A tutorial on cellular stochasticity and Gillespie's algorithm (DRAFT)

F. Hayot[1] and C. Jayaprakash[2]
1. Department of Neurology, Mount Sinai School of Medicine, New York, NY, 10029
2. Department of Physics, The Ohio State University, Columbus, OH 43210
email addresses: fernand.hayot@mssm.edu, jay@mps.ohio-state.edu

April 18, 2006

1 Introduction

There is a plethora of ways to model biological systems, depending on size, detail required and questions asked. One method consists in writing down a collection of coupled ordinary differential chemical equations, where each equation describes a number of reactions. The variables are the time dependent concentrations of participating molecules, and the parameters are reaction rate constants. In this approach, where dependences on spatial location are neglected, except for the consideration of different cellular compartments such as cytoplasm or nucleus, reactions are assumed to occur homogeneously throughout the compartmental volume. Concentrations are defined for large numbers of molecules, such that when numbers change by one or two units in a reaction, these changes can be treated differentially. Moreover when the number of molecules is large any two reactions can take place at the same time. The system of ordinary differential equations for concentrations thus represents a collections of reactions occurring simultaneously all through the reaction volume.

The simplifying features of this approach break down when the numbers of molecules become small, and reactions now occur in some random order rather than simultaneously. One then needs to adopt a new language for the description of the system: probabilities for the state of the system defined by the number of molecules of each type at a given time, replace the differentiable concentrations. These probabilities evolve in time as such or such reaction takes place randomly among all possible reactions. Gillespie's algorithm (Gillespie, 1977), which is the subject of this tutorial, is a way of implementing consistently this probabilistic description of a biological system. The probabilistic description by its very nature applies to single cells. The connection with molecular concentrations appears when, in the probabilistic formalism, averages are taken over many cells. These averages satisfy the same equations as concentrations. Thus the behavior of concentrations can be interpreted as that of a population average, provided that fluctuations around the average are small.

Noise can play an important role in single cell behavior (for a recent review see Rao et al., 2002), which is hidden when only population average is measured. An example is the all-or-none single cell response observed by Ferrell and Machleder (1998) in Xenopus oocytes under progesterone stimulation, whereas average response is graded. Another example is based on NF-κB oscillations observed both for cell populations (Hoffmann et al. 2002) and single cells (Nelson et al., 2004) in cells stimulated by TNFα. It is illustrated in Figure 1, which shows how average oscillatory behavior (the full line) is a poor description of single cell oscillations (broken lines) that differ in both amplitude and phase (Hayot and Jayaprakash, 2006) because of fluctuations in the signaling cascade components set into motion inside the cell by TNFα.

The tutorial is based on the study of a very simple model of gene transcription and translation (Thattai and Van Oudenaarden, 2001). Section 2 describes the model, its implementation as a set of ordinary differential equations, and proposes a MATLAB program to investigate it numerically. Section 3 contains a series of remarks on Poisson processes, and provides an introduction to the probabilistic approach to biological reactions based on the Master equation. In section 4 the Master equation for the Thattai-Van Oudenaarden model (2001) is established. Section 5 contains a general description of Gillespie's algorithm, and its application to the Thattai-Van Oudenaarden model, which for the purpose of numerical simulations is written down in a MATLAB program. In section 6, an extended version of the model allows the distinction between intrinsic and extrinsic fluctuations. Section 7 contains a few remarks on the numerical efficiency of Gillespie's algorithm.

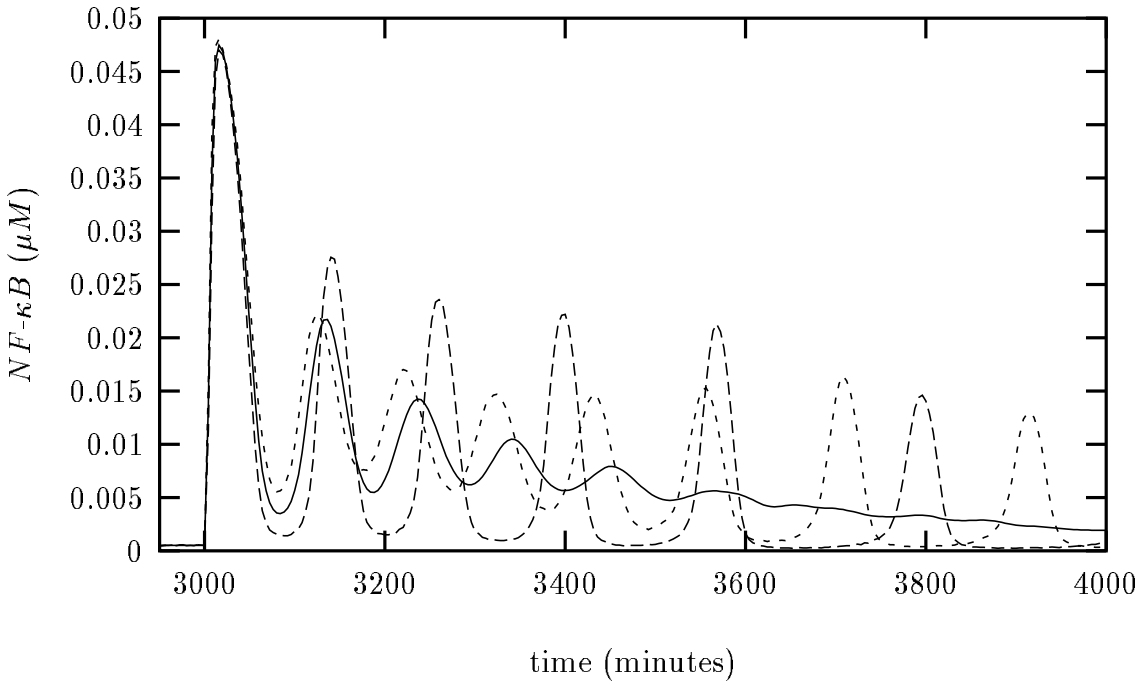

Figure 1:

2 A simple model of cellular transcription and translation (Thattai and Van Oudenaarden, 2001). Population or concentration behavior.

Consider the following set of chemical reactions for transcription of a gene D into mRNA, and subsequent translation of the latter into proteins:

$$D \stackrel{k_1}{\rightarrow} D + M \quad (1)$$

$$M \stackrel{k_2}{\rightarrow} M + P \quad (2)$$

with reaction rates k_1 and k_2, respectively. The mRNA, called M, as well as the protein, denoted by P, are degraded through the reactions

$$M \stackrel{k_3}{\rightarrow} \emptyset \quad (3)$$

$$P \stackrel{k_4}{\rightarrow} \emptyset \quad (4)$$

with k_3 and k_4 the corresponding rate constants, or equivalently with respective half-lives $\tau_3 = \frac{ln2}{k_3}$ and $\tau_4 = \frac{ln2}{k_4}$.

