

Modelling and Simulation of IntraCellular Dynamics: Choosing an Appropriate Framework

Olaf Wolkenhauer, Mukhtar Ullah, Walter Kolch, and Kwang-Hyun Cho.

Abstract—Systems biology, that is, mathematical modelling and simulation of biochemical reaction networks in intracellular processes has gained renewed interest in recent years. For most simulation tools and publications they are usually characterized by either preferring stochastic simulation or rate equation models. The use of stochastic simulation is occasionally accompanied with arguments against rate equations. Motivated by these arguments, in this paper we discuss the relationship between these two forms of representation. Towards this end we provide a novel compact derivation for the stochastic rate constant that forms the basis of the popular Gillespie algorithm. Comparing the mathematical basis of the two popular conceptual frameworks of generalized mass action models and the chemical master equation, we argue that some of the arguments that have been put forward are ignoring subtle differences and similarities that are important for answering the question in which conceptual framework one should investigate intracellular dynamics.

Index Terms—intracellular dynamics, generalized mass action models, chemical master equation, Gillespie algorithm.

I. INTRODUCTION

Mathematical modelling and simulation of intracellular dynamics has gained renewed interest in the area of systems biology [1]. For most simulation tools and publications they are usually characterized by either using stochastic simulation (e.g. [2]) or rate equations (e.g. [3], [4]). While there are good reasons for hypothesizing stochastic mechanisms [5], [6], some authors have tried to argue their use of stochastic simulation by suggesting differential equations were not suitable. In the present paper we are going to provide a critical discussion of some of these arguments and highlight subtle differences and relationships which have been ignored in some discussions. Towards this end we are going to investigate the mathematical basis and close relationship between generalized mass action models and stochastic simulation.

The paper is organized as follows. In the following section we introduce the two most commonly employed conceptual frameworks for modelling intracellular dynamics: the generalized mass action approach, using rate equations and the chemical master equation. This is followed by a derivation of the key elements of the Gillespie algorithms in Sections III and IV. In Section V we draw conclusions from the mathematical derivation and discuss the arguments that have been used in other publications. This is supported by a simulated example.

Correspondence should be addressed to O.Wolkenhauer, Systems Biology and Bioinformatics Group, University of Rostock, 18051 Rostock, Germany. www.sbi.uni-rostock.de

Manuscript received January 10, 2004.

II. RATE VERSUS MASTER EQUATIONS

We are considering a reaction network or pathway involving N molecular species S_i . A network, which may include reversible reactions, is decomposed into M unidirectional basic reaction channels R_μ

$$R_\mu: l_{\mu 1} S_{p(\mu,1)} + l_{\mu 2} S_{p(\mu,2)} + \cdots + l_{\mu L_\mu} S_{p(\mu,L_\mu)} \xrightarrow{k_\mu} \cdots$$

where L_μ is the number of reactant species in channel R_μ , $l_{\mu j}$ is the stoichiometric coefficient of reactant species $S_{p(\mu,j)}$, $K_\mu = \sum_{j=1}^{L_\mu} l_{\mu j}$ denotes the molecularity of reaction channel R_μ , and the index $p(\mu, j)$ selects those S_i participating in R_μ .

Assuming a constant temperature and that diffusion in the cell is fast, such that we can assume a homogeneously distributed mixture in a fixed volume V , we consider generalized mass action (GMA) models, consisting of N ordinary differential rate equations

$$\frac{d}{dt}[S_i] = \sum_{\mu=1}^M \nu_{\mu i} k_\mu \prod_{j=1}^{L_\mu} [S_{p(\mu,j)}]^{l_{\mu j}} \quad i = 1, 2, \dots, N \quad (1)$$

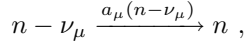
where the k_μ 's are rate constants and ν_μ denotes the change in molecules of S_i resulting from a single R_μ reaction. We write

$$[S] = S/V \quad \text{and} \quad \#S = S \cdot N_A,$$

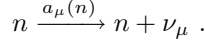
where N_A is the Avogadro number. The units of $[S]$ are mol per liter, $M = \text{mol/L}$. In this context then, S denotes the number of moles. GMA models have been widely used in modelling biochemical reactions and metabolic engineering [7], [8]. The mathematical representation (1) of a biochemical network does not account for noise on the states, which would lead to stochastic ODEs. Neither does it consider measurement noise, and we may call the model 'deterministic'. It is however not deterministic in the sense that it models molecules in a phase-momentum space and in fact it is rooted in the stochastic setting of Boltzmann's kinetic theory of gases and the $[S_i]$ are thus most probable values. In modelling intracellular dynamics, it has been argued that such "deterministic" models are not suitable for processes with relatively low numbers of molecules [9]–[12]. We are going to investigate this argument in detail.

In a stochastic framework, we are looking at populations of molecules and wish to determine for each molecular species S_i the probability $\text{Prob}\{\#S_i(t) = n_i\}$ that at time t there are n_i molecules. For N molecular species, let n denote the N -dimensional state-vector, whose values are positive integers, $n \in \mathbb{Z}_+^N$. $\nu_\mu \in \mathbb{Z}^N$ are the step-changes occurring for elementary reactions indexed by μ . If S is an N -dimensional

variable, we write $\text{Prob}\{\#S=n\} = P_n(t)$. Describing the changes in random variable S , we consider the following two state-transitions: a) from other states to state n , denoted



where $a_\mu(n - \nu_\mu)$ is referred to as the *propensity* of reaction channel R_μ , that is the probability per unit time, of a change ν_μ occurring, given that we are in state $n - \nu_\mu$. Secondly, b) moving away from state n is given as



From these definitions we arrive at the *chemical master equation* (CME)

$$\frac{\partial P_n(t)}{\partial t} = \sum_{\mu=1}^M [a_\mu(n - \nu_\mu)P_{(n-\nu_\mu)}(t) - a_\mu(n)P_n(t)] . \quad (2)$$

The first term on the right-hand side describes the change to state n , while the second term describes the changes away from state n . The product of the propensity with the probability should be read as an “and”. The multiplication of a propensity and probability makes sense in light of the derivative against time on the left, in that a propensity, multiplied with dt gives a probability. Note that the CME (and the therefore the Gillespie algorithm) does account for individual reaction channels but not for individual molecules. This issue was taken up, for example, in [11].

