power-law line is chosen arbitrarily; it only compares the slope of data for short times obtained from different methods. Colour version of this figure is available as Supplementary material online.

The fact that MesoRD and GMP are able to reproduce not only the average but also the correct variance of the solution as compared to the reference result obtained from GFRD, shows their capability to give good results in the regime where spatial effects are important (the diffusion-limited regime with small numbers of molecules).

5.3 Reversible reaction of an isolated pair

In order to explain the differences in noise level predicted by various methods (Fig. 4) we look at a simple example of an isolated pair of molecules undergoing a reversible reaction, the same type of reaction as the first step in the gene expression model. This will allow us to examine the distribution of times between successive reactive events.

The target molecule is fixed in the center of the unit volume V, the second molecule is diffusing freely with diffusion coefficient D. The molecules can undergo a reversible reaction with association and dissociation rates, κ_a and κ_d, respectively. We look at the time between consecutive bindings of the molecules. Simulations with GFRD, Smoldyn, GMP and SSA reveal that the distribution of inter-binding times is different for methods with and without space, and also that methods with spatial detail treat the diffusion-limited reactions differently.

It is known (Redner, 2001) that for a particle diffusing in an infinite 3D space the probability that it reaches a specific target at a specified time (the first-passage probability) has a power-law distribution. This is depicted in the log–log plot in Figure 5. All spatial methods reproduce the power-law behavior (straight line) for times shorter than the average time needed to approach the boundary; a classical result for diffusion in an infinite 3D space. On such a short timescale, the molecule is not

[2] The reaction mean free path is defined as , where is the mean time between reactions, D_{rel} is the relative diffusion coefficient (Baras and Mansour, 1996; Rodríguez *et al.*, 2006; Togashi and Kaneko, 2005).

influenced by the finite boundary because it simply did not have time to travel that distance. On the other hand, diffusion in a finite volume results in an exponential decay of the first-passage probability for long times. Exponential behavior is characteristic to mean-field approaches like the chemical master equation, where the time of a next reaction is independent of the molecules' position (the well-mixed assumption). Note that, the exponential part of the result obtained with GFRD can be reproduced by changing the forward reaction rate in SSA from the intrinsic rate κ_a to the overall *on* rate coefficient k_{ON} (Agmon and Szabo, 1990; van Zon and ten Wolde, 2005a), which 'includes' the time needed to reach the target by diffusion and the time to undergo a chemical reaction.[3] It is important to note that the average of the first-passage distribution SSA (k_{ON}) in Figure 5 differs from the rest of the experiments. This fact indicates that the simple change of the reaction rates from the intrinsic to the *overall* rates which include the effects of diffusion will not preserve the average time between bindings. The plot also shows that an increase of spatial resolution of an RDME-class method like GMP results in the distribution which can be as close as desired to the exact solution given by GFRD. For clarity we draw only intermediate distributions with small (5) and average (30) number of sub-volumes per unit volume. The simulation with n_{sv}=50 overlaps with the result from GFRD.

5.4 CheY diffusion

We have simulated the chemotaxis pathway in *E.coli* (Lipkow *et al.* (2005)) using Smoldyn, MesoRD and GMP. GFRD is omitted in this case study because it is not suited for solving such a complex problem. A detailed description of the geometry and input files for the simulations can be found in the Supplementary Material.

Since Smoldyn allows for placing and tracking every molecule in the system, the implementation of the geometry of the receptor and motors is straightforward: CheA and FliM molecules can be fixed exactly at their positions. However, one needs to make an assumption about the placement of the motor and receptor molecules because Smoldyn does not account for excluded volume interactions and does not have any special treatment of reactions near walls. Following the choice made in Lipkow *et al.* (2005), we placed the receptor array of CheA dimers inside the cytoplasm, 20 nm from the anterior wall. The FliM motor proteins form a cuboid consisting of three layers at 16, 25 and 35 nm distance from the cell wall. Although molecules are modeled as points in Smoldyn, the macroscopic reaction rates prescribe the microscopic binding radii for every bimolecular reaction channel. For the given parameters and a simulation time step of 0.2 ms the binding distances are 16 nm for CheY phosphorylation at the receptor, 4 nm for CheZ-mediated dephosphorylation and 6 nm for binding to the FliM proteins. Smaller time step of 0.1 ms did not affect the results.

