7-1-2004

Modelling Stochastic Gene Expression

Olienka Patricia Fernandez
The University of Texas at El Paso

MODELLING STOCHASTIC GENE EXPRESSION

OLIENKA PATRICIA DE LA O FERNANDEZ

Department of Physics

APPROVED:

Jorge A. López, Ph.D., Chair

Clarence H. Cooper, Ph.D.

Luis G. Valdez Sánchez, Ph.D.

MODELLING STOCHASTIC GENE EXPRESSION

by

OLIENKA PATRICIA DE LA O FERNANDEZ, B.S.

REPORT

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Physics

THE UNIVERSITY OF TEXAS AT EL PASO

July 2004

Contents

List of Figures

Chapter 1

Modelling Stochastic Gene Expression

1.1 Stochasticity

Stochastic motion, is motion which is partly generated by a force of random strength, or by a force at random times, or, perhaps, by both. Thus, for stochastic systems (as opposed to deterministic motion), 'it is not possible', to determine *exactly* the state of the system at later times given its state at the current time.

A manner to describe a stochastic system, it is by the probabilities that the system is in certain states and how these probabilities evolve with time.

Often, such a calculation is difficult and we need focus, on finding the moments of the probability distribution, such as the mean and variance, both of which are convenient to compare with experimental observations.

Stochastic behaviour is often called noisy behaviour in the and the extent of stochasticity as the amount of noise in the system.

There are two types of stochasticity [1,2]:

1. Intrinsic (sometimes called internal), which is noise that arises as an inherent part of the motion of the system, and,

2. Extrinsic (or external), which results from a random force acting on an otherwise deterministic system.

Intrinsic noise is make up by thermal fluctuations, where, for the case of a chemical system for example, collisions of potential reactants with solvent molecules leads to random changes in the energy of the reactants. The probability of reaction will thus be different at different times and for different pairs of reactants depending on their individual behaviours of molecular collisions. Intrinsic noise therefore can be considered to arise from fluctuating reaction rate "constants".

Extrinsic noise, however, is generated by fluctuating chemical species created outside but that act on the system. These species cause additional system stochasticity by varying reaction probabilities through fluctuations in their con-

centrations rather than through varying rate constants.

1.2 A stochastic description of chemical reactions

For two molecules to interact they first have to physically find each other in solution. This finding process is stochastic, being driven by diffusion and responding to temperature and intermollecular collisions. Once the molecules are in close vicinity with each other, they still need to have enough energy to react (i.e. to go beyond the activation barrier of the reaction) and, again, their particular energetic state will depend, amongst other things, on their collision history. Both effects lead to chemical reactions being, in general, stochastic processes. However, from a modelling perspective, including that stochasticity is only important when fluctuations are, in some sense, significant. When noise is small, a set of deterministic differential equations (based on the law of mass action) is an entirely appropriate approximation. It is therefore important to understand the factors that set the magnitude of chemical stochasticity.

1.3 The probability of a chemical reaction

The stochasticity inherent in chemical reactions leads to the concept of the *probability* that a reaction will occur in a small time interval δt, in contrast to a deterministic way, where the reaction will always occur for any interval δt. In reality, and for a stochastic model, this deterministic assumption is not true because for very small δt, reactants will simply not have enough time to find each other. Consider the second - order reaction shown in Figure 1.1.

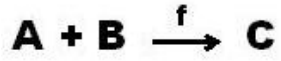

Figure 1.1: *A simple binary (second - order) reaction.*

Denoting $\mathcal{P}$ as probability, we can write

$\mathcal{P}$(reaction in δt)= $\mathcal{P}$(an A and a B molecule colliding)$\times\mathcal{P}$(orientation and energy of collision sufficient for a reaction)

If the last probability is nearly always close to one, the reaction is called diffusion-limited. Considering both A and B molecules as hard spheres of radius r_1 and r_2, respectively, the probability of a collision will depend on the relative velocity of A to B, denoted as v_{AB}. From Figure 1.2, a and b, the collision

volume, δV, swept out by B in time δt is

$$\delta V = \pi r_{AB}^2 v_{AB} \delta t \tag{1.1}$$

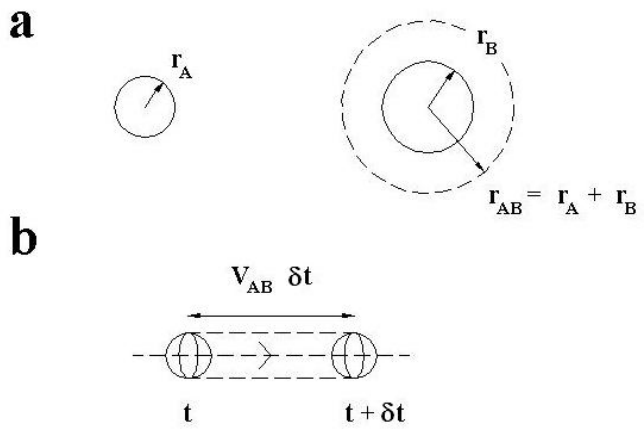

Figure 1.2: *a. Collision between two potential reactants. If A comes within a radius of $r_{AB} = r_A + r_B$ of B, a collision occurs. b. The volume swept out by one hard sphere in time δt is $\pi r_{AB}^2 v_{AB} \delta t$.*

If the system is at thermal equilibrium, then a single A is equally likely to be anywhere in V (the volume of the system) and the probability of finding an A in the collision volume is just $\delta V \setminus V$. Thus,

$$\mathcal{P}(\text{a particular A and a B molecule colliding}) = \frac{\pi r_{AB}^2 \bar{v}_{AB} \delta t}{V} \tag{1.2}$$

where we have replaced the relative velocity, v_{AB}, by its mean value. The probability of a collision somewhere in V, is (1.2) multiplied by the number of pairs of possible reactants,

$$\mathcal{P}(\text{an A and a B molecule colliding}) = \frac{n_A n_B \pi r_{AB}^2 \bar{v}_{AB} \delta t}{V} \tag{1.3}$$

for n_A and n_B molecules of A and B, respectively. Finally, defining q to be the probability that, given a collision, a reaction occurs

$$\begin{aligned} \mathcal{P}(\text{reaction in } \delta t) &= \frac{n_A n_B}{V} \times q\pi r_{AB}^2 \bar{v}_{AB} \delta t \\ &\sim \frac{n_A n_B}{V} \end{aligned} \tag{1.4}$$

The reaction probability thus increases as the numbers of A and B molecules rise and decreases with larger system volume (as it is then harder for molecules to find each other to react).

Then defining the probability that a particular A and B will react in unit time, f, by

$$f = \frac{q\pi r_{AB}^2 \bar{v}_{AB}}{V} \tag{1.5}$$

which is measured in inverse seconds and is related to the 'usual' rate constant of the reaction (measured in $M^{-1}\ s^{-1}$).

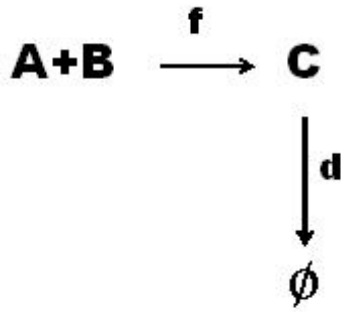

Figure 1.3: *A simple reaction scheme. A and B bind irreversibly to form complex C with probability f per unit time and individual C molecules degrade with probability d per unit time*

1.4 The Master equation

Stochastic chemical reactions, which occur with a certain probability per unit time, imply that once a reaction has started we can no longer know with certainty the numbers and types of molecules present at later times but can only adopt a probabilistic description. For the reactions shown in Figure 1.3, for example, the system can be described by

$\mathcal{P}(n_A$ molecules of A, n_B molecules of B, and n_C molecules of C at time $t)$

and how this probability evolves with time.

