Single cell experiments and Gillespie's algorithm

Fernand Hayot
Department of Neurology, Mount Sinai School of Medicine, New York, NY, 10029
email: fernand.hayot@mssm.edu

August 18, 2008

1 Introduction[1]

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) 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

[1]This lecture is part of the Yale Summer School in Computational Immunology (August 2008) which is supported by PRIME (NIAID contract HHSN266200500021C). It is based on and overlaps with a tutorial given in April 2006 by C. Jayaprakash and the author at Princeton University as part of the joint educational program of PICASso (NSF IGERT grant DGE-9972930) and PRIME. The MATLAB program referred to was run and studied by the students in a Discussion session. The program is a MATLAB version of a FORTRAN program written by the author.

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) when stimulated by TNFα. Here again average behavior is a poor description of single cell oscillations that differ in both amplitude and phase (Hayot and Jayaprakash, 2006) because of fluctuations in the signaling cascade components set into motion by TNFα.

This lecture is based on the study of a very simple model of gene transcription and translation (Thattai and Van Oudenaarden, 2001). Section 2 contains a number of remarks on measurements in single cell experiments, as well as some results on the Poisson probability distribution. Section 3 describes the Thattai-Van Oudenaarden model (2001) for gene transcription and mRNA translation. In section 4, isolating from the model the reactions of production and degradation of mRNA, we derive the relevant the Master equation and discuss some of its features. This is followed in section 5 by the complete Master Equation of the Thattai-Van Oudenaarden model. Section 6 contains a description of the Gillespie algorithm, which implements the Master Equation, followed in section 7 by some remarks concerning the computational efficiency of the algorithm. Section 6 calls the MATLAB program that runs Gillespie's algorithm for the Thattai-Van Oudenaarden model, and produces data that enableone to appreciate the role of stochasticity in single cell experimnets. The stochasticity observed in the simulations of the Thattai-Van Oudenaarden model is due to the occurrence of small numbers of molecules, mRNAs in particular. This is so-called "small copy number" noise. There exist other possible sources of noise (transcriptional, extrinsic) that can be implemented in Gillespie's algorithem.

2 Single cell experiments

Most cellular measurements involve very many cells, where protein or mRNA concentrations are obtained across a population of lysed cells, through techniques such as Western blot (proteins) or microarrays (mRNA). There is the implicit assumption that the measured average behavior over many cells represents the typical behavior of any single cell, in other words that cell to cell variations, in protein number for example, are small or/and not significant for the understanding of cellular function. The examples mentioned in the Introduction show that this is not always true (see also Rao et al, 2002) and that single cell experiments can reveal new aspects of cellular behavior.

How does one typically proceed in a single cell experiment? Let us suppose that there are N identical cells, prepared under the same conditions, and that the measurement is the determination - at some point in time - of the number of copies of a protein P in each cell. Thus one measures p_1 protein P in cell number 1, p_2 in cell 2, and so on up to p_N in cell N. From these numbers one can calculate two quantities, an average number $< p >$ of proteins per cell, and a standard deviation σ from the average number which indicates the level of variability between

cells:

$$< p > = \frac{1}{N}(p_1 + p_2 + p_3 + ...p_N) \quad (1)$$

$$< p^2 > = \frac{1}{N}({p_1}^2 + {p_2}^2 + {p_3}^2 + ...{p_N}^2) \quad (2)$$

From these two quantities one obtains the variance σ^2

$$\sigma^2 =< p^2 > - < p >^2 \quad (3)$$

and the standard deviation σ

$$\sigma = \sqrt{\sigma^2} \quad (4)$$

If we now assume that the number of cells is large, the calculated quantities can be given a statistical interpretation, where the histogram of cellular responses (the number of cells for which the measured protein number is the same within a small interval around some value) becomes a good approximation to the probability distribution of the number of proteins across cells. This probability distribution can then be compared to known ones, such as gaussian, Poisson, or gamma distribution, which can reveal something of the underlying mechanisms of stochasticity. One can also - by changing the amount of stimulus - vary $< p >$ and σ^2, and plot σ^2 versus $< p >$: if, for instance, one finds that $\sigma^2 =< p >$ the probability distribution is poissonian (see next section). The Poisson distribution is entirely determined by the knowledge of its average value. Generally, however, the full probability distribution (or histogram) contains much more information then what can be gathered from average value and standard deviation.