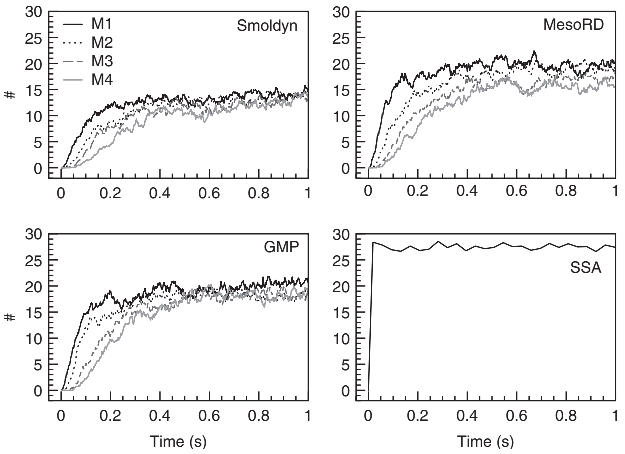

Fig. 6. Change in motor occupancy in time. Time step used in Smoldyn is 0.2 ms. The side of the sub-volume L_{sv} in MesoRD and GMP is 40 nm. Results do not change when L_{sv}= 20 or 80 nm is used. Colour version of this figure is available as Supplementary material online.

To reach the same level of spatial detail with RDME methods like MesoRD and GMP, a very fine discretization is required, since the exact positions of the molecules are known only up to the size of the sub-volume. We performed simulations where the side of the sub-volume L_{sv} equals 20, 40 and 80 nm. For L_{sv} =20 nm, the receptor consists of one boundary layer of 26 by 26 sub-volumes. In this case a cube constructed out of 8 sub-volumes approximates a motor. Four such cubes are positioned on the long sidewalls of the cell's surface. When bigger sub-volumes are used, motors occupy only one sub-volume. The receptor is a one-layer array of 13×13 and 7×7 sub-volumes for L_{sv}=40 and 80 nm, respectively.

5.5 Dynamics of CheY diffusion

In the computer simulations we first measure the time required to reach a given level of motor occupancy. Initially all CheA dimers in the receptor are in phosphorylated form, CheY and CheZ are freely diffusing in the cytoplasm. The results in Figure 6 show averages over 10 runs of 1 second for every method. The number of motor-bound CheYp is growing visibly slower for motors placed further along the cell. The time required to reach a threshold of 10 CheYp molecules bound to a FliM motor cluster predicted by MesoRD is the same within statistical error for all three sizes of the sub-volumes (Table 3). GMP produces slightly higher averages which can be attributed to the splitting error between reaction and diffusion. Molecules diffuse in 'bursts' due to the fixed diffusion time step which affects the first-passage properties of the diffusing front while the macroscopic mean square displacement is reproduced correctly. Finally, the average time to reach the given threshold is noticeably higher for Smoldyn. This difference cannot be explained by a wrong treatment in RDME of the non-linear reactions due to sub-volumes larger than the correlation length. Simulations for different sizes of the sub-volumes yielded the same results within statistical error. We contribute the discrepancy between Smoldyn and other methods to the

[3] The overall *on* reaction rate equals , where κ_a is the intrinsic association rate, and K_D is the diffusion-limited reaction rate given by $4\pi\sigma D$, dependent on the reaction distance σ and the relative diffusion coefficient D of two reacting molecules. Note that inverses of rates are equivalent to quantities with a dimension of time.