Consider a time interval δt small enough so that at most one reaction can occur and write $P(n_A, n_B, n_C, t)$ for the probability. If the system has n_A, n_B, and n_C molecules of A, B, and C, respectively, at time $t + \delta t$ then if reaction f occurred during the interval δt, the system must have been in the state $n_A + 1$, $n_B + 1$, and $n_C - 1$ at time t. Alternatively, reaction d could have occurred

during t and so the system then must have been in the state n_A, n_B, and n_C+1 at time t. Finally, no reaction may have occurred at all and so the system would be unchanged (at n_A, n_B, and n_C) at t. The probability of these reactions, from the state n_A, n_B, and n_C are

$$\begin{aligned} \mathcal{P}\,(f \text{ reaction}) &= f n_A n_B \delta t \\ \mathcal{P}\,(d \text{ reaction}) &= d n_C \delta t \\ \mathcal{P}\,(\text{no reaction}) &= 1 - f n_A n_B \delta t - d n_C \delta t \end{aligned} \tag{1.6}$$

Thus, we can write

$$\begin{aligned} P(n_A, n_B, n_C, t+\delta t) = \\ P(n_A+1, n_B+1, n_C-1, t)(n_A+1)(n_B+1) f \delta t \\ + P(n_A, n_B, n_C+1, t)(n_C+1) d \delta t \\ + P(n_A, n_B, n_C, t)(1 - n_A n_B f \delta t - n_C d \delta t) \end{aligned} \tag{1.7}$$

Dividing (1.7) through by δ t and taking the limit $\delta t \to 0$ gives

$$\begin{aligned} \frac{\partial}{\partial t} P_{n_A,n_B,n_C} &= f\left[(n_a+1)(n_B+1) P_{n_{A+1},n_{B+1},n_{C-1}} - n_a n_B P_{n_A,n_B,n_C}\right] \\ &\quad - d[n_C P_{n_A,n_B,n_C} - (n_C+1) P_{n_A,n_B,n_{C+1}}] \end{aligned} \tag{1.8}$$

where we have abbreviated $P(n_A, n_B, n_C, t)$ to P_{n_A,n_B,n_C} (1.8) is called a Master equation, because from it all the moments of the distribution can be derived and describes how the probability of the system changes with time.

Solving the Master equation can be done for simple (linear) systems but often only at steady - state (see [4] and [5] for methods of solution). We can use (1.8), however, to derive the equation of motion for the mean of C, which is defined as

$$\langle C \rangle = \sum_{n_A,n_B,n_C} n_C P_{n_A,n_B,n_C} \tag{1.9}$$

Multiplying (1.8) by n_C and summing over n_A, n_B, and n_C gives

$$\begin{aligned} \frac{\partial}{\partial t}\langle C \rangle = \quad & f \sum (n_c - 1 + 1)(n_A + 1)(n_B + 1) P_{n_A+1,n_B+1,n_C-1} \\ - \quad & f \sum n_A n_B n_C P_{n_A,n_B,n_C} - d \sum n_C^2 P_{n_A,n_B,n_C} \\ + \quad & d \sum (n_c + 1 - 1)(n_C + 1) P_{n_A,n_B,n_C+1} \end{aligned} \tag{1.10}$$

where terms in round brackets have been factored to follow the subscripts of P. Therefore, by using results such as

$$\sum_{n_A,n_B,n_C=0}^{\infty} n_A, n_B, n_C P_{n_A,n_B,n_C} \equiv \langle ABC \rangle$$
$$= \sum_0^\infty (n_A+1)(n_B+1)(n_C-1)$$
$$\times P_{n_A+1,n_B+1,n_C-1} \tag{1.11}$$

given that Pn_A, n_B, n_C must be zero if any of n_A, n_B or n_C are negative (minus numbers of A molecules, for example, can not exist), we have

$$\frac{\partial}{\partial t}\langle C \rangle = f\,[\,\langle ABC \rangle + \langle AB \rangle\,] - f\,\langle ABC \rangle - d\,\langle C^2 \rangle + d\,[\,\langle C^2 \rangle - \langle C \rangle\,]$$
$$= f\,\langle AB \rangle - d\,\langle C \rangle \tag{1.12}$$

It is interesting to compare (1.12) with a deterministic model of Figure (1.3). Using the law of mass action, $[C]$, the concentration of C, obeys

$$\frac{d}{dt}[C] = \widetilde{f}[A][B] - \widetilde{d}[C] \tag{1.13}$$

where $\widetilde{f}$ and $\widetilde{d}$ are the macroscopic (deterministic) rate constants and do not necessarily have units of inverse time. The concentration of a species can

be related to the underlying stochastic processes generating the reactions by

$$[C] = \frac{\langle C \rangle}{V} \tag{1.14}$$

and so the deterministic equations are equations for the rate of change of the *means* of the different chemical species involved. Using (1.14), (1.13) becomes

$$\frac{d}{dt}\langle C \rangle = \frac{\widetilde{f}}{V}\langle A \rangle \langle B \rangle - \widetilde{d}\langle C \rangle \tag{1.15}$$

Comparing this equation with (1.12), we find the relationship between the stochastic probabilities of reaction per unit time to the deterministic reaction rate constants:

$$\begin{aligned} \widetilde{f} &= \frac{V\langle AB \rangle}{\langle A \rangle \langle B \rangle} \cdot f \text{ second-order reaction} \\ \widetilde{d} &= d \text{ first-order reaction} \end{aligned} \tag{1.16}$$

Hence, for first-order reactions both the kinetic rate and the probability are the same. If the deterministic approximation is adopted, then

$$\langle AB \rangle \simeq \langle A \rangle \langle B \rangle \tag{1.17}$$

is necessarily part of that approximation and so

$$\widetilde{f} \simeq fV \tag{1.18}$$

for second-order reactions. This relationship determines the equivalent macroscopic rate given the probability of reaction, f. Equations (1.16) and (1.18) generally provide the inter-conversion between reaction rate constants and reaction probabilities.

Some numbers for *Escherichia coli*

Usually, any reaction rates that have been measured experimentally have implicity assumed the deterministic approximation in the measurement process. For modelling purposes, we then often have to calculate f from measured $\widetilde{f}$. A diffusion-limited reaction, for example, is expected to have

$$\widetilde{f} = 10^9 M^{-1} s^{-1}$$

for concentrations measured in molar units [6]. The volume of a typical *E. coli* bacterium is approximately 2×10^{-15} liters [1], which implies that 1 molecule has a concentration of

$$\frac{1/N_{Avo}}{V} = \frac{1}{6 \times 10^{23} \times 2 \times 10^{-15}} \simeq 10^{-9} M$$

i.e. approximately 1nM (here, N_{Avo} is Avogadro's Number).

Thus,

$$\begin{aligned} f &= \frac{\tilde{f}}{V} \\ &= \frac{10^9}{N_{Avo}V} \\ &\simeq 1 \end{aligned} \tag{1.19}$$

i.e. a rate of $10^9 M^{-1}s^{-1}$ corresponds to a probability of almost unity for a single reaction per second, which is a useful relationship to remember.