A major difficulty with the CME is that the dimension of these sets of equations depends not only on the number of chemical species N but for any possible number of molecules of any species we have n differential-difference equations. To avoid these difficulties, Gillespie [13]–[16] developed an algorithm to simulate a CME model efficiently. The Gillespie approach to stochastic *simulation* has in recent years become popular in *simulating* intracellular dynamic processes [2], [6], [10], [17]. Some authors have unfortunately confused the simulation of a stochastic model with a stochastic model. The Gillespie algorithm simulates the CME and, as we are going to show, does this in most cases based on the knowledge of the rate equation model. To argue a case for an alternative can thus be mistaken. While a formal analysis of (2) is difficult, it is possible to approximate the CME by a truncated Taylor expansion, leading to the Fokker-Planck equation, for which there exist some results [9], [18], [19]. Comparing (1) and (2), we note that while rate equations are deterministic in the sense that they employ differential equations, they are based on a probabilistic description of molecular kinetics. On the other hand, the CME is a stochastic formulation, but based on differential equations, with probabilities as variables. One should therefore avoid identifying differential equations as synonymously with ‘deterministic’ representations.

In the next section we discuss the key elements of the Gillespie algorithm with respect to the number of molecules and possible approximations. This is followed, in Section IV, by a derivation of the stochastic rate constant. The derivation is going to highlight the relationship between generalized mass action models and a stochastic simulation of the CME.

III. STOCHASTIC SIMULATION

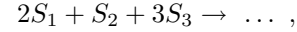
The Gillespie algorithm determines for each iteration first the propensity a_μ for all of the elementary reactions R_μ

$$a_\mu = h_\mu \cdot c_\mu \quad \mu = 1, \dots, M \quad (3)$$

where h_μ defines the number of distinct combinations of R_μ reactant molecules, which varies over time

$$h_\mu(n) = \begin{cases} \prod_{j=1}^{L_\mu} \binom{n_{p(\mu,j)}}{l_{\mu j}} & \text{for } n_{p(\mu,j)} > 0 , \\ 0 & \text{otherwise} . \end{cases} \quad (4)$$

For example, let R_μ be defined as



then $L_\mu = 3$, $l_{\mu 1} = 2$, $l_{\mu 2} = 1$, and $l_{\mu 3} = 3$, such that

$$h_\mu = \prod_{j=1}^{L_\mu} \binom{n_{p(\mu,j)}}{l_{\mu j}} = \binom{n_1}{2} \cdot \binom{n_2}{1} \cdot \binom{n_3}{3} .$$

If $n_{p(\mu,j)}$ in (4) is large and $l_{\mu j} > 1$, terms like $(n_{p(\mu,j)} - 1)$, \dots , $(n_{p(\mu,j)} - l_{\mu j} + 1)$ will not be much different from $n_{p(\mu,j)}$ and we may write

$$h_\mu \cong \prod_{j=1}^{L_\mu} \frac{(n_{p(\mu,j)})^{l_{\mu j}}}{l_{\mu j}!} = \frac{\prod_{j=1}^{L_\mu} (n_{p(\mu,j)})^{l_{\mu j}}}{\prod_{j=1}^{L_\mu} l_{\mu j}!} . \quad (5)$$

It should however be noted that this is an approximation, which can effect results in studies that compare GMA models with $l_{\mu j} > 1$ and stochastic simulations for small molecular populations. In fact, as we are going to show in Section IV, it is misleading to compare a GMA model with stochastic simulation *as alternatives*.

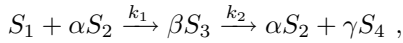
Akin to a rate constant k_μ , Gillespie introduced the *stochastic rate constant*, c_μ , which only depends on physical properties of the molecules and the temperature of the system. $c_\mu dt$ is the probability that a particular selected combination of R_μ reactant molecules at time t will react in the next infinitesimal time interval $(t, t + dt)$. A reaction requires two separate phenomena: a collision to occur and for the collision to be reactive. In [20], [21], for bimolecular reactions, Gillespie derived an expression for c_μ that contains a probability that a colliding pair of R_μ reactant molecules will chemically react. This probability is generally unknown. For trimolecular reactions the only relationship that can be derived from physical principles is the proportionality $c_\mu \propto V^{-K_\mu+1}$, where $K_\mu = 1, 2$, or 3 , and even this requires further unrealistic assumptions as Gillespie admits in [21]. Since a physical derivation for c_μ is in general not possible, implementations of c_μ in algorithms are relying on other arguments. Such a derivation will be the subject of Section IV. It turns out that such derivations rely on the rate constants k in GMA models. In [21] Gillespie also showed how the linear relationship $c_\mu dt$ is justified on a mathematical basis. A consequence of this derivation is that c_μ must be analytical. This can, for example, be achieved by keeping c_μ constant. If we multiply the probability $c_\mu dt$, which applies to a particular selected

combination of reactant molecules, by the total number of distinct combinations of R_μ reactant molecules in V at time t , we obtain the probability that an R_μ will occur somewhere inside V in the next infinitesimal time interval $(t, t+dt)$. This leads us to $c_\mu \cdot h_\mu dt \equiv a_\mu dt$ as the probability that an R_μ reaction will occur in V in $(t, t+dt)$, given that the system is in state S at time t .