Table 3. Average time (in seconds) to reach motor occupancy of 10 CheYp molecules

Method	Comments	T_{M1}	T_{M2}	T_{M3}	T_{M4}
Smoldyn	$\Delta t = 0.2$ ms	0.11	0.19	0.22	0.29
MesoRD	$L_{sv} = 20$ nm	0.06	0.10	0.15	0.21
MesoRD	$L_{sv} = 40$ nm	0.06	0.11	0.15	0.22
MesoRD	$L_{sv} = 80$ nm	0.06	0.10	0.14	0.19
GMP	$L_{sv} = 40$ nm	0.06	0.08	0.17	0.23

Results are averaged overss 10 runs for every method.

Table 4. Average and noise in the level of motor occupancy in steady-state

Method	M1		M2		M3		M4	
	$\langle N \rangle$	η	$\langle N \rangle$	η	$\langle N \rangle$	η	$\langle N \rangle$	η
Smoldyn	13.9	0.20	13.2	0.21	12.5	0.23	12.4	0.23
MesoRD	19.1	0.17	18.0	0.18	16.5	0.20	16.0	0.22
GMP	20.3	0.15	19.0	0.16	18.6	0.16	18.1	0.16
SSA	27.5	0.08						

Averages were computed from simulations of length 21 s after the equilibration period of 1 s. Parameters of the simulations are the same as in Figure 6. MesoRD and GMP used $L_{sv} = 40$ nm.

difference in the modeling of receptor and motors. In MesoRD and GMP a reaction may occur when reactants are in the same sub-volume. For BD-based methods like Smoldyn, two reacting molecules need to be within the binding radius in order for an event to occur. This is a stricter constraint because it is possible that two molecules may simply pass each other despite being in a very close vicinity which would result in a reaction in RDME-level methods (unless the discretization is such that each sub-volume contains just one motor molecule).

Another property of the CheYp diffusion model we have studied is the average and the noise in the motor occupancy in steady-state (Table 4). For MesoRD and GMP we pick the simulations with 40 nm sub-volumes. Both RDME-level methods yield very similar results, although again averages from GMP are slightly higher than those obtained from MesoRD; Smoldyn computes ~20% lower averages. Note the interesting effect regarding noise in the motor occupancy, which increases for motors placed further from the receptor. This behavior of noise can be attributed to the concentration gradient of CheYp (high at the receptor and low at the posterior end). A smaller CheYp concentration at motor four compared to motor one results in a drop of the average motor occupation, and causes the fluctuations in binding to the FliM cluster to increase.

Additionally, we provide the SSA result for the motor occupancy (lower-right plot in Fig. 6 and Table 4) with the same reaction rates as in the other simulations. The occupancy for only one motor cluster is shown because all of them are equivalent if space is not included. This is clearly a wrong approach to model CheY diffusion, nevertheless it gives an indication of the error one can make when spatial information is omitted either by not accounting for geometry or by lack of correction in the diffusion-limited reaction rates. The average occupation which is higher than in the other methods is a direct consequence of a lack of delay due to the diffusion of CheYp towards the motors. Lower noise, on the other hand, results from the constant, and immediate supply of reactants while in case of simulations accounting for space, the supply is considerably lower due to the appearance of the CheYp concentration gradient.

6 DISCUSSION

The computational methods for modeling biochemical systems with single-particle and spatial detail compared in this study are based on BD and RDME models. In principle they are all suited for application to systems with a low molecule number and a short correlation length. The need for methods in this regime is steadily increasing as new results of experiments on biochemical reactions in single biological cells appear (Acar *et al.*, 2005; Golding *et al.*, 2005; Pedraza and van Oudenaarden, 2005; Rosenfeld *et al.*, 2005). A theoretical study of simple systems as the gene expression shows that fluctuations arise in diffusion-limited processes not only due to the small number of reactants but also resulting from spatial effects (van Zon *et al.*, 2006).