The chemical rate equations for the concentrations [M] and [P] of mRNA M and protein P are:

$$d[M]/dt = k_1[D] - k_3[M] \quad (5)$$

$$d[P]/dt = k_2[M] - k_4[P] \quad (6)$$

This is a simple system, linear in the concentrations, such that in steady state $[M] = \frac{k_1}{k_3}[D], [P] = \frac{k_2}{k_4}[M]$. An important parameter is b, the average number of proteins produced in a mRNA lifetime, which is $b = k_2/k_3$. b is called burst factor; the distribution of proteins in a burst has been recently measured (Cai et al., 2006; Yu et al., 2006).

Programs in MATLAB: beginTVO.m and TVO.m

Instead of concentrations we use number of molecules (system is linear); we take the number of D's equal to 1 (haploid cell).

Simulations:

Figure 2

We plot the number of proteins P as a function of time until steady state is reached. Start from the initial state $D = 1, P = M = 0$, and use parameter values $k_1 = 0.01\, sec^{-1}, k_3 = 0.00577\, sec^{-1} (\tau_3 = 2\, min), k_4 = 0.0001925\, sec^{-1} (\tau_4 = 1\, hour), b = 20$.

Notice that in order to reach steady state from the given initial state, one must run in time of the order of 10 times the longest time scale of the system, which here is the protein lifetime. At 6 times the protein lifetime, one is still a few percent away from the calculated steady state value.

The predicted steady state value are (see expressions following equation 6) $M = 1.73$, $P = 1039$. Note that because M is very small ($M = 1.73$), the interpretation is that on

average over a population M=1.73 (see the second paragraph of the introduction). For the model considered, the deterministic formalism is not appropriate for the description of cellular mRNA content.

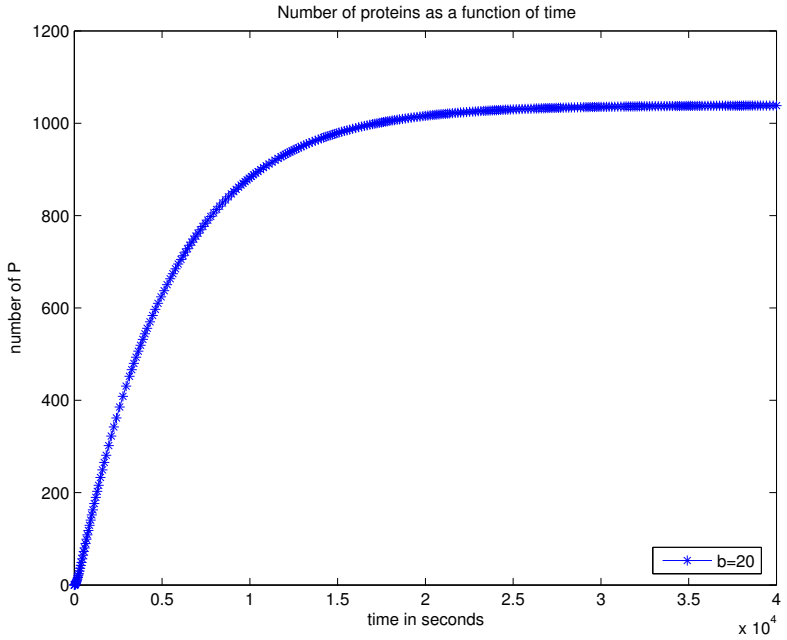

Figure 2:

Figure 3

Do the same numerical experiment for $b = 2$ by dividing the value of k_2 (rate of translation of mRNA) by 10. The steady state value of M is the same as previously, whereas that of P is divided by 10.

3 Notes on Poisson distribution

The expression of a Poisson probability distribution $P(n)$ for n events is

$$P(n) = \frac{\bar{n}^n \exp -\bar{n}}{n!} \tag{7}$$

Properties:

- $\sum_n P(n) = 1$
- $\sum_n nP(n) = \bar{n} =< n >$, the **average** number of events
- $\sum_n n^2 P(n) = \bar{n}^2 + \bar{n} =< n^2 >$

Thus for a Poisson distribution the **variance** $\sigma^2 =< n^2 > - < n >^2 =< n >$, and the **coefficient of variation** $C_v = \sigma / < n >= 1/\sqrt{< n >}$. One also finds in the literature the **Fano factor** $F = \sigma^2 / < n >$, which is equal to 1 for a Poisson distribution.

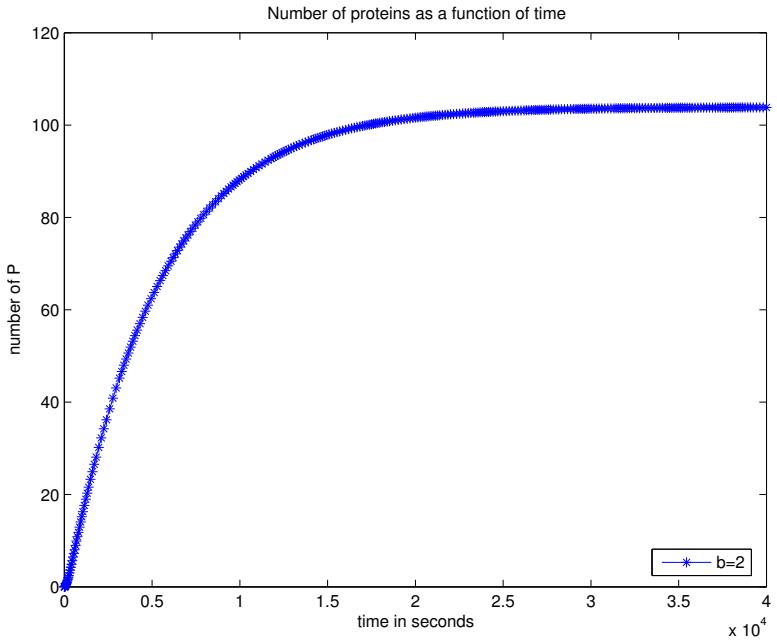

Figure 3:

3.1 Remarks

1. **Homogeneous and inhomogeneous Poisson process**

If the Poisson process takes place at a rate r for a time T, then in the above expression $\bar{n} = r\,T$. For a constant rate the Poisson process is called homogeneous, for a time dependent rate inhomogeneous. One then has

$$P(n) = \frac{(rT)^n \exp -rT}{n!} \tag{8}$$

2. **Numerical implementation of a Poisson process**.