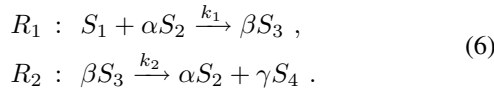
IV. DERIVING THE STOCHASTIC RATE CONSTANT

In the present section we are going to present a novel compact derivation for c_μ , which is also going to provide a means to discuss the relationship between rate equations and stochastic simulation.

Using the following example for a chemical reaction



which is split into two reaction channels



When a reaction occurs, the changes to molecule populations are

$$\nu_1 = (-1, -\alpha, \beta, 0) , \quad \nu_2 = (0, \alpha, -\beta, \gamma) .$$

From (6) and from the definition of reaction velocity we have the following relationships

$$\left\{ \begin{aligned} \left[\frac{\dot{S}_1}{-1} = \frac{\dot{S}_2}{-\alpha} = \frac{\dot{S}_3}{\beta} \right] &= k_1 [S_1][S_2]^\alpha , \\ \left[\frac{\dot{S}_3}{-\beta} = \frac{\dot{S}_2}{\alpha} = \frac{\dot{S}_4}{\gamma} \right] &= k_2 [S_3]^\beta . \end{aligned} \right. \quad (7)$$

The rate equations are then easily derived as

$$\left. \begin{aligned} d[S_1]/dt &= -k_1 [S_1][S_2]^\alpha \\ d[S_2]/dt &= -\alpha k_1 [S_1][S_2]^\alpha + \alpha k_2 [S_3]^\beta \\ d[S_3]/dt &= \beta k_1 [S_1][S_2]^\alpha - \beta k_2 [S_3]^\beta \\ d[S_4]/dt &= \gamma k_2 [S_3]^\beta . \end{aligned} \right\} \quad (8)$$

Looking at the structure of (8), we recognize in this set of equations the GMA representation (1). Substituting $[S] = S/V = \langle \#S \rangle / (N_A V)$ in (1), gives

$$\frac{d}{dt} \left(\frac{\langle \#S_i \rangle}{N_A V} \right) = \sum_{\mu=1}^M \nu_{\mu i} k'_\mu \prod_{j=1}^{L_\mu} \left(\frac{\langle \#S_{p(\mu,j)} \rangle}{N_A V} \right)^{l_{\mu j}} ,$$

which can be rewritten as

$$\frac{d}{dt} \langle \#S_i \rangle = \sum_{\mu=1}^M \frac{\nu_{\mu i} k'_\mu}{(N_A V)^{K_\mu-1}} \prod_{j=1}^{L_\mu} \langle \#S_{p(\mu,j)} \rangle^{l_{\mu j}} , \quad (9)$$

where we made use of the fact that

$$\prod_{j=1}^{L_\mu} (N_A V)^{l_{\mu j}} = (N_A V)^{\sum_{j=1}^{L_\mu} l_{\mu j}} = (N_A V)^{K_\mu} .$$

The differential operator is justified only with the assumption of large numbers of molecules involved, such that near continuous changes are observed. Let us now assert for the temporal evolution of $\langle \#S_i \rangle$ a ‘‘particle-ODE’’, :

$$\frac{d}{dt} \langle \#S_i \rangle = \sum_{\mu=1}^M \nu_{\mu i} k'_\mu \prod_{j=1}^{L_\mu} \langle \#S_{p(\mu,j)} \rangle^{l_{\mu j}} . \quad (10)$$

Comparing (10) with (9), we find

$$k'_\mu = \frac{k_\mu}{(N_A V)^{K_\mu-1}} , \quad (11)$$

This equation then describes the interpretation of the rate constant, dependent on whether we consider concentrations or counts of molecules.

To arrive at a general expression for the propensity a_μ from (10) it follows that

$$\langle \#R_\mu \rangle = k'_\mu \cdot \prod_{j=1}^{L_\mu} \langle \#S_{p(\mu,j)} \rangle^{l_{\mu j}} dt \quad (12)$$

gives the average number of R_μ reactions occurring in $(t, t+dt)$. Note that ν_μ has been excluded above since we would otherwise have an expression for the number of molecules not the number of reactions. Considering $\#R_\mu$, the number of R_μ reactions, as a discrete random variable with probability mass function $p_{r_\mu} = \text{Prob}\{\#R_\mu = r_\mu\}$. The expectation value $\langle \#R_\mu \rangle$ is given by

$$\langle \#R_\mu \rangle = \sum_{r_\mu} r_\mu \langle p_{r_\mu} \rangle \quad r_\mu = 0, 1, 2, \dots \quad (13)$$

where

$$p_{r_\mu} = \begin{cases} a_\mu dt + o(dt) & : r_\mu = 1 \\ 1 - a_\mu dt + o(dt) & : r_\mu = 0 \\ o(dt) & : r_\mu > 1 . \end{cases} \quad (14)$$

where $o(dt)$ denotes a negligible probability for more than one R_μ reaction to occur during dt . Since a_μ is a function of n , p_{r_μ} is randomly varying and hence the averaging $\langle p_{r_\mu} \rangle$ over the ensemble in (13). Equation (13) thus becomes

$$\langle \#R_\mu \rangle = 0 \cdot p_0 + 1 \cdot p_1 + \sum_{r_\mu > 1} r_\mu \langle p_{r_\mu} \rangle .$$