2.1 Notes on the Poisson probability distribution

The Poisson distribution is particularly important for this lecture, since it characterizes both the birth-and-death process of production and decay of a mRNA, and also plays a crucial role in the implementation of Gillespie's algorithm.

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

$$P(n) = \frac{\bar{n}^n \exp -\bar{n}}{n!} \quad (5)$$

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.

2.2 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. For a *homogeneous* process of rate r one has

$$P(n) = \frac{(rT)^n \exp -rT}{n!} \tag{6}$$

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 $t+\tau+d\tau$), or formally P(τ)= $\exp(-r\,\tau) \times r\,d\tau$, where r is the constant rate of the Poisson process. The two terms in the expression of P(τ) can be read off equation (6).

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

$$p(\tau) = r\,\exp(-r\,\tau) \tag{7}$$

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 $C_v = \sigma / < \tau >= 1$

3 A simple model of cellular transcription and translation (Thattai and Van Oudenaarden, 2001).

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

$$D \xrightarrow{k_1} D + M \tag{8}$$

$$M \xrightarrow{k_2} M + P \tag{9}$$

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

$$M \stackrel{k_3}{\rightarrow} \emptyset \tag{10}$$

$$P \stackrel{k_4}{\rightarrow} \emptyset \tag{11}$$

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:

$$d[M]/dt = k_1[D] - k_3[M] \tag{12}$$

$$d[P]/dt = k_2[M] - k_4[P] \tag{13}$$

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).

In order to simulate the system of reactions (8)-(11), whether deterministically or stochastically, one needs to know the initial amounts (numbers or concentrations) of D, P, M and the reaction rate constants. In the following we will take in the initial state the numbers $D = 1, P = M = 0$, and use parameter values $k_1 = 0.01\, sec^{-1}, k_3 = 0.00577\, sec^{-1} (\tau_3 = 2\, min), k_4 = 0.0001925\, sec^{-1} (\tau_4 = 1\, hour), b = 2$.
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 τ_4, 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 values for the concentrations of P and M are obtained from the chemical rate equations (12) and (13). Steady state means that concentrations no longer vary in time and therefore concentration steady state values are obtained by putting the left-hand sides of equations (12) and (13) equal to zero. In this way one finds (after multiplying by the volume) $M = 1.73$, $P = 104$ (the number of proteins is proportional to b). Note that because the number of M per cell is very small the interpretation is that on average over a population M=1.73 (see the second paragraph of the introduction), which means that one can expect cells to have between 0 and say 4 mRNA (see the distribution in Figure 1). For the model considered, the deterministic formalism gives only an approximate description of cellular mRNA content.

4 Master equation (Van Kampen, 1992)

"Birth-and-death" process: production and decay of mRNA

Consider equations (8) and (10) which describe production and decay ("birth-and-death") of mRNA M. Let $P(n, t)$ be the probability of having n molecules of mRNA at time t. The

corresponding so-called Master equation describes the variation in time of $P(n,t)$ and is (for $n_D = 1$)

$$\partial P(n,t)/\partial t = k_1[P(n-1,t) - P(n,t)] + k_3[(n+1)P(n+1,t) - nP(n,t)] \qquad (14)$$

Where does this equation come from?
Suppose at time t the state of the system, i.e. the number of mRNA molecules, is known. What then is the probability P(n,t+dt) of there being n molecules at time t+dt, where dt is considered sufficiently small that only one of the reactions of production or decay can occur. There are three possibilities:
- there are n-1 molecules at time t, and production takes place in time dt which gives a contribution to P(n,t+dt) equal to $k_1\, dt\, P(n-1,t)$ (we take $n_D = 1$)
- there are n+1 molecules at time t, and decay takes place in time dt which gives a contribution to P(n,t+dt) equal to $k_3\, dt\, (n+1)\, P(n+1,t)$. The factor of (n+1) occurs because the decay can take place from anyone of (n+1) molecules
- no reaction takes place, the probability of which is given by $P(n,t)$ multiplied by $[1 - k_1\, dt - k_3\, dt\, n]$, where the latter term is given by the total probability equal to 1 minus the probability that one of the two reactions takes place.
By putting these three term together and replacing $\frac{P(n,t+dt)-P(n,t)}{dt}$ by the derivative $\partial P(n,t)/\partial t$, one obtains the above Master equation (14).