In the comparison for the simple gene expression test case we show that not all methods compute fluctuations correctly, although the average is the same as those given by the mean-field models (CME, ODE, PDE). Note that in general, if reactions are non-linear, one cannot expect RDME methods to give a correct estimate of the average when the sub-volumes are larger than the correlation length because concentration gradients are not represented within a sub-volume. Smoldyn yields much smaller fluctuations compared to GFRD, even though both methods are BD-based (Fig. 3).The reason for the incorrect prediction of the second moment by Smoldyn lies in the way it deals with diffusion-limited reversible reactions. The assumption that every collision is reactive leads to the introduction of the *unbinding radius* such that it reconstructs the macroscopic geminate recombination probability (more details on Smoldyn in the Supplementary Material). By doing so part of the spatial fluctuations is 'averaged' and the resulting first-passage probability lies between the exact solution of the Smoluchowski diffusion equation obtained with GFRD and the mean-field result from SSA for the well-mixed system (Fig. 5).

The methods at the level of the RDME, MesoRD and GMP, are able to predict correctly the fluctuations. The key issue here is to choose the space discretization (division into sub-volumes) such that the size of a single sub-volume is of the order of the system's correlation length. Otherwise the assumption about the local independence of the reaction probability from the inter-particle distance does not hold. If the requirement of well-mixed sub-volumes is not satisfied, spatial fluctuations are averaged which is clearly visible in Figure 4. For a small number of volume sub-divisions both the noise in the protein level and the distribution of times between bindings approach the prediction from the CME model. On the other hand, if the number of sub-volumes increases up to the limit where two molecules fill completely one sub-volume,[4] the first-passage probability gradually recovers the desired characteristics typical for the diffusion process in a closed volume: power-law behavior for short times and exponential decay for long times (Fig. 5).

A word of caution about the notion of *exact prediction of fluctuations* needs to be added here. Although we treat noise computed with GFRD as a reference value, one should bear in mind that this is a result of fluctuations with a rather simple BD model for chemical reactions. We are ignoring here other, possibly important, microscopic effects, like hydrodynamic interactions, electrostatic forces or molecular crowding. These certainly affect the diffusion process (Echevería *et al.*, 2006) but their significance for enhancing noise in biological systems is an open issue. Methods like GFRD which solve numerically the Smoluchowski model for diffusion-limited chemical reactions will provide an upper bound for the magnitude of fluctuations if compared to mesoscopic methods based on the master equation. The latter contain additional physical assumptions (Table 1) in order to simplify computations at the cost of averaging microscopic phenomena.

In Section 5.3 we argue that for a reversible reaction of a pair of particles the methods reproduce the first-passage probability differently, which is the cause of the variation in noise for the gene expression case. The power-law behavior for short departures from the target diminishes as the spatial detail is decreased, which is equivalent to: increasing the size of sub-volumes in MesoRD and GMP, increasing the difference between the binding and the unbinding radius for the reversible reaction in Smoldyn, and obviously the well-mixed postulate in SSA. The power-law region constricts also with increasing number of molecules or with accelerating the diffusion process (not shown in this study). Then the system moves away from the *low-molecule-number* and the *short-correlation-length* regime and the distribution converges to the mean-field exponential behavior. This can be properly approximated either by RDME-level methods with a coarse discretization or simply by the SSA algorithm. If the *overall* forward reaction rate k_{ON} is taken instead of the intrinsic κ_{a}, the SSA is also able to reproduce the exponential decay of the first-passage probability equal to the one obtained from GFRD or RDME methods with a large number of sub-volumes. However, for obvious reasons the power-law part is not recovered (in SSA the next event is drawn from the Poisson distribution), and hence the average inter-binding time is larger than that of GFRD.