For both homogeneous and inhomogeneous Poisson processes, and for a choice of time step Δt such that $r\,\Delta t < 1$, one draws at each time step a uniformly distributed random number x_{rand} between 0 and 1. If $r\,\Delta t > x_{rand}$ an event takes place. This method is based on the fact that for sufficiently small Δt, $P(1) = r\,\Delta t$, and $P(0) = 1 - r\,\Delta t$.

3. **Time distribution between successive events for a constant rate Poisson process**.

Suppose an event occurred at time t. What is the probability $P(\tau)$ that the next one takes place between $t + \tau$ and $t + \tau + d\tau$? One can decompose the probability in the following way:

P(τ)=(probability that no event takes place between t and $t + \tau$) $\times$ (probability that one event occurs between $t + \tau$ and $t + \tau + d\tau$)= $\exp(-r\,\tau)\,r\,d\tau$, where r is the constant rate of the Poisson process.

The probability density function for successive time intervals is therefore a decaying exponential

$$p(\tau) = r\,\exp(-r\,\tau) \tag{9}$$

Properties:

- $\int_0^\infty d\tau\, p(\tau) = 1$

- $< \tau >= 1/r$

- $< \tau^2 >= 2/r^2$

- $\sigma^2 =< \tau^2 > - < \tau >^2= 1/r^2$, and therefore the coefficient of variation $C_v = \sigma / < \tau >= 1$

4. **Numerical implementation of an exponential probability density distribution**

If x is a random number uniformly distributed between 0 and 1 ($p(x) = 1$) then $y = -(1/r)\ln x$ is randomly distributed according to $p(y) = r\exp(-ry)$, for y varying from 0 to ∞.

Proof: since y is a monotonic function in x, $p(y)dy = p(x)dx = dx$. Thus $p(y) = |dx/dy| = r\exp(-ry)$.

5. **Master equation for "one-step" (or "birth-and-death") processes (Van Kampen, 1992)**

Consider a continuous time stochastic process where one takes steps on the integers, with permissible jumps between adjacent integers only, and (time independent) transition rates r_n for $n \to n-1$ and g_n for $n \to n+1$. (Examples: absorption and emission of particles, arrival and departure of customers)

Probability $p_n(t + \Delta t)$ to be at site n at time $t + \Delta t$ is given by:

$$p_n(t + \Delta t) = r_{n+1}\Delta t\, p_{n+1}(t) + g_{n-1}\Delta t\, p_{n-1}(t) + (1 - r_n\Delta t - g_n\Delta t)p_n(t) \tag{10}$$

The first two terms on the right-hand side describe jumping to position n from $n+1$ and $n-1$ respectively, the last term expresses the probability that during a (sufficiently small) time Δt, the jumper, at n at time t, remains at position n from t to $t + \Delta t$.

For $\Delta t \to 0$ the preceding equation takes the form

$$\partial p_n/\partial t = r_{n+1}p_{n+1} + g_{n-1}p_{n-1} - (r_n + g_n)p_n \tag{11}$$

This is the Master equation of the process, an equation for the time evolution of probability distributions. A solution requires the specification of initial time values of the probabilities.

EXAMPLES:

1.**Random walk in continuous time**

Here $r_n = p$, $g_n = q$, with $p + q = 1$

2. **Homogeneous Poisson process**

Here $r_n = 0$, $g_n = r$ with initial condition $p_n(0) = \delta_{n,0}$.

From comparison with equation (11), the Master equation is

$$\partial p_n/\partial t = r(p_{n-1} - p_n) \tag{12}$$

Solution:

the system of equations (12) can be solved recursively starting with the equation for p_0, namely $\partial p_0/\partial t = -r\, p_0$ ($p_{-1}(t) = 0$), and then proceeding to the solution for p_1, and so on, ending up

with $p_n(t) = (rt)^n \exp(-rt)/n!$, the probability of being at position n at time t.

Above, for a homogeneous Poisson process (see equation 8), we described the same probability as the probability of having n events in time t. If we replace "events" by "molecules" and interpret transition rates r_n and g_n as chemical rate constants (multiplied by the number n of molecules) for reactions describing respectively transitions from a state with n molecules to one with $n-1$, and from a state of n molecules to one with $n+1$, the stochastic "one step process" Master equation (for space and time independent rate constants) becomes an example and the paradigm for dealing with stochasticity in chemical reactions.

A case in point is provided by the following example of production and decay of mRNA taken from the Thattai-Van Oudenaarden model (see section 2).

3. **"Birth-and-death" process": production and decay of mRNA**

Consider equations (1) and (3) which describe production and decay of mRNA M. Let $p_n(t)$ be the probability of having n molecules of mRNA at time t. The corresponding Master equation (see equation 11) is (for $n_D = 1$)

$$\partial p_n/\partial t = k_1[p_{n-1} - p_n] + k_3[(n+1)p_{n+1} - np_n]$$

The steady state solution, which satisfies equation $(n+1)p_{n+1} = np_n + k_1/k_3[p_n - p_{n-1}]$ can be found recursively, such that

$$p_1 = k_1/k_3\, p_0;\; p_2 = 1/2(k_1/k_3)^2\, p_0;\; p_n = 1/n!(k_1/k_3)^n\, p_0$$

with $p_{-1} = 0$.

Normalization of this probability function then gives $p_0 = \exp(-k_1/k_3)$.

The steady state distribution of M thus follows a Poisson distribution, with average steady state value equal to k_1/k_3 (see equation 5).

4. **Generating function**

Before deriving the full Master equation for the Thattai-Van Oudenaarden model in the next section, we show how the homogeneous Poisson process Master equation (12) can be solved by a general method, that of generating function. The generating function for the $p_n(t)$'s is defined as

$$F(z,t) = \sum_n z^n p_n(t) \tag{13}$$

such that

$F(1,t) = 1, (\partial F/\partial z)_{z=1} =< n(t) >, (\partial^2 F/\partial z^2)_{z=1} =< n(t)(n(t)-1) >$

We now multiply equation (12) by z^n and sum over n. We obtain the following equation satisfied by the generating function of a homogeneous Poisson process of constant rate r

$$\partial F(z,t)/\partial t = r(z-1)F(z,t) \tag{14}$$

with initial condition $F(z,0) = 1$ corresponding to $p_n(0) = \delta_{n,0}$.

The solution of (14) is

$F(z,t) = \exp[r(z-1)t] = \exp(-rt)\sum_n (rtz)^n/n!$

from which one recovers the usual expression for $p_n(t)$, namely expression (8).