From (13) and (14) we then have

$$\langle \#R_\mu \rangle = \langle a_\mu dt \rangle + o(dt) . \quad (15)$$

Now, from (12) and (15) the propensity of R_μ reactions to occur in dt is given as

$$\langle a_\mu \rangle = k'_\mu \prod_{j=1}^{L_\mu} \langle \#S_{p(\mu,j)} \rangle^{l_{\mu j}} . \quad (16)$$

To proceed, we consider another alternative expression for a_μ , by substituting (5) into (3), and considering the average

$$\langle a_\mu \rangle = c_\mu \cdot \left\langle \frac{\prod_{j=1}^{L_\mu} (\#S_{p(\mu,j)})^{l_{\mu j}}}{\prod_{j=1}^{L_\mu} (l_{\mu j}!)} \right\rangle , \quad (17)$$

where $\#S_{p(\mu,j)}$ denotes the random variable whose value is $n_{p(\mu,j)}$. Note that this implied the assumption of a large number of molecules for all species S_i and $l_{\mu j} > 1$. Comparing (17) and (16)

$$k'_\mu \prod_{j=1}^{L_\mu} \langle \#S_{p(\mu,j)} \rangle^{l_{\mu j}} = \frac{c_\mu \left\langle \prod_{j=1}^{L_\mu} (\#S_{p(\mu,j)})^{l_{\mu j}} \right\rangle}{\prod_{j=1}^{L_\mu} (l_{\mu j}!)}$$

Making the same notorious assumption¹ of zero covariance as in [21], gives

$$k'_\mu = \frac{c_\mu}{\prod_{j=1}^{L_\mu} (l_{\mu j}!)}, \quad (18)$$

which can be turned into an expression for c_μ :

$$c_\mu = k'_\mu \cdot \prod_{j=1}^{L_\mu} (l_{\mu j}!) \quad (19)$$

Inserting (11) for k'_μ , we arrive at

$$c_\mu = \left(\frac{k_\mu}{(N_A V)^{K_\mu - 1}} \right) \cdot \prod_{j=1}^{L_\mu} (l_{\mu j}!) \quad (20)$$

Equation (20) establishes a relationship between the stochastic constant c_μ and rate constant k_μ and is used in most implementations of Gillespie's algorithm. Note that if above we substitute S/V in (1) for $[S]$ instead of $\langle \#S \rangle / (N_A V)$, the only difference to (11) and (20) is that the N_A would not appear in these equations.

The difference of our derivation to the one given by Gillespie in [21] is that we introduced the average number of reactions (12) to move from the general GMA representation (1), which is independent of particular examples, to an expression that allows us to derive parameter c_μ of the stochastic simulation (20) without making reference to the temporal evolution of moments from the CME. In [21], the temporal evolution of the mean is derived for examples of bi- and tri-molecular reactions. Taking the variance of $\#S(t)$ to be zero to make it a 'sure variable', this equation is compared to the GMA model to derive c_μ .

Equation (20) is at the heart of the Gillespie algorithm and its implementations. There are two conclusions from the derivation. First, using the approximation (5) for h_μ is valid for a large number of molecules with $l_{\mu j} > 1$. Although in most practical cases this will not lead to significant differences, this has been ignored by some authors. More important however is the fact that the derivation of (20) relies on the rate constants of the GMA model. In this respect, it does not make sense to compare a GMA model and a stochastic simulation *as alternatives* if the stochastic rate constant c_μ is derived from the rate constants of the GMA model.

¹The assumption of zero covariance such that $\langle \#S_i \#S_j \rangle = \langle \#S_i \rangle \langle \#S_j \rangle$ means for $i \neq j$ nullifying correlation, and for $i = j$ nullifying random fluctuations. The same assumption is required if one compares the temporal evolution of the mean of the CME model with the GMA model. This then demonstrates that a GMA model does *not* always arise as the mean of the CME model [21] (page 363).

So how do we compare deterministic and stochastic models? First, we ought to compare models with models and simulations with simulations. The advantage of the GMA model (1) is that its terms and parameters have a precise meaning, they are a direct translation of the biochemical reaction diagrams that capture the biochemical relationships of the molecules involved. For a formal analysis of the model, as opposed to a simulation, rate equations are in virtually all cases simpler than the CME. One might argue that for any realistic pathway model a formal analysis is not feasible for either model and a simulation (numerical solution) is the way to go forward. In this case the Gillespie algorithm provides an efficient implementation to generate realizations of the CME. An advantage of simulations is furthermore that, in principle, it is possible to vary temperature and volume over time.

V. DISCUSSION

As concentrations and the number of molecules becomes small, the variability of molecular populations in biochemical reaction networks increases. It is frequently argued that in this case differential equation models do not account for the observed variability and a stochastic approach should be preferred. To account for variability in chemical master equations (2) and rate equations (1), for *both* conceptual frameworks the identification of the model and its parameters requires a set of replicate experimental time series over which to average and estimate the moments of the distributions that account for the variability. While there are indeed good reasons to hypothesize stochastic mechanisms in intracellular dynamics (see [6] for a recent overview), the arguments used for stochastic simulation and against differential equations are occasionally misguided.

One ought to differentiate between a hypothesized principle or molecular mechanism and the observations we make from experimental data. While rate equations are deterministic in the sense that they employ differential equations, they are based on a probabilistic description of molecular kinetics. On the other hand, the chemical master equation is a stochastic formulation, but based on differential equations, with probabilities as variables. The Gillespie algorithm, as is used in most publications, realizes the chemical master equation but thereby makes explicit use of the rate constants that define the generalized mass action model.