Whereas the previous examples show the biophysical behavior of the methods for diffusion-limited reactions, we discuss in Sections 5.4 and 5.5 a more realistic biological problem, the chemotaxis pathway in *E.coli* (Lipkow, 2006; Lipkow *et al.*, 2005). There it is shown that modeling aspects and their consequences for the computational approaches can result in different predictions of averages and noise.

Table 5. Scaling of the computational cost for the spatial discrete methods presented in this comparison

Method	Computational cost	Comment
GFRD (event-driven)	$\sim \sum_S N_S$ $\sim \sum_{N_R} \prod_{S \in R} N_S$	Diffusive movements. Reactive distances.
Smoldyn (fixed time step)	As GFRD	As GFRD.
MesoRD (event-driven)	$\sim \log N_R$ $\sim \log N_{sv}$ $\sim \langle \tau_D \rangle^{-1}$ $\sim \langle \tau_R \rangle^{-1}$	Gibson and Bruck (2000). A sub-volume adds a diffusive reaction.
GMP*	$\sim \sum_S N_S$ ** $\sim N_R$ $\sim \tau_D^{-1}$ $\sim \langle \tau_R \rangle^{-1}$	Diffusive movements. As in the SSA. Fixed diffusion time step.

where:
$\langle \tau_D \rangle \propto L_{sv}^2/D \cdot N_{sv}/N_S, \quad \langle \tau_R \rangle \propto L_{sv}^3/k_R \cdot \prod_{S \in R} N_{sv}/N_S, \; \tau_D \propto L^2/D \cdot N_{sv}^{-2/3}$.

*The scheme is event-driven for reactions but the diffusion time step τ_D is fixed. The diffusion time step is assigned for every diffusing species.

**If $N_S/N_{sv} > 90$ molecules are moved in bulk, otherwise one-by-one in τ_D. Note that for event-driven schemes, the cost of diffusive movements or of computing reactive distances is given per iteration time step.

Abbreviations: N_S—number of molecules of a given species, N_R—number of reaction channels, N_{sv}—the total number of sub-volumes, $\langle \tau_R \rangle$—average time between reactions, $\langle \tau_D \rangle$—average time between diffusive movements, D—diffusion coefficient, k_R—rate of reaction R, L_{sv}—length of the sub-volume, L—length of the total volume.

Qualitative estimates addressing the computational cost of the mesoscopic methods considered here are given in Table 5. GFRD and Smoldyn scale in a manner typical for BD-based methods. Their main computational cost lies in computing the next position for every molecule (involves drawing few random numbers) and computing distances between reacting molecules, typically a N^2 operation if no neighbor list technique is applied. Differences in computational time may arise, however, because GFRD is an event-driven scheme while Smoldyn uses a fixed time step. Choosing the right Δt in the latter is not a completely arbitrary procedure since one has to assure that the probability of events per time step is *small*. In GFRD the maximum simulation time step during an iteration depends on the distance of the molecules to the target. If the total number of molecules decreases, the inter-particle distances increase, thus making a larger time step possible. Using similar arguments one can explain differences in the cost of performing diffusion in MesoRD and GMP (the first is an event-driven scheme, the latter uses a fixed Δt). Obviously the number of molecules of a given species in the sub-volume has to be used instead of the inter-particle distance. Then, the average time between diffusive jumps, $\langle \tau_D \rangle$, in MesoRD is inversely proportional to that quantity (see the caption of Table 5. Additionally, thanks to the next sub-volume method, MesoRD finds sub-volumes where the next event occurs instead of looping over the whole volume. On the other hand, GMP favors higher densities because, contrary to all other methods, particles can be diffused in bulk rather than one-by-one. The computational cost of the two

[4] In the physical picture, if the size of the sub-volume is further decreased, the probability of finding two molecules, which in reality have a certain diameter, is zero. This can be done in RDME-level methods since they do not model single molecules but their population in sub-volumes. Such an operation is not physical, however [see Materials and Methods section in Fange and Elf (2006)].