1.5 An exception: homo-dimerization reactions

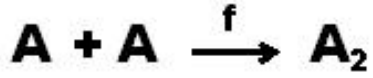

Figure 1.4: *The formation of a homo-dimer. Two A monomers combine to form an A dimer.*

For homo-dimerization reactions (reactions between 2 equal molecules), where two like-molecules come together, as illustrated in Figure (1.4), the number of

unique pairs of possible reactants is not n_A^2 but $n_A^2 = \frac{1}{2}n_A(n_A - 1)$. This change alters (1.3) and so leads to (1.16) becoming

$$\widetilde{f} = \frac{\frac{1}{2}V\langle A(A-1)\rangle}{\langle A\rangle\langle A\rangle} \cdot f \tag{1.20}$$

However, consistency of the deterministic approximation implies

$$\langle A(A-1)\rangle \simeq \langle A\rangle^2 \tag{1.21}$$

and so

$$\widetilde{f} \simeq \frac{fV}{2} \tag{1.22}$$

which is the inter-conversion formula for dierization reactions.

1.6 Simulating stochastic biochemical reactions

The algorithm that is most commonly used to stochastically simulate biochemical systems is that of Gillespie[3]. The equivalent of two dice are rolled on the computer; one to choose which reaction will occur next and the other to decide when that reaction will occur. Let us assume that we have a system in which n reactions can take place, then

$\mathcal{P}$ (reaction i occurs between $t+\tau$ and $t+\tau+\delta\tau$) $\equiv P_i(\tau)\delta\tau$

needs to be calculated for each reaction. For example, if reaction i corresponds to the reaction shown in Figure (1.1), then

$$\begin{aligned} \mathcal{P}(\text{reaction } i \text{ in time } \delta\tau) &= n_A n_B f \delta\tau \\ &\equiv a_i \delta\tau \end{aligned} \tag{1.23}$$

where a_i is referred to as the propensity of reaction i. Therefore,

$$\begin{aligned} rclP_i(\tau)\delta\tau &= \mathcal{P}(\text{no reaction for time } \tau) \\ &\quad \times \mathcal{P}(\text{reaction } i \text{ happens in time } \delta\tau) \\ &\equiv P_0(\tau) a_i \delta\tau \end{aligned} \tag{1.24}$$

with $P_0(\tau)$ the probability that no reaction occurs during the interval τ. This probability satisfies

$$P_0(\tau + \delta\tau) = P_0(\tau)[1 - \sum_{j=1}^{n} a_j \delta\tau] \tag{1.25}$$

which implies

$$\frac{dP_0}{d\tau} = -\sum_{j=1}^{n} a_j \tag{1.26}$$

and so

$$P_0(\tau) = exp\left(-\sum a_j \tau\right) \tag{1.27}$$

Equation (1.27) implies that

$$P_i(\tau) = a_i e^{-\sum a_j \tau} \tag{1.28}$$

and so to choose in the simulation which reaction happens next an n-sided die must be rolled with each side corresponding to a reaction and weighted appropriately by that reaction's propensity. A second die is used to sample from (1.27) to determine the time when the reaction will occur. The chemical species and time variable in the simulation are updated to reflect the occurrence of the reaction and the process is then repeated. See[3] for more details.

1.7 Definition of noise

A definition of the noise of a probability distribution is: **the coefficient of variation**, which is defined as the ratio of the standard deviation of the distribution to its mean, or in physics language, the inverse of the signal-to-noise

ratio. We will denote this definition of noise by the symbol η

$$\eta = \frac{\sqrt{\langle N^2 \rangle - \langle N \rangle^2}}{\langle N \rangle} \tag{1.29}$$

for the random variable N. The noise, η, has the advantage of being dimensionless and so easy to interpret. It measures the magnitude of a typical fluctuation of a quantity relative to its mean and so seems most relevant to the engineering 'design' of biochemical networks, where deviations of chemical components of the network away from their mean levels is likely to significantly affect network function. One might expect that biological networks have evolved to limit such fluctuations so that the concentrations of important components are reliably maintained close to their mean values.

An alternative definition is the **Fano factor**

$$\text{Fano factor} = \frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle} \tag{1.30}$$

which is the variance of the distribution divided by the mean and so can potentially have dimensions. It is mainly used to compare the noise in a particular stochastic process to the noise in a Poisson process (a simple 'birth-and-death' process) for which the Fano factor is equal to unity (see Figure (1.5)).

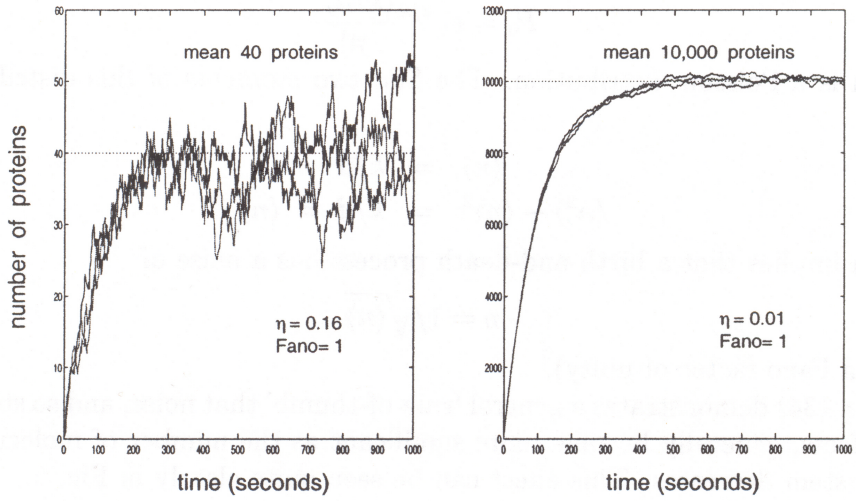

Figure 1.5: *Three simulation runs of two simple (birth-and-death) models of protein production. Each model involves an identical set of reactions but different parameters values leading to different mean protein levels. In this case, the probability distribution for protein numbers is Poisson and the Fano factor is always one. The coefficient of variation, η, does however determine different levels of noise in the two processes.*

1.8 Poisson ('birth-and-death') processes

A very simple model of gene expression can be obtained from the reaction scheme of Figure (1.3) by letting n_A and n_B become fixed. For example, n_A and n_B could be considered the number of molecules of DNA and RNA polymerase, respectively, which can be approximated as fixed at constant concentrations.

By defining $k = fn_An_B$, Figure (1.3) collapses to the scheme of Figure (1.6) with protein C being born on average every $1/k$ seconds and being degraded ('dies') with rate d.

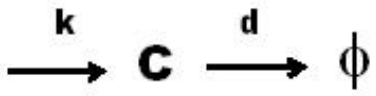

Figure 1.6: *A simple model of gene expression*

The Master equation for Figure (1.6) is a simplified version of (1.8)

$$\frac{\partial}{\partial t}P_n = k[P_{n-1} - P_n] - d[nP_n - (n+1)P_{n+1}] \tag{1.31}$$

with $P_n(t)$ the probability of having n molecules of protein C at time t. While we won't solve (1.31) explicitly here (we will derive the moments of P_n later), the solution can be found using moment generated functions [4] and is

$$P_n = e^{-k/d}\frac{k/d^n}{n!} \tag{1.32}$$

which is a Poisson distribution. The first two moments of this distribution

are

$$\begin{aligned} \langle n \rangle &= \frac{k}{d} \\ \left\langle n^2 \right\rangle - \langle n \rangle^2 &= \frac{k}{d} = \langle n \rangle \end{aligned} \tag{1.33}$$

which implies that a birth-and-death process has a noise of

$$\eta = \frac{1}{\sqrt{\langle n \rangle}} \tag{1.34}$$

(and a Fano factor of unity). (1.34) demonstrates a general 'rule-of-thumb' that noise, and so stochastic effects, generally become more significant as the number of molecules in the system decreases. This effect can be seen quite clearly in Fig. (1.5)

1.9 An improved model of gene expression

The scheme of Figure (1.6) lumps together the processes of transcription and translation into one first-order reaction k. To determine the causes of noise in biological systems, it is essential to model these two processes individually. Figure (1.7) shows a model which makes this distinction but is still simple enough to be mathematically tractable. Transcription and translation are both

approximated as first-order processes, leading to the model containing both $mRNA$, M, and protein, N, species. Each of these species has a half-life set by their degradation rate (d_0 and d_1, respectively). Typically, $d_1 \ll 0$ reflecting protein life-times of hours compared to those of $mRNA$ which are only a few minutes.