A common argument is that if the concentration or the number of molecules of the chemical species involved in a biochemical reaction is low, a stochastic approach in form of the chemical master equation is a more accurate representation than rate equations [9], [10], [12], [17]. In case of [10], [17] and [12] this discussion is not done on the basis of the chemical master equation but using the Gillespie algorithm for stochastic simulation. A question is what is meant by "low concentrations" or the consequences of small numbers of molecules? In [10] a figure of the order of less than a few thousand is given. In [9] the copy number of proteins is cited as less than a hundred. Since the number of molecules for most reactant species reduces either to very small values or increases steadily for others, we assume that authors, speaking of 'numbers of molecules' refer to initial numbers

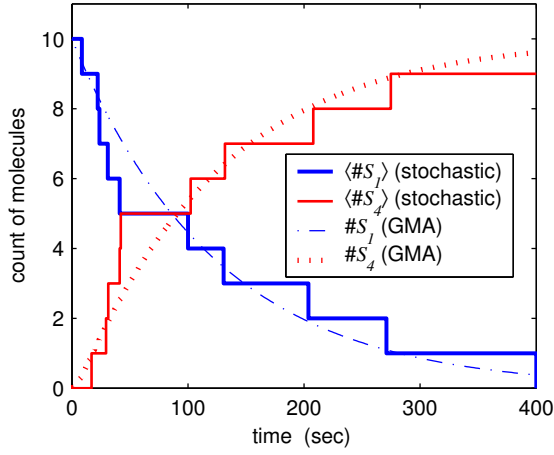


Fig. 1. Single run of a stochastic simulation using Gillespie's Direct Method [14] for Example (6) in the text. The parameters used are $V = 1$ pL, $k_1 = 0.5$ (nM \cdot sec) $^{-1}$, $k_2 = 0.2$ sec $^{-1}$, $\alpha = 1$, $\beta = 1$, $\gamma = 1$, $\#S_2(0) = \#S_3(0) = 0$.

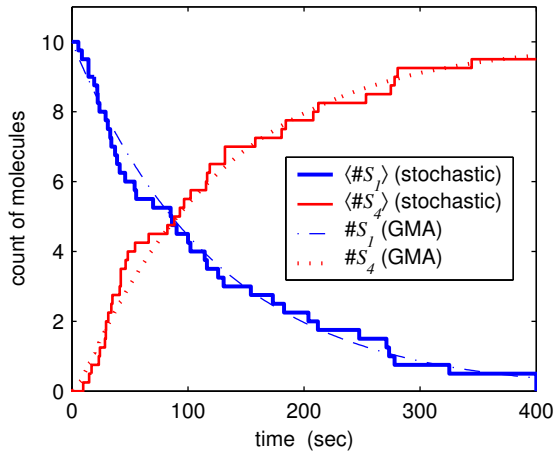


Fig. 2. Average over four realizations of a stochastic simulation of Example (6), using the same parameters as in Figure 1. The GMA solutions have been multiplied by $N_A \cdot V \cdot 10^{-9}$ to convert concentrations into a count of molecules.

at time zero. Subject to approximation (5), Figures 1 and 2 compare realizations of the stochastic simulation of Example (6) with solutions of the rate equations². Figure 3 shows the temporal evolution of a_μ for a volume of 1 pL and initial numbers of 10 molecules. The simulations demonstrate that even for very small numbers of molecules single realizations of stochastic simulations show steadily changing patterns that can be modelled well using a continuous representation. The close similarity between the numerical solution of the ODEs and the stochastic simulation is no surprise since the rate constants of the GMA model are also integral part of the stochastic simulation, as shown by equation (20).

In fact, plots shown in those publications that argue for stochastic simulation in case of small molecule populations, are almost always displaying steady increases and decreases that are well approximated by ordinary differential equations.

²MATLAB functions to simulate the GMA model (1) and the CME (2) are available from www.sbi.uni-rostock.de

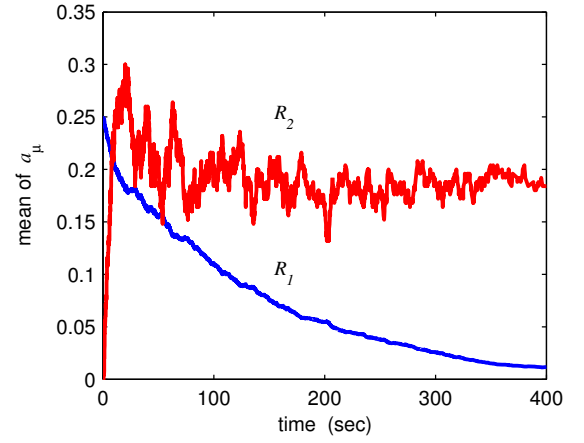


Fig. 3. Example (6). Temporal evolution of the a_μ . The parameters are the same as in Figure 1. What is shown is an average of the a_μ over realizations in order to illustrate the overall trend, free of the variability in single realizations.

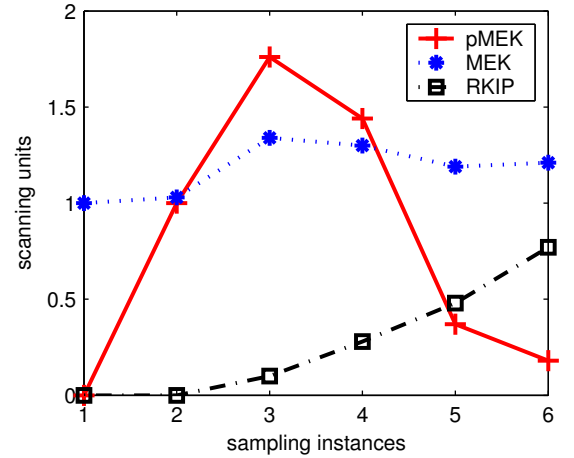


Fig. 4. Data from western blot time course experiments.