RDME-level methods differs also in scaling with the number of reaction channels N_R. The usage of the SSA scheme in GMP results in linear scaling with N_R; MesoRD achieves approximately log N_R scaling. Note that a diffusion event in the latter method is treated similarly as a reaction, and is also entered into an event queue.

Results for the test cases give an additional indication of performance. In case of the gene expression model general tools like Smoldyn, MesoRD or GMP will not outperform significantly the supposedly more expensive GFRD since they are not optimized for this very specific problem. On the other hand GFRD needs tailoring to every new problem and in general is implementation-wise difficult to adjust in order to tackle bigger problems. Computations of the CheY model showed a similar performance of all the methods. Smoldyn was only approximately twice slower than MesoRD and GMP, and appeared to be the most flexible method from the modeling point of view. For example, it allows to construct a more realistic, disc shape of the FliM motor cluster without any additional efficiency penalty. Such geometry of the motor implemented in RDME-level methods requires a much finer spatial resolution, which adds a significant computational cost.

ACKNOWLEDGEMENTS

The authors would like to thank Jeroen van Zon and Marco Morelli for the code of GFRD, Karen Lipkow for the Smoldyn input files for the CheY diffusion model. This work was supported by the Netherlands Organization for Scientific Research, project NWO-CLS-635.100.007, and by the Dutch BSIK/BRICKS project.

Conflict of Interest: none declared.

REFERENCES

Acar,M. *et al.* (2005) Enhancement of cellular memory by reducing stochastic transitions. *Nature*, **435**, 228–232.

Agmon,N. and Szabo,A. (1990) Theory of reversible diffusion-influenced reactions. *J. Chem. Phys.*, **92**, 5270–5284.

Allen,M. and Tildesley,D. (2002) *Computer Simulation of Liquids*. Oxford University Press, Oxford.

Andrews,S.S. and Bray,D. (2004) Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Phys. Biol.*, **1**, 137–151.

Arkin,A. *et al.* (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected Escherichia coli cells. *Genetics*, **149**, 1633–1648.

Baras,F. and Mansour,M.M. (1996) Reaction-diffusion master equation: A comparison with microscopic simulations. *Phys. Rev. E*, **54**, 6139–6148.

Becskei,A. *et al.* (2005) Contributions of low molecule number and chromosomal positioning to stochastic gene expression. *Nat. Genet.*, **37**, 937–944.

Bhalla,U.S. (2004) Signaling in small subcellular volumes. I. Stochastic and diffusion effects on individual pathways. *Biophys. J.*, **87**, 733–744.

Chopard,B. *et al.* (1994) Multiparticle lattice gas automata for reaction diffusion systems. *Int. J. Mod. Phys. C*, **5**, 47–63.

Doubrovinski,K. and Howard,M. (2005) Stochastic model for Soj relocation dynamics in Bacillus subtilis. *Proc. Natl Acad. Sci. USA*, **102**, 9808–9813.

Echevería,C. *et al.* (2007) Diffusion and reaction in crowded environments. *J. Phys.: Condens. Matter*, **19**, 065146–065158.

Elf,J. and Ehrenberg,M. (2004) Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. *IEE Sys. Biol.*, **1**, 230–236.

Fange,D. and Elf,J. (2006) Noise-induced min phenotypes in E. coli. *PLoS Comp. Biol.*, **2**, 0637–0648.

Francke,C. *et al.* (2003) Why the phosphotransferase system of Escherichia coli escapes diffusion limitation. *Biophys. J.*, **85**, 612–622.

Gardiner,C.W. (1983) *Handbook of Stochastic Methods*. Springer-Verlag, Berlin.

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

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

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

Gillespie,D.T. (1992) A rigorious derivation of the chemical master equation. *Physica A*, **188**, 404–425.

Golding,I. *et al.* (2005) Real-time kinetics of gene activity in individual bacteria. *Cell*, **123**, 1025–1036.