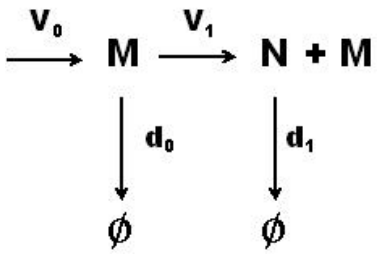

Figure 1.7: *A model of gene expression, which explicitly includes transcription (rate v_0) and translation (rate v_1). $mRNA$ and protein are denoted M and N, respectively.*

1.10 The Langevin solution

Langevin theory is a way of avoiding solution of the Master equation for a system by explicitly adding noise terms (stochastic variables) to the macroscopic (deterministic) equations of motion. For the model of Figure (1.7), the deterministic equations are

$$\begin{aligned} \frac{dM}{dt} &= v_0 - d_0 M \\ \frac{dN}{dt} &= v_1 M - d_1 N \end{aligned} \tag{1.35}$$

where M and N denotes the numbers of $mRNA$ and protein, respectively. A Langevin model adds a stochastic variable, $\xi(t)$, to each of these equations

$$\begin{aligned} \frac{dM}{dt} &= v_0 - d_0 M + \xi_1(t) \\ \frac{dN}{dt} &= v_1 M - d_1 N + \xi_2(t) \end{aligned} \tag{1.36}$$

and is only fully specified when the probability distributions for the ξ_i are also given.

There are two classes of choices for the properties of the ξ_i:

Extrinsic noise The moments of the ξ_i can be arbitrarily chosen to match the strength of the extrinsic noise.

Intrinsic noise The ξ_i are carefully specified so that they mimic thermal fluctuations and so successfully model intrinsic noise. The solution of the Langevin equation should then be a good approximation to that of the Master equation (and an exact solution in some limit).

Understanding noise: auto-correlations

To specify the ξ_i correctly we need to be able to characterize a noise distribution. Often, it is sufficient, for example if we wish to know only the variance of a fluctuating variable, to find the first two moments of ξ_i and its auto-correlation time. This time-scale, denoted τ, in general describes, on average, the life-time of a typical fluctuation, as well as the approximate average time separating the occurrence of such fluctuations. Figure (1.8) shows typical behaviour of a fluctuating variable obeying a Poisson distribution, where time has been rescaled by the auto-correlation time. On average, the number of molecules changes significantly only over the auto-correlation time (one in this units), or longer.

To estimate the typical life-time of fluctuations (and so to characterize the behaviour of fluctuating quantity over time), one can ask how correlated a deviation of the variable of interest away from its mean at time t_2 is with the deviation from the mean at a later time t_1. The auto-correlation function,

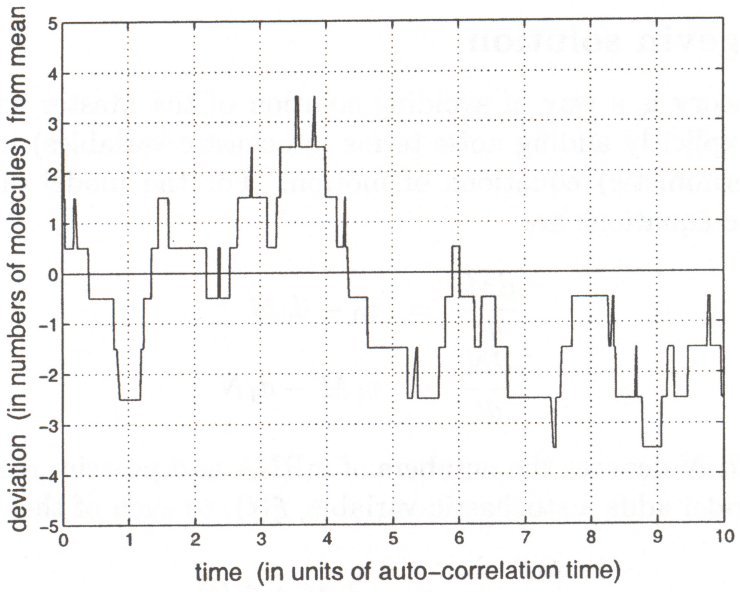

Figure 1.8: *A time-series of a Poisson (birth-and-death) process (shown in Figure (1.6). Time has been rescaled by the auto-correlation time so that on average fluctuations should last for approximately one unit. The deviation away from the mean, i.e.* $n - \langle n \rangle$*, is plotted on the y-axis for clarity.*

defined as

$$\begin{aligned} C_\xi(t_1, t_2) &= \langle [\xi(t_1) - \langle \xi(t_1) \rangle] [\xi(t_2) - \langle \xi(t_2) \rangle] \rangle \\ &= \langle \{ \xi(t_1)\xi(t_2) - \langle \xi(t_1) \rangle \xi(t_2) - \xi(t_1) \langle \xi(t_2) \rangle + \langle \xi(t_1) \rangle \langle \xi(t_2) \rangle \} \rangle \\ &= \langle \xi(t_1)\xi(t_2) \rangle - \langle \xi(t_1) \rangle \langle \xi(t_2) \rangle \end{aligned} \tag{1.37}$$

provides a quantitative answer to this question. Note that when $t_1 = t_2$,

(1.37) is just the variance of $\xi(t)$ and often the correlation function is normalized by dividing by this quantity (so that it has a value of unity for $t_1 = t_2$).

For stationary processes (processes that are invariant under time translations and so are statistically identical at all time points), such as the steady state behaviour of chemical systems, the correlation function obeys

$$C_\xi(t_1, t_2) = C_\xi(t_1 - t_2) \tag{1.38}$$

and is just a function of one variable, the time difference between the two times points considered. Figure (1.9) shows the auto-correlation function for the Poisson model of gene expression (1.6) at steady-state. the correlation function is normalized by the variance (at $t = 0$) and is well-fit by an exponential decay, $e^{-t\backslash\tau}$, where τ defines the auto-correlation time. A typical fluctuation only persists for this time-scale, as it allows enough new events to occur (production or degradation of molecules, in our case) to change behaviour to such an extent as to break correlation with earlier times. After the auto-correlation time, there is no memory left of early behaviour and it is impossible to predict the state of system (i.e. its deviation away from the mean) at times greater than an auto-correlation time away given the state of the system at the current time.