Its steady state solution, which satisfies the 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 $P(0) = \exp(-k_1/k_3)$, and the expected form of a Poisson distribution for P(n) (see expression (5)), namel $P(n) = \frac{(k_1/k_3)^n \exp(-k_1/k_3)}{n!}$.
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), which is precisely the value obtained directly from solving equation (12) of the deterministic system, an example of the correspondence between deterministic chemical reaction behavior and that of the population avarage of a stochastic formulation. This correspondence becomes clearer still if one multiplies both sides of equation (14) by n and sums over all n ($\sum P(n,t) = 1, \sum nP(n,t) =< n(t) >$), which leads to

$$d < n(t) > /dt = k_1 - k_3 < n(t) \qquad (15)$$

This equation has the same form as (12): thus the average value over all cells $< n(t) >$ follows the same equation as the mRNA concentration of the deterministic equation.

5 Master equation for the Thattai-Van Oudenaarden model

The reactions of the model, given in section 3, 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 section 4). Here the relevant probability 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 4) for the mRNA birth-and-death process. One calculates $P(n_P, n_M, t + dt)$ from knowledge of the state of the system at time t: there are contributions from each reaction, as well as from the situation where, during time dt, the state of the system does not change. The result is:

$$\begin{aligned} \partial P(n_P, n_M, t)/\partial t &= k_2 n_M [P(n_P - 1, n_M, t) - P(n_P, n_M, t)] \\ &+ k_3[(n_M + 1)P(n_P, n_M + 1, t) - n_M P(n_P, n_M, t)] \\ &+ k_1 n_D [P(n_P, n_M - 1, t) - P(n_P, n_M, t)] \\ &+ k_4[(n_P + 1)P(n_P + 1, n_M, t) - n_P P(n_P, n_M, t)] \end{aligned} \quad (16)$$

This is the Master equation for the Thattai-Van Oudenaarden model (2001). Lines 2 and 3 on the right-hand side of equation (16) correspond to the two reactions involving M but not P derived previously (see section 4). We will put $n_D = 1$ (n_D is the number of DNA molecules) from now on. This Master 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.

Using the same approach as in section 4 one obtains the equations satisfied by $< n_M >$ and $< n_P >$

$$d < n_M > /dt = -k_3 < n_M > +k_1 \quad (17)$$

$$d < n_P > /dt = -k_4 < n_P > +k_2 < n_M > \quad (18)$$

We verify again that these equations for population averages are similar to the concentration equations (cf. equations 12 and 13) 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 the "birth-and-death" process seen before (section 4.). As to protein number fluctuations, there is an additional term besides the Poisson term, corresponding to the fact that the number of 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 3.

6 Gillespie's algorithm (Gillespie, 1977)

Generally biological systems are much more complex than what is represented in 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, it is 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 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 the time interval (t, t+dt). $a_\mu(t)$ is the product of two parts: the reaction rate c_μ for the 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

$$P_1 \; + \; P_2 \stackrel{c_\mu}{\rightarrow} Z \tag{19}$$

where Z is the heterodimer formed of P_1 and P_2, one has
$a_\mu(t) = c_\mu P_1 P_2$
where the product $P_1 P_2$ represents the product of the numbers of P_1 and of P_2 molecules.
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

$$P_1 + P_2 \stackrel{k}{\rightarrow} Z \tag{21}$$

for which the chemical rate equation would read

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

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.

6.1 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 $a_0(t) = \sum a_\mu(t)$ the sum of all $a_\mu(t)$. $a_0(t)\,dt$ is the probability of any reaction occurring in (t, t+dt). 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 2.2.