4 Master equation for the Thattai-Van Oudenaarden model

The reactions of the model, given in section 2, are

$$\begin{aligned} D &\xrightarrow{k_1} D + M \\ M &\xrightarrow{k_2} M + P \\ M &\xrightarrow{k_3} \emptyset \\ P &\xrightarrow{k_4} \emptyset \end{aligned}$$

We have already derived the Master equation for the two reactions involving M only (see example 3 in 3.1.5). The relevant probability here is $P(n_P, n_M, t)$, the probability of having at time t in the volume considered n_P proteins P and n_M mRNA M. The procedure for deriving the Master equation satisfied by P is the same as that illustrated above (section 3.1.5.) for "one-step" processes (equation (10)). One calculates $P(n_P, n_M, t + \Delta t)$ ($\Delta t \to 0$) from the state of the system at time t: there are contributions from each reaction, as well as from the situation where, during time Δt, the state of the system does not change. The result is:

$$\begin{aligned} \partial P(n_P, n_M, t)/\partial t &= k_2 n_M [P(n_P - 1, n_M, t) - P(n_P, n_M, t)] \\ &+ k_3[(n_M + 1)P(n_P, n_M + 1, t) - n_M P(n_P, n_M, t)] \\ &+ k_1 n_D [P(n_P, n_M - 1, t) - P(n_P, n_M, t)] \\ &+ k_4[(n_P + 1)P(n_P + 1, n_M, t) - n_P P(n_P, n_M, t)] \end{aligned} \tag{15}$$

This is the Master equation for the Thattai-Van Oudenaarden model (2001). We will put $n_D = 1$ (n_D is the number of DNA molecules) from now on. This equation is relatively simple, because the number of reactions is small, and reactions are linear in the components. A result of the latter is that reaction constants are simply rates, without any volume dependence. We will discuss more general cases later when describing Gillespie's algorithm. Much can be learned about first and second moments by multiplying both sides of the Master equation by n_P, or n_P^2 and similarly for n_M, and summing over all n_P and n_M to obtain $< n_P >$, or $< n_P^2 >$, and so on. The angular brackets correspond to the average over a population of cells. One can also obtain first and second order (or higher) moments from the generating function $F(z_1, z_2, t) = \sum_{n_P, n_M} z_1^{n_P} z_2^{n_M} P(n_P, n_M, t)$, which here satisfies the equation

$$\partial F/\partial t = (z_1 - 1)(k_2 z_2 \partial F/\partial z_2 - k_4 \partial F/\partial z_1) + (z_2 - 1)(-k_3 \partial F/\partial z_2 + k_1 F) \tag{16}$$

For the first order moments $< n_M >$ and $< n_M >$, obtained respectively from $(\partial F/\partial z_2)_{z_1 = z_2 = 1}$ and $(\partial F/\partial z_1)_{z_1 = z_2 = 1}$ we find

$$\partial < n_M > /\partial t = -k_3 < n_M > + k_1 \tag{17}$$

$$\partial < n_P > /\partial t = -k_4 < n_P > + k_2 < n_M > \tag{18}$$

These equations for population averages are similar to the concentration equations (cf. equations 5 and 6) derived previously. When fluctuations around averages are small, concentrations

represent averages over cell populations divided by cell volume.
In all cases there are fluctuations embodied in the second and higher moments. At steady state one finds the following for the Thattai-Van Oudenaarden model
$\sigma_M^2 =< n_M^2 > - < n_M >^2 =< n_M >$
$\sigma_P^2 / < n_P >^2 = \frac{1}{<n_P>}[1 + k_2/(k_3 + k_4)]$
The stochastic behavior of mRNA is Poisson. It corresponds to a "birth-and-death" process, as we have seen before (see 3.1.5). As to protein number fluctuations, there is an additional term besides the Poisson term, corresponding to the fact that mRNA, from which protein is translated, is itself stochastic. Often mRNA lifetimes (of the order of minutes) are much smaller than protein lifetimes (of the order of hours): thus $k_4/k_3 << 1$ and one has simply
$\sigma_P^2 / < n_P >^2 \simeq \frac{1+k_2/k_3}{<n_P>} = \frac{1+b}{<n_P>}$,
where b is the average number of proteins produced in a mRNA lifetime, as defined in section 2.

5 Gillespie's algorithm (Gillespie, 1977)

Generally biological systems are much more complex than the Thattai- Van Oudenaarden model (2001). The number of reactions in a single cell can be in the tens and larger; many reactions such as dimerization or the binding of an enzyme to its substrate have nonlinear components. Though one can in principle write down the Master equation, or the partial differential equation satisfied by the generating function, these are unwieldy, too complicated to be solved by means other than numerical. Fortunately there is the straightforward numerical algorithm developed by Gillespie (1977), which he showed to be equivalent to solving the Master equation of a system of chemical reactions in a well stirred container. The crux of the algorithm is the drawing of two random numbers at each time step, one to determine after how much time the next reaction will take place, the second one to choose which one of the reactions will occur.

Suppose there are $\mu = 1, 2, ...$ reactions. We consider reactions with at most two species (the probability of three species reacting at the same spot and time is considered negligible). The quantity characterizing each reaction is the probability $a_\mu(t)dt$ that given the state of the system at time t, reaction μ will occur in volume V in (t, t+dt). $a_\mu(t)$ is the product of two parts: the reaction rate c_μ for a particular reaction μ, which is related to the chemical rate constants for that reaction, and the number of possible reactions μ in volume V. For example, for reaction

$$P_1 + P_2 \rightarrow Z \tag{19}$$

where Z is the heterodimer formed of P_1 and P_2, one has
$a_\mu(t) = c_\mu P_1 P_2$
If P_2 is identical to P_1, then

$$a_\mu(t) = c_\mu P_1(P_1 - 1)/2 \tag{20}$$

where $P_1(P_1 - 1)/2$ is the number of distinct pairs of P_1.

Remark: relation between c_μ's and chemical constants k

By definition the c_μ's are rates with dimension of an inverse time. When for a given reaction the chemical constant has the dimension of an inverse time, as is the case of the reactions of the Thattai-Van Oudenaarden model, the c_μ is simply equal to the corresponding k. However for reaction (19), namely

$$P_1 + P_2 \xrightarrow{k} Z \tag{21}$$

which in a deterministic approach with concentrations reads

$$d[Z]/dt = k[P_1][P_2] \tag{22}$$

the chemical constant k has dimension of volume divided by time. Therefore here

$$c_\mu = k/V \tag{23}$$

where V represents the volume of the region in which the reaction takes place.

If P_1 and P_2 are the same as in (20) then $c_\mu = 2k/V$.

In cases like these chemical rate constants are expressed in inverse molars and inverse seconds. It is useful to note that 1 nM corresponds to 1 particle in a volume of 1.6 μ^3.

Implementation of Gillespie's algorithm (Gillespie, 1977)

Suppose the system is known at time t, which means the number of molecules of each type is known, and consequently the quantities $a_\mu(t)$ are known for each reaction. Call $a_0(t)$ the sum of all $a_\mu(t)$.