For simple, linear systems, such as the Poisson process of Figure (1.6), it is

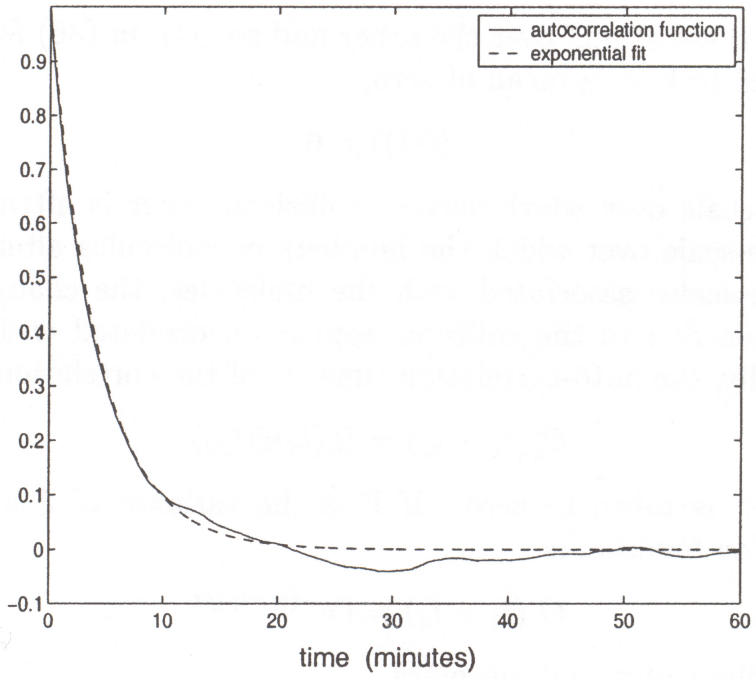

Figure 1.9: *Auto-correlation function for the Poisson process of Figure (1.6) with a typical time series extract shown in Figure (1.8). The dotted line is an exponential fit using an auto-correlation time of* $1 \simeq 4.2$ *minutes.*

the time-scale associated with degradation which sets the auto-correlation time at steady-state; as it is degradation that provides the restoring force that keeps the number of proteins fluctuating around their mean steady-state value. The probability of degradation in time , dn, *changes* as the number of proteins, n, changes. It increases as the number of molecules rises above the mean value, raising the probability of degradation, and so increasing the probability of a return to mean levels. Similarly, fluctuations which take n below mean levels, lower the probability of degradation and again raise the probability of recovering

mean values. Thus, fitting and exponential to Figure (1.9) recovers the inverse of the degradation rate as the auto-correlation time (which can also be shown to be true analytically).

1.11 White noise

Langevin theory was originally applied to describe thermal fluctuations, i.e. noise that arises from changes in energy of the molecule of interest due to its constant collisions with particles of surrounding gas or solvent. Such collisions can either act to increase the probability of the reaction that produce the molecules of interest or to increase the probability that these molecules are degraded. There is no reason, *a priori*, why thermal fluctuations should favour one of these effects over the other and so $\xi(t)$, in (1.36) for example, is usually defined to have a mean of zero,

$$\langle \xi(t) \rangle = 0 \tag{1.39}$$

The time-scale over which solvent collisions occur is often much faster than the time-scale over which the numbers of molecules alters, and so, at the long time-scales associated with the molecules, the changes in energy of the molecule due to the collision appear uncorrelated with each other. Mathematically, the

auto-correlation time, τ, of the correlation function

$$C_\xi(t_1 - t_2) = \langle \xi(t_1)\xi(t_2) \rangle \tag{1.40}$$

as $\langle \xi(t) \rangle = 0$, is taken to zero. If Γ is the variance of ξ at time t, the auto-correlation function is

$$C_\xi(t_1 - t_2) = \Gamma e^{-(t_1 - t_2)\backslash\tau} \tag{1.41}$$

which in the limit of $\tau \to 0$, becomes

$$\langle \xi(t_1)\xi(t_2) \rangle = \begin{cases} 0 & \text{for } t_1 \neq t_2 \\ \Gamma & \text{for } t_1 = t_2 \end{cases} \tag{1.42}$$

or

$$\langle \xi(t_1)\xi(t_2) \rangle = \Gamma \delta(t_1 - t_2) \tag{1.43}$$

where $\delta(t)$ is the Dirac delta function. A stochastic variable that obeys (1.39) and (1.43), i.e. that is completely uncorrelated with zero mean, is referred to as 'white' noise (noise variables with a finite auto-correlation time, in contrast, are 'coloured'). Usually, Langevin models that deal with thermal fluctuations introduce white noise terms into the equations of motion. This is

precisely the approach we will take with (1.36), our model of gene expression. The parameter Γ is referred to as the noise strength and needs to be carefully specified.

1.12 Langevin theory for stochastic gene expression

We now return to modelling gene expression, via Figure (1.36) which is shown again below

$$\frac{dM}{dt} = v_0 - d_0 M + \xi_1(t)$$
$$\frac{dN}{dt} = v_1 M - d_1 N + \xi_2(t) \quad (1.44)$$

While from the arguments presented above, we expect ξ_1 and ξ_2 to have zero mean and zero auto-correlation times, we can show that this assumptions are true explicitly by considering the steady-state solution of (1.44) in the absence of the stochastic variables, ξ_i,

$$M_s = \frac{v_0}{d_0} \quad ; \quad N_s = \frac{v_1}{d_1} M_s \quad (1.45)$$

If we assume that the system is at (or near) steady-state and consider a time

interval δt small enough such that at most only one reaction can possibly occur, then, ξ_1 and ξ_2, which account for any reactions that might take place, can only have the values

$$\xi_i \delta t = \begin{cases} +1 \\ 0 \\ -1 \end{cases} \quad (1.46)$$

where $i = 1$ or 2, as the number of N or M molecules can only increase by one, decrease by one, or remain unchanged in the small time δt.

Define

$$P(i, j) = \mathcal{P}(\xi_1 = i, \xi_2 = j)$$

i.e. the probability that the number of mRNAs and proteins change by the amounts i and j, respectively, then Figure (1.7), at steady-state, implies

$$\begin{aligned} P(+1, 0) &= v_0 \delta t \\ P(+1, -1) &= 0 \\ P(+1, -1) &= 0 \end{aligned}$$

$$\begin{aligned} P(-1,0) &= d_0 M_s \delta t \\ P(-1,+1) &= 0 \\ P(-1,-1) &= 0 \end{aligned}$$

$$\begin{aligned} P(0,+1) &= v_1 M_s \delta t \\ P(0,0) &= 1 - v_0 \delta t - v_1 M_s \delta t - d_0 M_s \delta t - d_1 N_s \delta t \\ P(0,-1) &= d_1 N_s \delta t \end{aligned} \tag{1.47}$$

and so we can use these probabilities to calculate the moments of the ξ_i. Firstly,

$$\begin{aligned} \langle \xi_1 \delta t \rangle &= (+1)_0 \delta t + (-1)_0 M_s \delta t + (0) \times (1 - v_0 \delta t - d_0 M_s) \\ &= (v_0 - d_0 M_s) \delta t \\ &= 0 \end{aligned} \tag{1.48}$$

and

$$\begin{aligned} \langle \xi_2 \delta t \rangle &= (+1) \times v_1 M_s \delta t + (-1) \times d_1 N_s \delta t \\ &= (v_1 M_s - d_1 N_s) \delta t \\ &= 0 \end{aligned} \tag{1.49}$$

using (1.45). The means are both zero, as expected, and the ξ_i act to keep the system at steady-state (as they should).