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) $P_1 \rightarrow P_1 - 1$, $P_2 \rightarrow P_2 - 1$, and $Z \rightarrow Z + 1$. 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 1

We run the Gillespie simulation for the Thattai-Van Oudenaarden model with 200 cells, with the parameter values given in section 3. The results are in figure 1 where a number of quantities characterizing cell-to-cell variability are shown, in particular the comparison of average protein number and individual cell behavior for three randomly chosen cells in the top left panel. Each

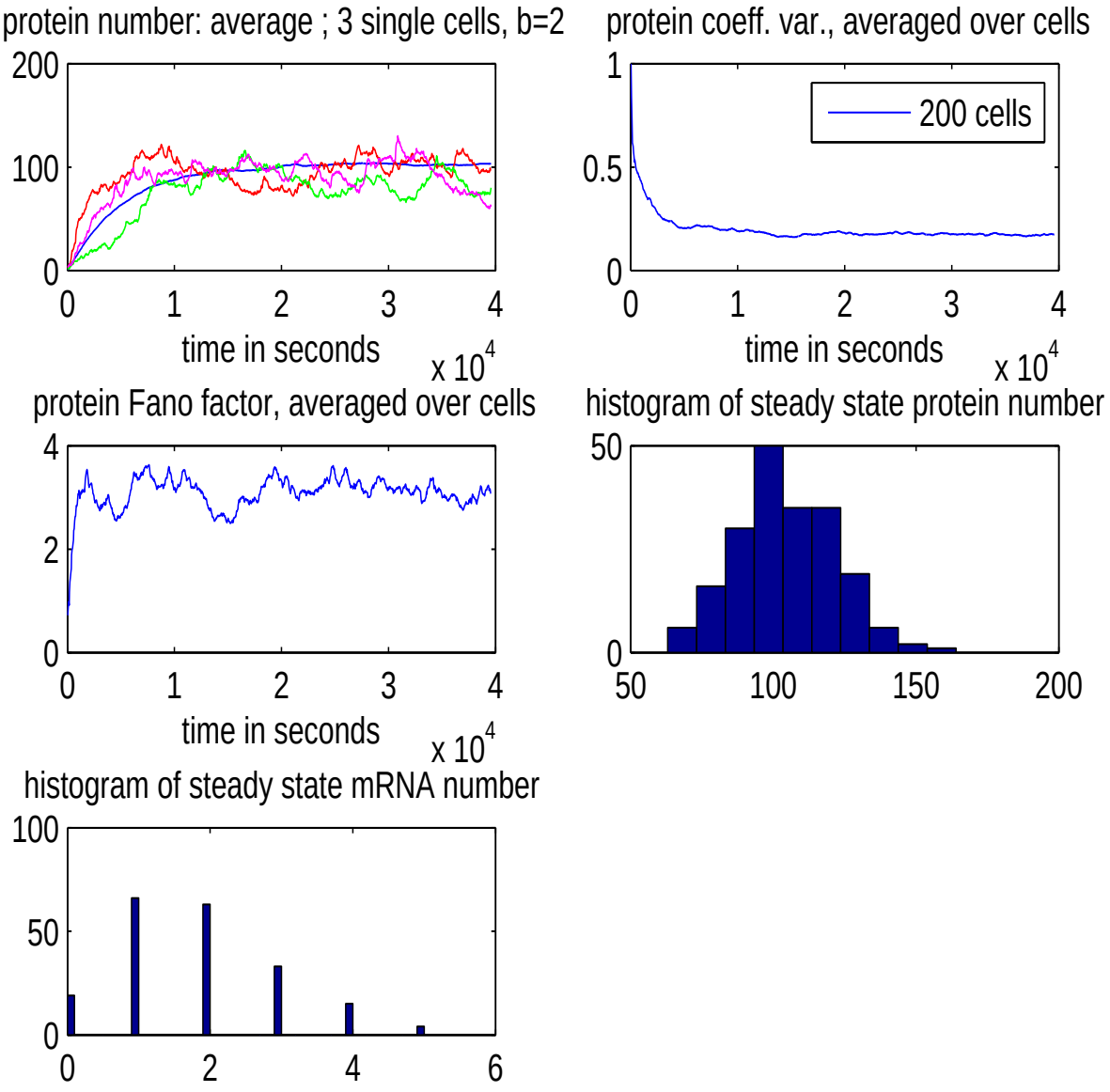

Figure 1: Each panel offers a different view of single cell variability in both transient and steady state regimes. The graphs correspond to b=2, for which $< n_P >= 104$.