Then do the following steps:

1. find the time τ after t at which the next reaction will take place, by drawing a random number from an exponential probability density function of rate a_0 ($p(\tau) = a_0 \exp(-a_0\tau)$). The reasoning is the same as in point 3 of section 3.1.

2. choose now at random the reaction which will occur at time $t+\tau$. Draw a random number from a uniform distribution between 0 and 1. If that number falls between 0 and a_1/a_0 reaction 1 is chosen, between a_1/a_0 and $(a_1 + a_2)/a_0$ reaction 2 is chosen and so on.

3. the occurrence of the chosen reaction at time $t + \tau$ changes the numbers for molecules involved in the reaction, for example for the forward reaction of (19) $P_1 \to P_1 - 1$, $P_2 \to P_2 - 1$, and $Z \to Z + 1$. Thus the values of the a_μ which depend on any of these numbers change. One then goes back to point 1 of the algorithmic implementation with a new distribution of molecules at time $t + \tau$. The process is reiterated for as long as one wishes to follow the evolution of the system.

Program in MATLAB: tattai.m

Figure 4

For comparison with figure 2, we run the Gillespie simulation with 200 cells, with the same parameter values as in figure 2. The results are in figure 4 where average protein number as a function of time and other quantities are shown, in particular the comparison of average protein number and individual cell behavior for three randomly chosen cells. Each cell starts off from the same initial state, and ends up after a time chosen long enough to reach steady state, with

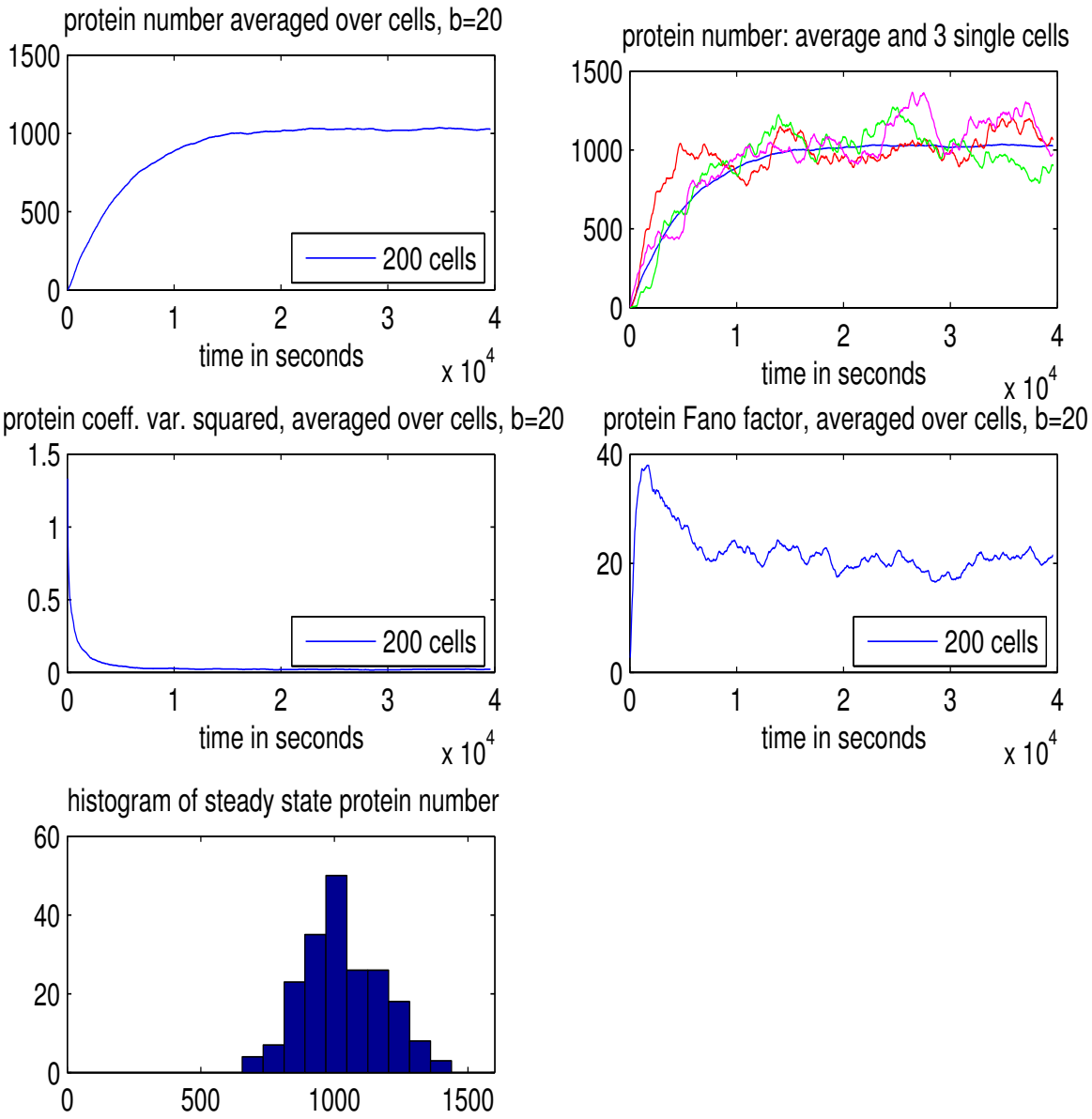
protein number averaged over cells, b=20
1500
1000
500
0
200 cells
0
1
2
3
4
time in seconds
x 10^4
protein number: average and 3 single cells
1500
1000
500
0
0
1
2
3
4
time in seconds
x 10^4
protein coeff. var. squared, averaged over cells, b=20
1.5
1
0.5
0
200 cells
0
1
2
3
4
time in seconds
x 10^4
protein Fano factor, averaged over cells, b=20
40
20
0
200 cells
0
1
2
3
4
time in seconds
x 10^4
histogram of steady state protein number
60
40
20
0
0
500
1000
1500

Figure 4:

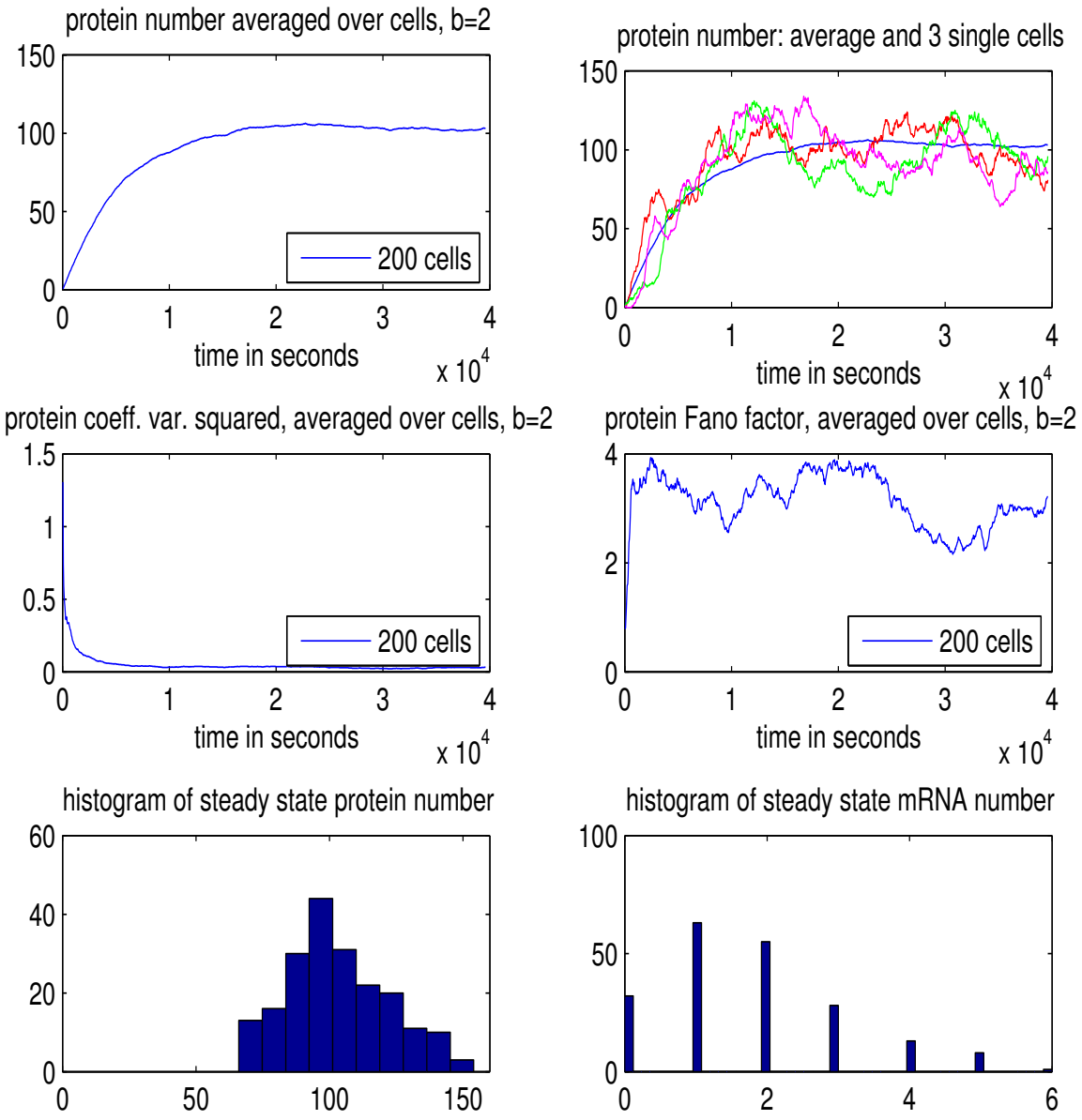
protein number averaged over cells, b=2
150
100
50
0
200 cells
0
1
2
3
4
time in seconds
x 10^4
protein number: average and 3 single cells
150
100
50
0
0
1
2
3
4
time in seconds
x 10^4
protein coeff. var. squared, averaged over cells, b=2
1.5
1
0.5
0
200 cells
0
1
2
3
4
time in seconds
x 10^4
protein Fano factor, averaged over cells, b=2
4
2
0
200 cells
0
1
2
3
4
time in seconds
x 10^4
histogram of steady state protein number
60
40
20
0
0
50
100
150
histogram of steady state mRNA number
100
50
0
0
2
4
6

Figure 5:

a different number of proteins, due to internal fluctuations. The histogram based on the protein number in each cell at the end of the run (40,000 seconds) shows a wide distribution between cells with as few as 800 proteins to cells with as many as 1400 proteins. The program also prints out the average number of proteins in steady state ($< n_P >= 1039$) for the run of figure 4, the Fano factor $\sigma^2/ < P >= 21.42$, standard deviation $\sigma_P = 148.9$), and coefficient of variation $C_v = 0.1439$. Compare these values with the analytical results of section 2. The statistical error on the average is equal to $\sigma_P/\sqrt{N_c}$, where N_c is the number of cells. Here this error is about 10. As one averages over more and more cells the statistical error decreases and one obtains results for the average number of proteins closer to the deterministic value of figure 2.

Figure 5

This figure corresponds to figure 3 where $b = 2$ and therefore the average number of proteins gets divided by 10 as compared to the case of figure 3. For the run of figure 5, $< n_P >= 103$, Fano factor=3.48, $\sigma_P = 18.96$ and $C_v = 0.1835$. The coefficient of variation here is 25% higher than when the number of proteins is ten times larger (figure 4), signaling increased fluctuations as the number of proteins decreases. The bottom right subfigure gives the mRNA histogram, for which the average value of $< M >= 1.73$ (see discussion of figure 2) is a poor description of what the actual number is in a single cell.

6 Intrinsic and extrinsic fluctuations

Rather than equation (1) of the Thattai-Van Oudenaarden model, namely

$$D \xrightarrow{k_1} D + M \tag{24}$$

consider the following two equations:

$$D + R \underset{k'}{\overset{k}{\rightleftharpoons}} D^* \tag{25}$$

$$D^* \xrightarrow{k_n} D + R + M \tag{26}$$

The remaining reactions of mRNA decay and protein P production and decay remain the same, as given in section 2, equations (2)-(4).

In the new set of reactions, the gene promoter region is bound by a polymerase or/and a transcription factor represented by protein R, giving complex D^*. One complex D^* is formed, it can transcribe mRNA.

Fluctuations in the previous system of reactions, considered as an entity by itself, are intrinsic to that system. If now the amount of R fluctuates between cells, then the corresponding fluctuations are extrinsic fluctuations, since R can be considered external to the system. Clearly, the division between what are intrinsic and what are extrinsic fluctuations is somewhat arbitrary. In practical cases however is is mostly clear where to draw the line.