For the mean square, we have

$$\begin{aligned} \left\langle \xi_1^2 \delta t^2 \right\rangle &= (+1)^2 \times v_0 \delta t + (-1)^2 \times d_0 M_s \delta t \\ &= (v_0 + d_0 M_s)\delta t \\ &= 2 d_0 M_s \delta t \end{aligned} \tag{1.50}$$

and, similarly,

$$\begin{aligned} \left\langle \xi_2^2 \delta t^2 \right\rangle &= 2 d_1 N_s \delta t \\ \langle \xi_1 \xi_2 \rangle &= 0 \end{aligned} \tag{1.51}$$

If the fluctuations are never large enough to drive M and N far from their steady-state values, i.e.

$$|M - M_s| \ll M_s \quad ; \quad |N - N_s| \ll N_s \tag{1.52}$$

then the probabilities, (1.47), for ξ_1 and ξ_2, will always be approximately true. In other words, if the system is close to steady-state and the steady-state values of M_s and N_s are large enough such that (1.52) is true, then we can

assume that (1.47) holds for all times, which implies that ξ_1 at time t_2, where $|t_2 - t_1| > \delta t$ (just as the throws of a die, whose outcomes are given by fixed probabilities, are also uncorrelated). Thus, we define as white noise terms

$$\begin{aligned} \langle \xi_1(t_1)\xi_2(t_2)\rangle &= 2d_0 M_s \delta(t_1 - t_2) \\ \langle \xi_2(t_1)\xi_2(t_2)\rangle &= 2d_1 N_s \delta(t_1 - t_2) \\ \langle \xi_1(t_1)\xi_2(t_2)\rangle &= 0 \end{aligned} \tag{1.53}$$

with their Γ_i given by (1.50) and (1.51).

This definition of ξ_1 and ξ_2 implies that the solution of (1.44) at steady-state for choices of the rate constants such that (1.52) holds, will recover the mean and variance of N and M that would be obtained by direct solution of the Master equation corresponding to Figure (1.7).

1.13 A future simplification

It is possible to solve directly the two coupled differential equations, (1.44), but we can also take advantage of the very different time-scales associated with mRNA and with protein. Typically, mRNA life-time is of order minutes while that of protein is often several hours. Figure (1.10)shows a time series extract with the much longer auto-correlation time protein $(1 \setminus d_1)$ compared to mRNA

$(1 \setminus d_0)$ clearly visible.

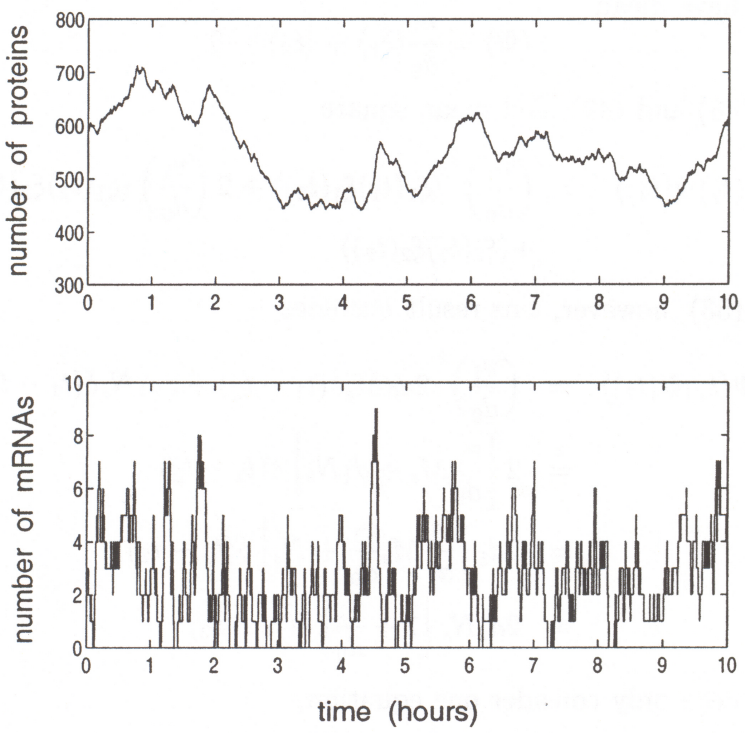

Figure 1.10: *Protein and mRNA numbers from a simulation of the scheme of Figure (1.7). Protein half-life is approximately 1 hour while that of mRNA is only 3 minutes resulting in very different behaviours.*

During one protein fluctuation, many mRNA fluctuations occur and so the mean level of mRNA during that fluctuation is the steady-state level. Therefore, we can set

$$\frac{dM}{dt} \simeq 0 \tag{1.54}$$

which implies that

$$\begin{aligned} M &= \frac{v_0}{d_0} + \frac{\xi_1}{d_0} \\ &= M_s + \frac{\xi_1}{d_0} \end{aligned} \tag{1.55}$$

Consequently, the equation for protein, (1.44), becomes

$$\frac{dN}{dt} = v_1 M_s - d_1 N + \frac{v_1}{d_0}\xi_1 + \xi_2 \tag{1.56}$$

and so is a function of the two stochastic variables, ξ_1 and ξ_2. To simplify (1.56), we can define a new stochastic variable

$$\Psi = \frac{v_1}{d_0}\xi_1 + \xi_2 \tag{1.57}$$

which will have mean

$$\langle \Psi \rangle = \frac{v_1}{d_0} \langle \xi_1 \rangle + \langle \xi_2 \rangle = 0 \tag{1.58}$$

from eqs (1.48) and (1.49), and mean square

$$\begin{aligned}\langle\Psi(t_1)\Psi(t_2)\rangle &= \left(\frac{v_1}{d_0}\right)^2\langle\xi_1(t_1)\xi_1(t_2)\rangle + 2\left(\frac{v_1}{d_0}\right)\langle\xi_1(t_1)\xi_2(t_2)\rangle \\ &\quad + \langle\xi_2(t_1)\xi_2(t_2)\rangle \end{aligned} \tag{1.59}$$

From eqs. (1.53), however, this result becomes

$$\begin{aligned}\langle\Psi(t_1)\Psi(t_2)\rangle &= \left(\frac{v_1}{d_0}\right)^2 2d_0M_s\delta(t_1-t_2) + 2d_1N_s\delta(t_1-t_2) \\ &= 2\left[\frac{v_1^2}{d_0}M_s + d_1N_s\right]\delta(t_1-t_2) \\ &= 2d_1\left[\frac{v_1}{d_0}M_s\frac{v_1}{d_0} + N_s\right]\delta(t_1-t_2) \\ &= 2d_1N_s\left[1+\frac{v_1}{d_0}\right]\delta(t_1-t_2) \end{aligned} \tag{1.60}$$

and so we need only consider one equation,

$$\frac{dN}{dt} = v_1M_s - d_1N + \Psi(t) \tag{1.61}$$

where the effects of the mRNA fluctuations have been absorbed into the protein noise term, Ψ, resulting in an increase in its magnitude - compare (1.60) and (1.53).

1.14 Solving the model

Eq. (1.61) can be written as

$$\frac{d}{dt}(Ne^{d_1 t}) = v_1 M_s e^{d_1 t} + \Psi e^{d_1 t} \tag{1.62}$$

and so integrated

$$N(t)e^{d_1 t} - N_s = \frac{v_1 M_s}{d_1}(e^{d_1 t} - 1) + \int_0^t \Psi(t')e^{d_1 t'} dt' \tag{1.63}$$

where we have assumed that initially we are at steady-state, i.e. $N = N_s$ when $t = 0$. Thus

$$N(t) = N_s + e^{-d_1 t} \int_0^t \Psi(t')e^{d_1 t'} dt' \tag{1.64}$$

Using the properties of $\Psi(t)$, (1.58) and (1.60), as well as (1.64), the mean protein number satisfies

$$\begin{aligned} \langle N(t) \rangle &= N_s + e^{-d_1 t} \int_0^t \langle \Psi(t') \rangle \, e^{d_1 t'} dt' \\ &= N_s \end{aligned} \tag{1.65}$$

and so the steady-state is stable to fluctuations (as expected).