cell starts off from the same initial state, and ends up after a time chosen long enough to reach steady state, with a different number of proteins, due to internal fluctuations. The average number is close to the value of 104 calculated for the deterministic system of section 3. 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 60 proteins to cells with as many as 150 proteins. The program, at the end of the run, also prints out the steady state values of the average number of proteins, Fano factor, and coefficient of variation. One can compare some of these values with the analytical results at the end of section 3, which give a value of 3 for the Fano factor, and 0.17 for the coefficient of variation. One should note that the numerical values for average number of proteins, or Fano factor fluctuate themselves from run to run. 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 1.2. As one includes more cells into each run, the statistical errors decrease, and one obtains results for the average number of proteins, for example, closer to the deterministic values. The bottom panel of Figure 1 shows the mRNA histogram. Whereas the deterministic approach (section 3 and equation (12)) gives a number of mRNA molecules equal to 1.73, the histogram shows that that number arises from a distribution over cells with 0,1,2,3,4, or 5 copy numbers. Here the poulation average gives a poor description of single cell behavior.

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. Transcriptional noise and bursting

The noise considered in the previous sections, evaluated in the Thattai-Van Oudenaarden model, is so-called "small copy number " noise. It is a consequence of the small numbers of molecules, mRNAs in particular, with fluctuations (coefficient of variation) typically decreasing as the square root of their number. There exist other sources of noise, classified usually as either intrinsic or extrinsic. An example of extrinsic noise would be cell-to-cell variability of some kinase that is activated by a stimulus and sets in motion a signaling pathway. Other sources of intrinsic noise are transcription or transcriptional bursting. Both are present in a model of interferon beta induction in single human dendritic cells infected by a virus (Hu et al., 2007,2009; Iyer-Biswas et al., 2009). The model that follows is a simplified version. It includes two transcription factors P1 and P2 that bind cooperatively to create complex Ds11. The latter goes back and forth to a state D* which produces mRNA (M) that also degrades. The reactions are the following:

$$D+P1 \leftrightarrows Ds1$$

$$Ds1+P2 \leftrightarrows Ds11$$

$$Ds11 \leftrightarrows D^*$$

$$D^* \rightarrow D^* + M$$

$$M \rightarrow 0$$

The first two reactions correspond to stochastic transcriptional binding, the last three represent busting during the time intervals transcription takes place from D*.

A MATLAB program simulates the reactions with Gillespie's algorithm. The simulation runs for 10 hours with 500 cells. Binding starts at 5 hours to mimic the situation where it takes about 5 hours after cell infection to activate transcription factors. The program produces 4 **figures:** a histogram of mRNA, the same histogram on a log-log plot with a linear fit, the exponent of which (printed out) determines the power law behavior of the distribution when it is long-tailed, the average mRNA as a function of time, and a histogram of times where each cell reaches the bursting state D* for the first time. The last figure makes clear that a fraction of the cells (printed out) never reaches induction stage. By changing forward or backward rates of the conversion of Ds11 to D*, and thus the probability of reaching the bursting state or the time spent in it, one can explore a range of mRNA distributions.

Acknowledgment:

I wish to thank the Summer School organizers,Uri Hershberg and Steven Kleinstein, and I am grateful to my colleague C. Jayaprakash (Department of Physics, Ohio State University) for many discussions and collaborationson stochastic phenomena. I also wish to acknowledge support from contract HHSN266200500021C of theNational Institute of Allergy and Infectious Diseases.

9 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.

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. 240, 583-591.

J. Hu, S.I. Biswas, S.C. Sealfon, J. Wetmur, C. Jayaprakash, and F. Hayot,Power laws in interferon mRNA distributions in virus-infected dendritic cells,Biophys. J. 97, 1984-1989 (2009).

J. Hu, S. Sealfon, F. Hayot, C. Jayaprakash, M. Kumar, A.C. Pendleton, A. Ganee, A. Fernandez-Sesma, T.M. Moran, and J.G. Wetmur, Chromosome-specific and noisy IFNB1 transcription in individual virus-infected human primary dendritic cells, Nucleic Acids Res. 35, 5232-5241 (2007).

S. Iyer-Biswas, F. Hayot, and C. Jayaprakash, Stochasticity of gene products from transcriptional pulsing, Phys. Rev. E, 031911 (2009).

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.

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.