Let us now investigate our new system. Let us assume that the reactions of binding and unbinding of R are fast fast compared to all others (an assumption which is often justified), so

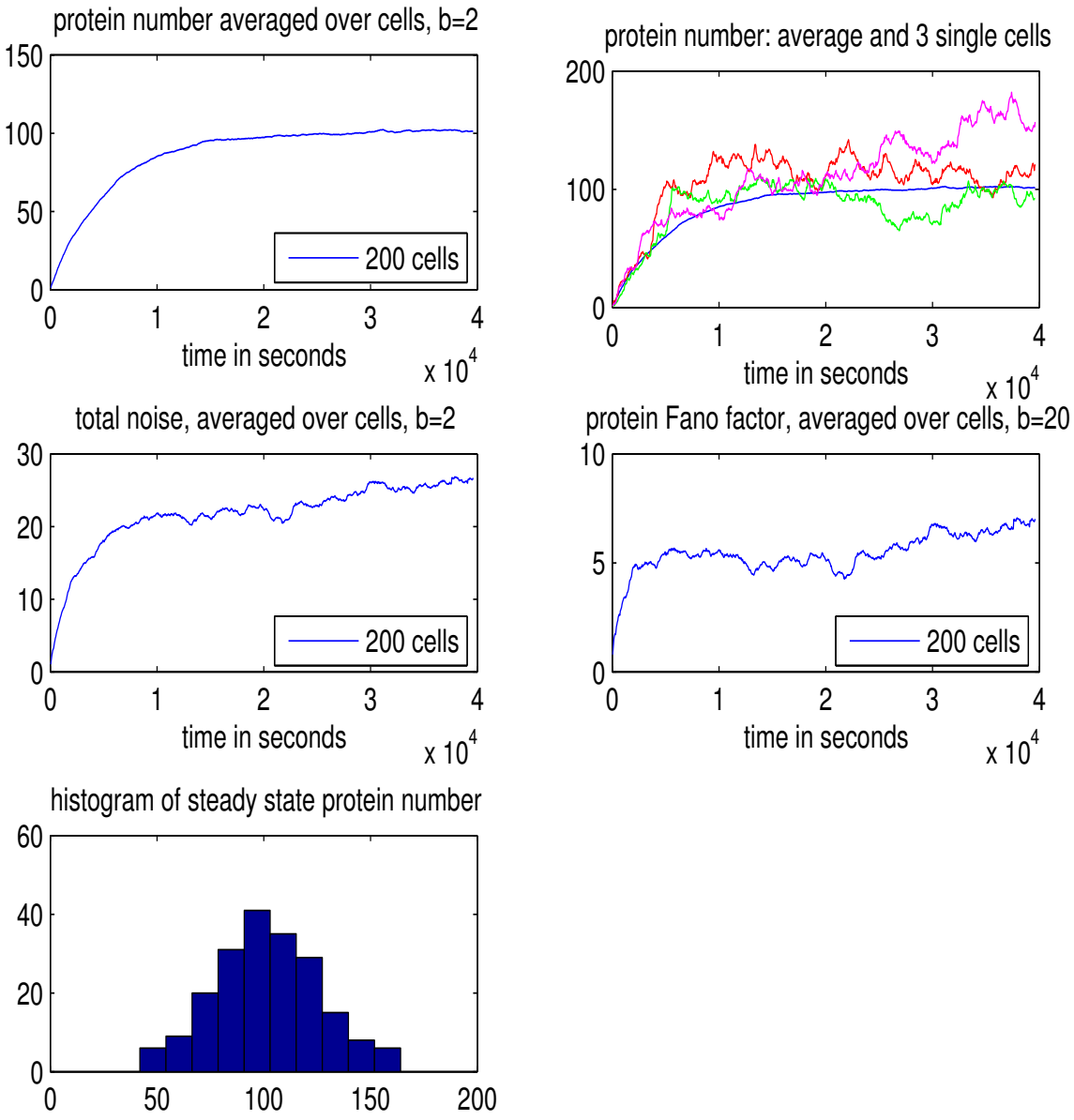
protein number averaged over cells, b=2
150
100
50
0
0
1
2
3
4
time in seconds
x 10^4
200 cells
protein number: average and 3 single cells
200
100
0
0
1
2
3
4
time in seconds
x 10^4
total noise, averaged over cells, b=2
30
20
10
0
0
1
2
3
4
time in seconds
x 10^4
200 cells
protein Fano factor, averaged over cells, b=20
10
5
0
0
1
2
3
4
time in seconds
x 10^4
200 cells
histogram of steady state protein number
60
40
20
0
0
50
100
150
200

Figure 6:

that on the longer time scales considered this reaction is at steady state. One then obtains from the equation for D^* the relation

$$k[R][D] = k'[D^*] \tag{27}$$

which because $[D] + [D^*] = 1/volume$, leads to the following expression for D^*

$$[D^*]\,(volume) = [R]/(K + [R]) \tag{28}$$

The two sides of this input $[R]$- output $[D^*]$ equation are dimensionless. $K = k'/k$ is called a Michaelis-Menten constant, which is such that when $[R] = K$, the output reaches half its maximum.

The new concentration equation for M is

$$d[M]/dt = k_n[D^*] - k_3[M] = k_n[R]/volume(K + [R]) - k_3[M] \tag{29}$$

This equation is similar to equation (5) (section 2) with the additional term $[R]/(K + [R])$ in the production of M. The assumption of fast binding of R and the conservation of the sum of D and D^* concentrations have led to an effective equation with a Michaelis-Menten term. It has been shown (Bundschuh et al., 2003) that such a form does not spoil fluctuations calculated with the Gillespie algorithm. We now make the following assumptions:
- proteins R constitute a reservoir, and therefore their dynamics can be neglected
- the number of R varies from cell to cell according to a gaussian, with average value chosen equal to the Michaelis-Menten constant K multiplied by the volume. Thus, if for this average value of R we take $k_n/2 = k_1$, the new enlarged model behaves on average the same way as the original model.

Program in MATLAB: tattaiR.m

Figure 6

This figure highlights the increase of protein P fluctuations, when external polymerase fluctuations are present. All cells start off with the same initial number of constituents, except for R whose value, for any cell, is drawn from a gaussian of given average. In program *tattaiR.m* the average is 30 and the gaussian standard deviation is chosen equal to 10. The parameters are the same as for figure 5, with in particular b=2. For the run of figure 6, the Fano factor=6.3, a doubling compared with that of figure 5. The coefficient of variation $C_v = 0.25$ has a corresponding increase of 40%. The probability distribution of P's is correspondingly wider.

Total and intrinsic noise

One can measure total noise by calculating σ_t, the width of the distribution of protein P, in program *tattaiR.m*. This total noise is shown in figure 6 as function of time. It is the total noise, because cell to cell fluctuations arise both internally, as in the original Thattai-Van Oudenaarden model, and externally because the amount of polymerase fluctuates from cell to cell. The measured total noise is the simultaneous average over both sources of noise

$$\sigma_t^2 = \overline{< P^2 >} - \overline{< P >}^2 \tag{30}$$

The brackets denote averaging over internal noise, the overbar averaging over external noise.