We can also use (1.64) to find the mean of the square of the protein number,

$$\begin{aligned}\langle N(t_1)N(t_2)\rangle \\ &= \left\langle \left[N_s + e^{-d_1 t_1}\int_0^{t_1}\Psi(t')e^{d_1 t'}dt'\right]\times\left[N_s + e^{-d_1 t_2}\int_0^{t_2}\Psi(t'')e^{d_1 t''}dt''\right]\right\rangle \\ &= N_s^2 + e^{-d_1(t_1+t_2)}\int_0^{t_1}e^{d_1 t'}dt'\int_0^{t_2}e^{d_1 t''}dt''\,\langle\Psi(t')\Psi(t'')\rangle \qquad (1.66)\end{aligned}$$

as $\langle\Psi\rangle = 0$. Then, from (1.60), we have

$$\begin{aligned}&\langle N(t_1)N(t_2)\rangle \\ &= N_s^2 + 2d_1N_s(1+\frac{v_1}{d_0})e^{-d_1(t_1+t_2)}\int_0^{t_1}dt'\int_0^{t_2}dt''e^{d_1(t'+t'')}\delta(t'-t'') \qquad (1.67)\end{aligned}$$

It helps to be a little careful when carrying out the double integral over the delta function, as we need to determine when t' is equal to t'' which will depend on the range of the two integrals. Let's assign $t_1 \geq t_2$, then the double integral can be decomposed into

$$\begin{aligned}\int_0^{t_1}dt'\int_0^{t_2}dt'' &= \left(\int_{t_2}^{t_1}dt' + \int_0^{t_2}dt'\right)\int_0^{t_2}dt'' \\ &= \int_{t_2}^{t_1}dt'\int_0^{t_2}dt'' + \int_0^{t_2}dt'\int_0^{t_2}dt'' \qquad (1.68)\end{aligned}$$

where we can now explicitly see that for the first term $t' > t''$ (and there will be no contribution from the delta function), while for the second t' can equal t''

(and the delta function will contribute). Therefore,

$$\begin{aligned}\int_0^{t_1} dt' \int_0^{t_2} dt'' e^{d_1(t'+t'')}\delta(t'-t'') & \\ &= \int_0^{t_1} dt' \int_0^{t_2} dt'' e^{d_1(t'+t'')}\delta(t'-t'') \\ &+ \int_0^{t_1} dt' \int_0^{t_2} dt'' e^{d_1(t'+t'')}\delta(t'-t'') \\ &= \int_0^{t_2} e^{2d_1 t'} dt' \\ &= \frac{1}{2d_1}(e^{2d_1 t_2}-1) \end{aligned} \tag{1.69}$$

as the first integral evaluates to zero.

Consequently, (1.67)becomes

$$\begin{aligned}\langle N(t_1)N(t_2)\rangle - N_s^2 &= 2d_1 N_s \left(1+\frac{v_1}{d_0}\right) e^{-d_1(t_1+t_2)} \frac{1}{2d_1}\left(e^{2d_1 t_2}-1\right) \\ &= N_s \left(1+\frac{v_1}{d_0}\right)\left(e^{-d_1(t_1-t_2)} - e^{-d_1(t_1+t_2)}\right) \end{aligned} \tag{1.70}$$

As $\langle N(t)\rangle = N_s$, we have, finally,

$$\langle N(t_1)N(t_2)\rangle - \langle N(t_1)\rangle \langle N(t_2)\rangle = N_s \left(1+\frac{v_1}{d_0}\right)\left(e^{-d_1(t_1-t_2)} - e^{-d_1(t_1+t_2)}\right) \tag{1.71}$$

Eq. (1.71) is the auto-correlation function for protein number, and, after

long times $t_1 > t_2 \gg 1$, becomes

$$C_N = N_s \left(1 + \frac{v_1}{d_0}\right) e^{-d_1(t_1 - t_2)} \tag{1.72}$$

and so the protein auto-correlation time can be read off as $1 \setminus d_1$. Notice that eq. (1.61) has the same structure as the equation for mRNA

$$\frac{dM}{dt} = v_0 - d_0 M + \xi_1(t) \tag{1.73}$$

i.e. a constant rate of production and first-order degradation. Therefore, the solution of (1.73) should be of the same form as (1.72) but with d_1 replaced by d_0 and the magnitude of the noise term being given by (1.53) rather than (1.60). This substitution implies that

$$C_M = M_s e^{-d_0(t_1 - t_2)} \tag{1.74}$$

and so that the auto-correlation time of the mRNA is given by $1 \setminus d_0$.

When $t_1 = t_2$, the auto-correlation becomes the variance and so we can calculate the noise in mRNA levels as

$$\eta_M^2 = \frac{\langle M(t)^2 \rangle - \langle M(t) \rangle^2}{\langle M(t) \rangle^2}$$

$$\begin{aligned} &= \frac{M_S}{M_s^2} \\ &= \frac{1}{\langle M \rangle} \end{aligned} \tag{1.75}$$

Eqs. (1.74) and (1.75) are the general solutions to a birth-and-death (Poisson) model, and correspond to the expressions given (without proof) in eqs. (1.33) and (1.34).

The protein noise is a little more complicated and satisfies

$$\begin{aligned} \eta_N^2 &= \frac{1}{N_s} + \frac{v_1}{d_0}\frac{1}{N_s} \\ &= \frac{1}{N_s} + \frac{d_1}{d_0}\frac{1}{M_s} \\ &= \frac{1}{\langle N \rangle} + \frac{d_1}{d_0}\frac{1}{\langle M \rangle} \end{aligned} \tag{1.76}$$

This result should be compared to that of the simple model of Figure (1.6), given by (1.34). The effect of the mRNA, which acts as a fluctuating source of proteins, is to increase the noise above the value for a Poisson model. Eq. (1.76) can be described as

$$(\text{protein noise})^2 = (\text{Poisson noise})^2 + \frac{\text{mRNA lifetime}}{\text{protein lifetime}} \times (\text{mRNA noise})^2 \tag{1.77}$$

and so the Poisson noise is augmented by a time average of the noise in

the mRNA. As the protein life-time increases relative to the mRNA life-time, each protein is able to average over more mRNA fluctuations, and so lower the protein noise, with η_N becomes closer and closer to the Poisson result as $d_1 \setminus d_0 \to 0$.

1.15 Typical numbers

Some typical numbers for constitutive (unregulated) expression in *E. coli* are

$$\begin{aligned} d_1 = 1 \setminus \text{hour} \quad &; \quad d_0 = 1 \setminus 3 \text{ minutes} \\ \langle N \rangle = 10^3 \quad &; \quad \langle M \rangle = 5 \end{aligned} \tag{1.78}$$

and so (1.76) becomes

$$\begin{aligned} \eta_N^2 &= 1 \setminus 1000 = 3 \setminus 60 \cdot 1 \setminus 5 \\ &= 0.001 + 0.01 \end{aligned} \tag{1.79}$$

The mRNA term can therefore be seen to set the overall magnitude of the noise, which is believed to generally true for most (constitutively expressed) genes in *E. coli.*

Eq. (1.76)can be re-written using (1.45) as

$$\eta_N^2 = \frac{d_1}{v_1 M_s} + \frac{d_1}{d_0} \cdot \frac{1}{M_s} \tag{1.80}$$

and only the first term can be seen to contain the parameter v_1 associated with translation. Therefore, transcription is the dominant source of noise if

$$\frac{d_1}{d_0} \cdot \frac{1}{M_s} \gg \frac{d_1}{v_1 M_s} \tag{1.81}$$

i.e. if the second term (arising from transcription) in (1.80) is greater than the first (which arises from transcription and translation). Eq. (1.81) simplifies to

$$v_1 \ll d_0 \tag{1.82}$$

Ribosomes are believed to translate at a rate of around 40 nt s^{-1} [7] and so for a 1000 nt protein, v_1 satisfies

$$\frac{1}{v_1} = \frac{1000 \text{ nt}}{40 \text{ nt } s^{-1}} \tag{1.83}$$

and so is $v_1 \simeq 0.04s^{-1}$ Eq. (1.82) then becomes

$$0.04 \gg \frac{1}{3 \times 60} \simeq 0.006 \tag{1.84}$$

which is certainly true and so transcription, rather than translation, is normally the dominant source of noise.