Figure 4 shows typical experimental data as obtained from western blot time course experiments to study proteins in signal transduction pathways. While there surely is measurement noise, it seems a reasonable assumption to believe the concentration profiles follow roughly the linear interpolations shown. If for the few time points that current experimental practice generates, we were not observing steadily increasing or decreasing pattern, and instead would argue for a truly random process, we would have a hard time to validate such a model from data like those shown in Figure 4. Figure 5 shows random simulations of time series with only six time points. How do we distinguish between random from deterministic pattern in data?

Western-blot time series, like those shown in Figure 4, are generated from a pool of about 10^7 cells although we are trying to understand what is happening in a cell. We could explain the deterministic pattern in experimental data as follows. Looking at the population of molecules of species S_i , from each reaction channel a change $\nu_{\mu i}$ arises for when the reaction channel R_μ is realized or active. The change $\nu_{\mu i}$ is a random variable, and the total change of S_i across all

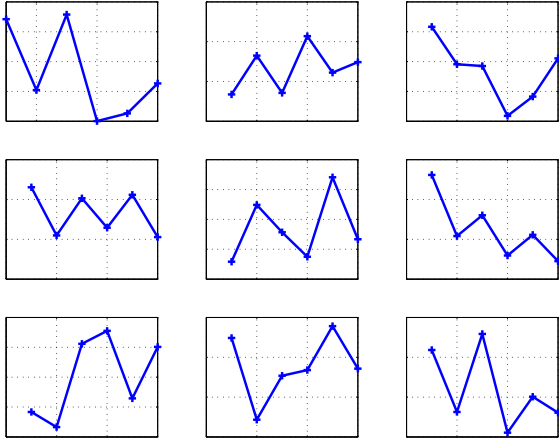


Fig. 5. Random simulations of time series. We can only start modelling by assuming that these curves are not random. If they are, we could not test this since there are not enough data points for a statistical test to be applicable. We call this the WYSIWYM principle: What You See Is What You Model!

reaction channels is a sum of random variables

$$\Delta(\#S_i) = \sum_{\mu=1}^M \nu_{\mu i}.$$

For more than one reaction channel, from the central limit theorem, $\Delta(\#S_i)$ is approximately normal distributed, $\Delta(\#S_i) \sim \mathcal{N}(\cdot, \sigma_v^2)$. For any further averaging process with say m elements, e.g., using 10^7 cells in immunoblotting, the variance of measurements is of the order σ_v^2/m . This means that if we are not considering single-cell measurements we are likely to observe relatively smooth patterns. If we do consider single-cell measurements, we ought to have in any case replicates to average out random variations.

If we are to consider a stochastic simulation and wish to validate it with experimental data, we get the following requirements for the experimenters. In Gillespie's algorithm, the time interval for the next reaction to occur is calculated as

$$\tau = (1/a^*) \cdot \ln(1/r_1),$$

where r_1 is a random number in the unit interval and

$$a^* = \sum_{\mu} a_{\mu}. \quad (21)$$

Note that τ is a function of state n and thus implicitly also a function of time. As $\#S_i$ goes down, there are fewer reactive collisions and the propensity a_{μ} decreases. This means that for all relevant reaction channels, (21) will also decrease. This does however mean that the ratio a_{μ}/a^* changes little. Since the probability of the next reaction occurring is given by [13]

$$P(\mu|\tau) = a_{\mu}/a^*, \quad (22)$$

this means that the resulting concentration levels are relatively similar. However, since τ , i.e., the time for the next reaction to occur is exponentially distributed,

$$P(\tau) = a^* \cdot \exp(-a^* \tau), \quad (23)$$

with mean $1/a^*$ and standard deviation $1/a^*$, the variance of τ increases more substantially. This in turn means, that

for a specified t the variance of the realizations will increase. Figure 6 illustrates the dependence of the variability on the initial number of molecules. A consequence is that for fewer molecules, more realizations are required to obtain an accurate picture through averaging across realizations. Also, the larger the number of reaction channels, M , the smaller is the average time to the next reaction τ , as shown by (23). However, at the same time, the number of possible transitions from state n will increase as can be seen from (22).

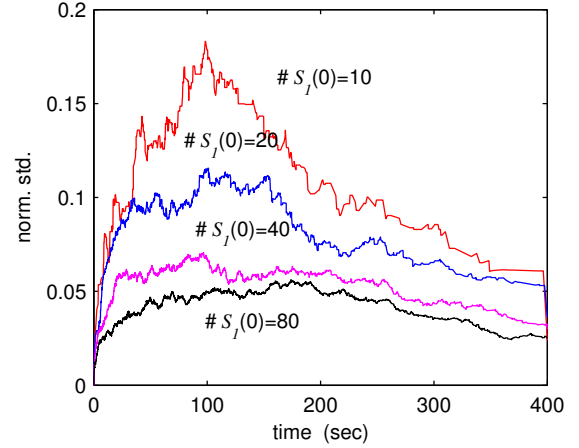


Fig. 6. Example (6). Temporal evolution of the normalized standard deviation $\sigma_{S_i}(t)/(\#S_i(0))$ over 50 realizations at t . $\alpha = \beta = \gamma = 1$, $k_1 = 0.6 \text{ (nM} \cdot \text{sec)}^{-1}$, $k_2 = 0.1 \text{ sec}^{-1}$. $\#S_1(0) = 10$, $\#S_2(0) = \#S_3(0) = 0$, $\#S_1(0) = 10, 20, 40, 80$. Note that the normalization is necessary to make the plots independent of the initial $\#S_i$ and thereby make them comparable.