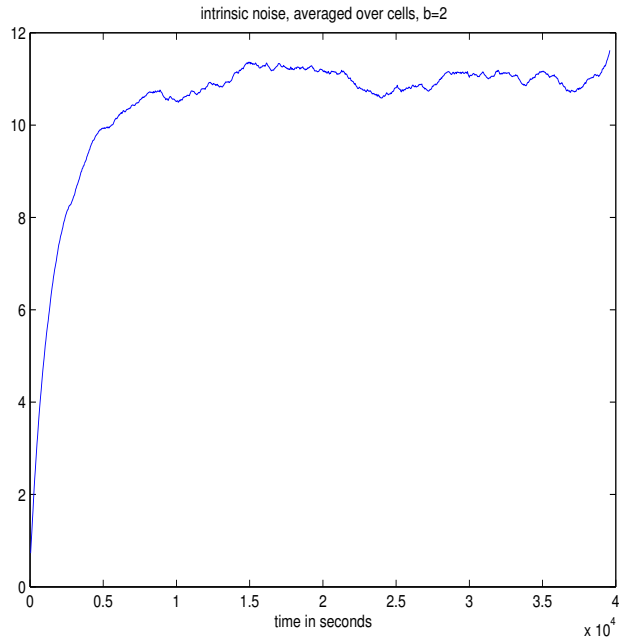

Figure 7:

The intrinsic noise by itself σ_{in} can be calculated (Swain et al., 2002) at each fixed value of external noise source (here the amount of polymerase) over all cells, and then averaged over the probability distribution of external noise (here a gaussian for polymerase). Thus

$$\sigma_{in}^2 = \overline{< P^2 > - < P >^2} \tag{31}$$

Once total and intrinsic noise are known, extrinsic noise σ_{ext} is obtained as

$$\sigma_{ext} = \sqrt{\sigma_t^2 - \sigma_{in}^2} \tag{32}$$

Program in MATLAB: tattaiRin.m

Figure 7

Here is the curve for intrinsic noise, σ_{in}, as a function of time, averaged over 200 cells, which is to be compared with the corresponding curve for σ_t in figure 6. σ_t is larger: the difference between the two curves is extrinsic noise, defined in equation (32).

Experimentally one measures fluctuations in P, which are the result of both intrinsic and extrinsic noise. What experimental procedure can one adopt to separate intrinsic and extrinsic noise contributions? A methodology of using 2 identical promoter regions, coding for different fluorescent proteins, was introduced and applied to E. coli by Elowitz et al. (2002), and similarly used in yeast (Raser and O'Shea, 2004). For human cells infected by a virus, allelic imbalance of interferon-β mRNA, allows a discussion and discrimination between intrinsic and extrinsic sources of noise (Hu et al., 2006) in interferon-β production.

7 Efficiency of Gillespie's algorithm

Gillespie's algorithm, when implemented in FORTRAN or C, leads to very efficient numerical computations for systems of several tens of reactions. The exception occurs when some reactions, such as a dimerization reaction, are very fast on the time scales for which the system is observed, which are typically time scales of the order of the longest time scales of reaction dynamics. In this case of some very large rate constant, coupled with a reasonably large number of molecules, Gillespie's algorithm spends a large fraction of time selecting for updating that very fast reaction. The computation then becomes inefficient. (For a discussion of this issue, remedies and problems, see Bundschuh et al. (2003)).

Several methods have been proposed to accelerate Gillespie's algorithm for large systems of reactions, such as the "tau-leap" stochastic algorithm (Gillespie and Petzold, 2003) and the algorithm of Gibson and Bruck (1999). The first scheme replaces serial updating of the state of the system through individual reactions by a probabilistic updating of many interactions in some given time interval (under certain conditions). In the second scheme, where the problem mentioned above with very fast reactions persists, updating takes place as in the usual Gillespie algorithm, albeit much more efficiently, A software package, called "Dizzy" (Ramsey et al., 2005), is available that implements Gillespie's algorithm and the above algorithmic improvements.

8 References

Bundschuh,R., Hayot, F., and Jayaprakash, C., 2003. Fluctuations and slow variables in genetic networks. Biophys. J. 84, 1606-1615.

Cai, L., Friedman, N., and Xie, X.S., 2006. Stochastic protein expression in individual cells at the single molecule level. Nature 440, 358-362.

Elowitz, M.B., Levine, A.J., Siggia, E.D., and Swain, P.S., 2002. Stochastic gene expression in a single cell. Science 297, 1883-1886.

Ferrell, J.E., and Machleder, E.M., 1998. The biochemical basis of an all-or-none cell fate switch in Xenopus oocytes. Science 280, 895-898.

Gibson, M.A., and Bruck, J., 1999. Efficient exact stochastic simulation of chemical systems with many species and many channels. Caltech Parallel and Distributed Systems Group technical report No 026.

Gillespie, D.T., 1977. Stochastic simulations of coupled chemical reactions. J. Phys. Chem. 81, 2340-2361.

Gillespie, D.T., 2001. Approximate accelerated stochastic stimulation of chemically reacting systems. J. Chem. Phys. 115, 1716-1733.

Gillespie, D.T., and Petzold, L.R., 2003. Improved leap-size selection for accelerated stochastic simulation. J. Chem. Phys. 119, 8229-8234.

Hayot, F., and Jayaprakash,C., 2006. NF-kB oscillations and cell-to-cell variability. J. Theor. Biol. (to appear), available on line at www.sciencedirect.com

Hu, J., Pendleton, A.C., Kumar, M., Ganee, A., Moran, T.M., Hayot, F., Jayaprakash, C., Sealfon, S.C., and Wetmur, J., 2006. Noisy induction of interferon β by viral infection of human dendritic cells. Submitted.

Ramsey, S., Orrell, D., and Bolouri, H., 2005. Dizzy: stochastic simulation of large-scale genetic regulatory networks. J. Bioinf. Comp. Biol. 3(2), 415-436.

Rao, C.V., Wolf, D.M., and Arkin, A.P., 2002. Control, exploitation and tolerance of intracellular noise. Nature 420, 231-237.

Raser, J.M., and O'Shea, E.K., 2004. Control of stochasticity in eukaryotic gene expression. Science 304, 1811-1814.

Swain, P.S., Elowitz, M.B., and Siggia, E.D., 2002. Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Nat. Ac. Sci. 99, 12795-12800.

Thattai, M., and Van Oudenaarden, A., 2001. Intrinsic noise in gene regulatory networks, PNAS 98, 8614-8619

Van Kampen N.G. Stochastic processes in physics and chemistry. North-Holland (1992)

Yu, J., Xiao, J., Ren, X., Lao, K., and Xie, S.X., 2006. Probing gene expression in live cells, one protein molecule at a time. Science 311, 1600-1603.