Appendix 1: Dirac delta Function

The Dirac delta function can be thought of as the limit of a normal distribution with mean zero as the standard deviation of that distribution also tends to zero

$$\delta(x) = \lim_{n\to\infty} \frac{n}{\sqrt{\pi}} e^{-n^2x^2} \tag{85}$$

This limit leads to a function, whose integral over all x is fixed at unity, but that becomes increasingly more and more spiked at zero (see Figure. (11)).

Ultimately

$$\delta(x) = 0 \text{ for all } x \neq 0 \tag{86}$$

and is not strictly at x = 0 but does retain the property

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{87}$$

These two characteristics imply that integrating the product of a delta function and another function, $f(x)$, will only give a non-zero result at $x = 0$ and so effectively selects the value $f(0)$,

$$\int_{-\infty}^{\infty} f(x)\delta(x)dx = f(0) \tag{88}$$

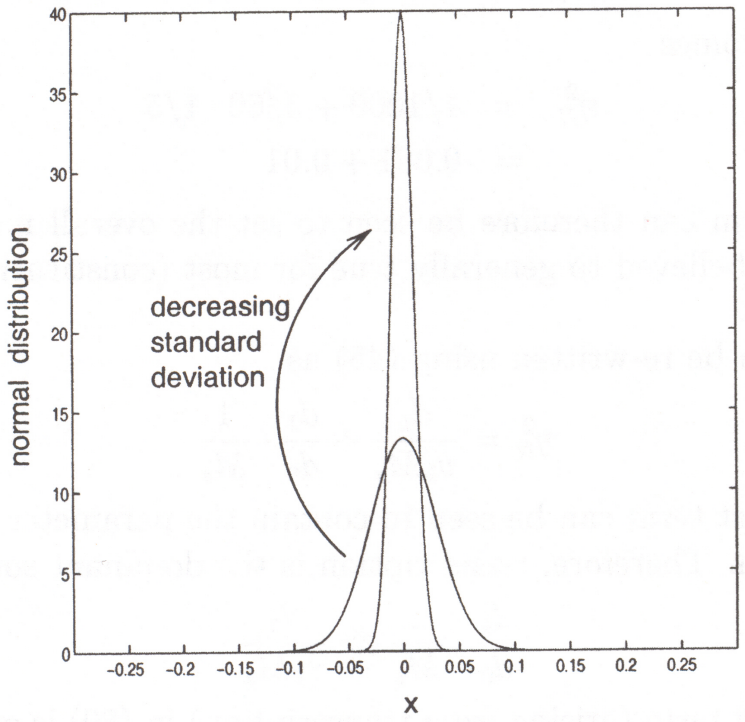

Figure 11: *The Dirac delta function is the 'spike' limit of a normal distribution as its standard deviation tends to zero.*

or, more generally,

$$\int_{-\infty}^{\infty} f(x-y)\delta(x)dx = f(y) \tag{89}$$

Appendix 2: Sampling from a probability distribution

Often, in simulation, we wish to sample from a particular probability distribution, $P(x)$ say. If we define the cumulative distribution of $P(x)$ by

$$F(x) = \int_{x_{min}}^{x} P(x')dx' \tag{90}$$

then

$$\begin{aligned} \mathcal{P}(x \leq x_0) &= \int_{x_{min}}^{x} P(x')dx' \\ &= F(x_0) \end{aligned} \tag{91}$$

A sketch of the typical behaviour of $F(x)$ is shown in Figure. (12). Notice that if x_0 then $F(x) \leq F(x_0)$ as $F(x)$, by definition, must be a monotonic function.

To sample from $P(x)$, first let y be a uniform random number (easily obtained on a computer) such that $0 \leq y \leq 1$, then

$$\mathcal{P}(y \leq y_0) = \int_{0}^{y_0} dy' = y_0 \tag{92}$$

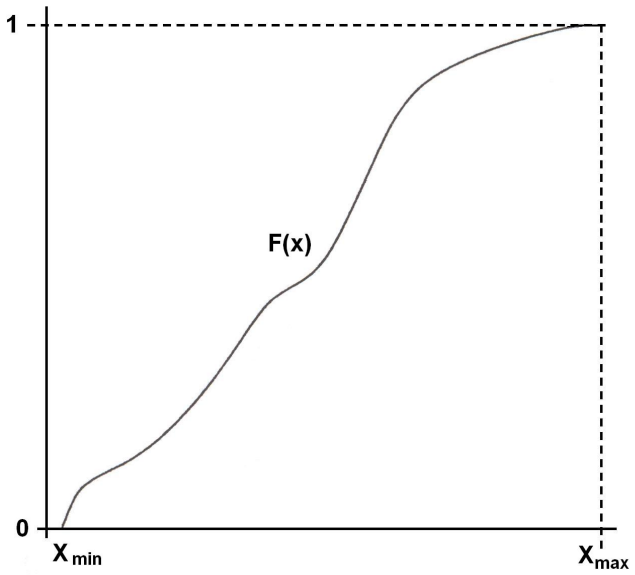

Figure 12: *A typical plot of cumulative frequency versus x*

for some $0 \leq y_0 \leq 1$. Define

$$x = F^{-1}(y) \tag{93}$$

where $F(x)$ is the cumulative frequency of $P(x)$. Consequently,

$$\begin{aligned} \mathcal{P}(x \leq x_0 &= \mathcal{P}\left(F^{-1}(y) \leq x_0\right) \\ &= \mathcal{P}\left(F \cdot F^{-1}(y) \leq F(x_0)\right) \end{aligned} \tag{94}$$

given that $F(x)$ is monotonic. As $F \cdot F{-}1(y) = y$, we have

$$\begin{aligned} \mathcal{P}(x \leq x_0 &= \mathcal{P}\left(y \leq F(x_0)\right) \\ &= F(x_0) \end{aligned} \tag{95}$$

as y is a sample, between 0 and 1, from the uniform distribution, see (92). Thus, x as defined in (93) obeys (91) and so is a sample from $P(x)$. Therefore, if we can calculate the inverse function of the cumulative frequency of a distribution $P(x)$, then acting this inverse function on a sample from the uniform distribution provides a sample from $P(x)$.

Bibliography

[1] Elowitz, M.B., Levine, A.J., Siggia, E.D. & Swain, P.S. (2002) *Science* **297**, 1183-1186.

[2] Swain, P.S., Elowitz, M.B., & Siggia, E.D. (2002) *Proc. Natl. Acad. USA* **99**, 12795-12800.

[3] Gillespie, D.T. (1977) *J. Phys. Chem.* **81**, 2340-2361.

[4] Van Kampen, N.G. (2001) *Stochastic Processes in Physics and Chemestry* (Elsevier Science, New York, New York).

[5] Gardiner, C.W. (2004) *Handbook of Stochastic Processes* (Springer Verlag, New York, New York).

[6] Lauffenburger, D.A. & Linderman, J.L. (1996) *Receptors: Models for Binding, Trafficking, and Signalling* (Oxford, U.K.).

[7] Bremer, H. & Dennis, P.P. (1996) in *Escherichia coli and Salmonella: Cellular and Molecular Biology,* ed. Neidhardt, F.C. (Am. Soc. Microbiol., Washington, D.C.).