In considering mathematical modelling and simulation, most important are the context and the purpose of modelling. Do we wish to use a model to hypothesize a fundamental molecular mechanism, or are we trying to model the observed consequences of these molecular mechanisms? Is the phenomena we observe an aggregate of a group of dependent subcomponents (e.g. molecules or cells) that combine individual, discrete responses into graded response at higher levels of organization (e.g. in tissue and organs)?

In some cases, authors who wished to argue for their use of stochastic simulation, have unfortunately missed some of the subtleties of our foregoing discussion. Let us look at some examples. In [17] it is argued that

“The availability of a *huge* amount of molecular data concerning various biochemical reactions provoked numerous attempts to study the dynamics of cellular processes by means of kinetic models and computer simulations.” (emphasis is ours).

To take western blot time course experiments as an example, the problem we face for modelling is anything but one of dealing with huge amounts of data. For a time series, usually only six to ten time points are available and replicates are the subject of hard fought negotiations between modellers and biologists. For realistic pathway models, because of the costs and time required, usually only a fraction of all pathway proteins can be covered by experiments.

The authors of [12] clearly missed the mark:

“There is also a problem of interpretation by users. Systems of differential equations have a number of parameters that must be fitted from experimental data. However, the parameters may have no meaning to the biologists, who are therefore unable to gauge whether the values are appropriate.”

Quite the opposite is true. The parameters of GMA model (1) have a very precise meaning, which can be fitted from experimental data. We would argue the fact that, for GMA models, we can identify parameters directly from experimental data is an advantage. Although this is not a trivial task, there are well established algorithms available for this purpose. Why would the authors of [12] think the CME (2) is more intuitive, and how would they validate their models?

Whether starts off with the GMA representation or the CME, it is often not possible to obtain all necessary parameter values from experimental data. For such practical reasons but also in order to simplify the mathematical analysis it is often desirable to make use of the quasi-steady-state assumption (QSSA) [22], [23]. The QSSA implies that for the time scale of interest the instantaneous rates of change of intermediate species are approximately equal to zero. Modelling signal transduction pathways, the consecutive activation of kinases is commonly described through phosphorylation and dephosphorylation steps, equivalent to enzyme kinetic reactions. Assuming the concentration of kinase-substrate complexes is small compared with the total concentration of the reactants, phosphorylation is modelled as a bimolecular reaction and assuming that the concentration of active phosphatase is constant, dephosphorylation can be modelled as a first order reaction. Such assumptions allow a formal analysis of various important aspects of cell signalling rooted in GMA models. See [24] for an outstanding example of such an analysis. These simplifications do of course also simplify the stochastic simulation since the k 's of the rate constants are implicitly used in the simulation. Alternatively, one considers the QSSA for the CME as discussed in [25].

We conclude that one should not argue the case for either rate equations or stochastic simulations with the numerical accuracy of a representation or physical realism but whether a biological principle is reflected by the model. Whether we are using the GMA or CME representation we make various assumptions about the physical context, including for example a constant volume, temperature, rapid diffusion etc. While these assumptions may seem outrageous in light of what we know and observe about intracellular dynamics, we are reminded of Box's dictum: “All models are wrong, but some are useful”.

ACKNOWLEDGMENT

This work was in parts supported by the U.K. Government Department for Environment, Food and Rural Affairs (DEFRA) as part of the *M.bovis* post-genomics programme. This project is conducted in collaboration with the Veterinary Laboratories Agency (VLA), Weybridge, U.K. K.-H.Cho acknowledges the support by a systems biology grant from the Korean Ministry of Science and Technology (M10309000006-03B5000-00211). Allan Muir contributed helpful discussions

on moving from probabilistic models to differential equations via expectations.