Halford,S.E. and Marko,J.F. (2004) How do site-specific DNA-binding proteins find their targets? *Nucleic Acids Res.*, **32**, 3040–3052.

Hattne,J. *et al.* (2005) Stochastic reaction-diffusion simulation with MesoRD. *Bioinformatics*, **21**, 2923–2924.

Kærn,M. *et al.* (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat. Genet.*, **6**, 451–464.

Kierzek,A.M. *et al.* (2001) The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *J. Biol. Chem.*, **276**, 8165–8172.

Krishna,S. *et al.* (2005) Stochastic simulations of the origins and implications of long-tailed distributions in gene expression. *Proc. Natl Acad. Sci. USA*, **102**, 4771–4776.

Lipkow,K. (2006) Changing cellular location of CheZ predicted by molecular simulations. *PLoS Comp. Biol.*, **2**, 0301–0310.

Lipkow,K. *et al.* (2005) Simulated diffusion of phosphorylated CheY through the cytoplasm of Escherichia coli. *J. Bacteriol.*, **187**, 45–53.

Marion,G. *et al.* (2002) Spatial heterogeneity and the stability of reaction states in autocatalysis. *Phys. Rev. E*, **66**, 051915 (9pp).

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

Metzler,R. (2001) The future is noisy: The role of spatial fluctuations in genetic switching. *Phys. Rev. Lett.*, **87**, 068103 (4pp).

Pedraza,J.M. and van Oudenaarden,A. (2005) Noise propagation in gene networks. *Science*, **307**, 1965–1969.

Redner,S. (2001) *A Guide to First-Passage Processes*. Cambridge University Press, New York.

Rodríguez Vidal,J. *et al.* (2006) Spatial stochastic modelling of the phosphoenolpyruvate-dependent phosphotransferase (PTS) pathway in Escherichia coli. *Bioinformatics*, **22**, 1895–1901.

Rosenfeld,N. *et al.* (2005) Gene regulation at the single-cell level. *Science*, **307**, 1962–1965.

Shnerb,N.M. *et al.* (2000) The importance of being discrete: Life always wins on the surface. *Proc. Natl Acid. Sci. USA*, **97**, 10322–10324.

Stundzia,A.B. and Lumsden,C.J. (1996) Stochastic simulation of coupled reaction-diffusion processes. *J. Comp. Phys.*, **127**, 196–207.

Takahashi,K. *et al.* (2005) Space in systems biology of signaling pathways – towards intracellular molecular crowding in silico. *FEBS Lett.*, **579**, 1783–1788.

Togashi,Y. and Kaneko,K. (2001) Transitions induced by the discreteness of molecules in a small autocatalytic system. *Phys. Rev. Lett.*, **86**, 2459–2462.

Togashi,Y. and Kaneko,K. (2004) Molecular discreteness in reaction-diffusion systems yields steady states not seen in the continuum limit. *Phys. Rev. E*, **70**, 020901 (4pp).

Togashi,Y. and Kaneko,K. (2005) Discreteness-induced stochastic steady state in reaction diffusion systems: Self-consistent analysis and stochastic simulations. *Physica D*, **205**, 87–99.

van Kampen,N.G. (1997) *Stochastic Processes in Physics And Chemistry*. Elsevier, Amsterdam.

van Zon,J.S. and ten Wolde,P.R. (2005a) Green's-function reaction dynamics: A particle-based approach for simulating biochemical networks in time and space. *J. Chem. Phys.*, **123**, 234910 (16pp).

van Zon,J.S. and ten Wolde,P.R. (2005b) Simulating biochemical networks at the particle level and in time and space: Green's function reaction dynamics. *Phys. Rev. Lett.*, **94**, 128103.

van Zon,J.S. *et al.* (2006) Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophys. J.*, **91**, 4350–4367.

Zhdanov,V.P. (2002) Cellular oscillator with a small number of particles. *Eur. Phys. J. B*, **29**, 485–489.