REFERENCES

- [1] O. Wolkenhauer, H. Kitano, and K.-H. Cho, “Systems biology: Looking at opportunities and challenges in applying systems theory to molecular and cell biology,” *IEEE Control Systems Magazine*, vol. 23, no. 4, pp. 38–48, August 2003.
- [2] H. McAdams and A. Arkin, “Stochastic mechanisms in gene expression,” *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 814–819, 1997.
- [3] K.-H. Cho, S.-Y. Shin, H.-W. Lee, and O. Wolkenhauer, “Investigations into the analysis and modelling of the TNF α -mediated NF- κ B-signaling pathway,” *Genome Research*, vol. 13, pp. 2413–2422, 2003.
- [4] I. Swameye, T. Müller, J. Timmer, O. Sandra, and U. Klingmüller, “Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by data-based modeling,” *Proc. Natl. Acad. Sci.*, vol. 100, pp. 1028–1033, 2003.
- [5] H. McAdams and A. Arkin, “It’s a noisy business!” *Trends in Genetics*, vol. 15, pp. 65–69, February 1999.
- [6] C. Rao, D. Wolf, and A. Arkin, “Control, exploitation and tolerance of intracellular noise,” *Nature*, vol. 420, pp. 231–237, 2002.
- [7] R. Heinrich and S. Schuster, *The Regulation of Cellular Systems*. Chapman and Hall, 1996.
- [8] D. Fell, *Understanding the Control of Metabolism*. Portland Press, 1997.
- [9] J. Elf, P. Löfstedt, and P. Sjöberg, “Problems of high dimension in molecular biology,” in *17th GAMM-Seminar*, Leipzig, 2001, pp. 1–10.
- [10] C. van Gend and U. Kummer, “STODE - automatic stochastic simulation of systems described by differential equations,” in *Proceedings of the 2nd International Conference on Systems Biology*, Yi and Hucka, Eds. Pasadena: Omnipress, 2001, pp. 326–333.
- [11] N. Le Novère and T. Shimizu, “STOCHSIM: modelling of stochastic biomolecular processes,” *Bioinformatics*, vol. 17, p. 575, 2001.
- [12] X.-Q. Xia and M. Wise, “DiMSim: A discrete-event simulator of metabolic networks,” *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 1011–1019, 2003.
- [13] D. Gillespie, “General method for numerically simulating the stochastic time evolution of coupled chemical reactions,” *Journal of Computational Physics*, vol. 22, pp. 403–434, 1976.
- [14] —, “Exact stochastic simulation of coupled chemical reactions,” *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.
- [15] —, “Approximate accelerated stochastic simulation of chemically reacting system,” *Journal of Chemical Physics*, vol. 115, no. 4, pp. 1716–1733, 2001.
- [16] M. Gibson and J. Bruck, “Efficient exact stochastic simulation of chemical systems with many species and many channels,” *J. Phys. Chem. A*, vol. 104, pp. 1876–1888, 2000.
- [17] A. Kierzek, “STOCKS STOchastic Kinetic Simulations of biochemical systems with Gillespie algorithm,” *Bioinformatics*, vol. 18, pp. 470–481, 2002.
- [18] C. Gardiner, *Handbook of Stochastic Models*, 2nd ed. Springer, 1985.
- [19] N. van Kampen, *Stochastic Processes in Physics and Chemistry*. North-Holland, 1992.
- [20] D. Gillespie, “A rigorous derivation of the chemical master equation,” *Physica A*, vol. 188, pp. 404–425, 1992.
- [21] —, *Markov Processes*. Academic Press, 1992.
- [22] S. Schnell and C. Mendoza, “Closed form solution for time-dependent enzyme kinetics,” *J. theor. Biol.*, vol. 187, pp. 207–212, 1997.
- [23] L. Segel, *Modeling dynamic phenomena in molecular and cellular biology*. Cambridge University Press, 1984.
- [24] R. Heinrich, B. Neel, and T. Rapoport, “Mathematical models of protein kinase signal transduction,” *Molecular Cell*, vol. 9, pp. 957–970, May 2002.
- [25] C. Rao and A. Arkin, “Stochastic chemical kinetics and the quasi-steady-state assumption: Applications to the Gillespie algorithm,” *Journal of Chemical Physics*, vol. 118, pp. 4999–5010, March 2003.

Olaf Wolkenhauer received degrees in control engineering from the University of Applied Sciences, Hamburg, Germany and the University of Portsmouth, U.K. in 1994. He received his Ph.D. for work on possibility theory applied to data analysis in 1997 from the University of Manchester Institute of Science and Technology (UMIST) in Manchester, U.K. From 1997 to 2000 he was a lecturer in the Control Systems Centre, UMIST and held a joint senior lectureship between the Department of Biomolecular Sciences and the Department of Electrical Engineering & Electronics until 2003. He is now full professor (Chair in Systems Biology and Bioinformatics) in the Department of Computer Science at the University of Rostock, Germany. He is also a visiting professor in the Department of Mathematics at UMIST. His research interest is in statistical data analysis and mathematical modelling with applications to molecular and cell biology. The research focus is on dynamic modelling of gene expression and cell signalling.

Mukhtar Ullah received his BSc in Electrical Engineering from N.W.F.P University of Engineering and Technology Peshawar, N-W.F.P, Pakistan in 1999 and his MSc. in Advanced Control and Systems Engineering from the Department of Electrical Engineering and Electronics, at the University of Manchester Institute of Science and Technology (UMIST), Manchester, U.K. in 2002. He is now a research assistant in the Systems Biology and Bioinformatics group in the Department of Computer Science, University of Rostock, Germany. His research is concerned with the identification of dynamic models of gene/protein interactions from gene expression time series.

Walter Kolch studied Medicine at the University of Vienna, Austria. After doing postdoctoral work at the National Cancer Institute, Frederick, Maryland, USA, he became Head of the Molecular Biology Department at Goedecke-Parke Davies, Freiburg, Germany, and subsequently a Group Leader at the GSF Research Centre, Munich, Germany. Currently he is a Senior Group Leader at the Beatson Institute for Cancer Research, Glasgow, and Professor for Molecular Cell Biology and Director of the Sir Henry Wellcome Functional Genomics Facility at the University of Glasgow. He is interested in the molecular mechanisms of signal transduction with a research focus on MAPK and the role of protein interactions in the regulation of signalling pathways. A current aim is to understand signalling networks by using proteomics, real time imaging and mathematical modeling.

Kwang-Hyun Cho received his B.S. summa cum laude, M.S. summa cum laude, and Ph.D. in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST) in 1993, 1995, and 1998, respectively. He joined the School of Electrical Engineering at the University of Ulsan, Korea in 1999, where he is an Assistant Professor. From 2002 to 2003, he worked at the Control Systems Centre, Department of Electrical Engineering & Electronics at the University of Manchester Institute of Science and Technology (UMIST), U.K. as a visiting research fellow. His research interests cover the areas of systems science and control engineering including systems biology, bioinformatics, discrete event systems, nonlinear dynamics, and hybrid systems. The focus has been on applications in biotechnology and biological sciences, in particular, mathematical modeling and simulation of genetic- and signal transduction pathways, and DNA microarray data analysis.