Lecture notes on stochastic models in systems biology

Peter S. Swain

peter.swain@ed.ac.uk

Biological Sciences, University of Edinburgh

Abstract

These notes provide a short, focused introduction to modelling stochastic gene expression, including a derivation of the master equation, the recovery of deterministic dynamics, birth-and-death processes, and Langevin theory. The notes were last updated around 2010 and written for lectures given at summer schools held at McGill University's Centre for Non-linear Dynamics in 2004, 2006, and 2008.

Introduction

A system evolves stochastically if its dynamics is partly generated by a force of random strength or by a force at random times or by both. For stochastic systems, it is not possible to exactly determine the state of the system at later times given its state at the current time. Instead, to describe a stochastic system, we use the probability that the system is in a certain state and can predict how this probability changes with time. Calculating this probability is often difficult, and we usually focus on finding the moments of the probability distribution, such as the mean and variance, which are commonly measured experimentally.

Any chemical reaction is stochastic. Reactants come together by diffusion, their motion driven by collisions with other molecules. Once together, these same collisions alter the internal energies of the reactants, and so their propensity to react. Both effects cause individual reaction events to occur randomly.

Is stochasticity important in biology? Intuitively, stochasticity is only significant when typical numbers of molecules are low. Then individual reactions, which at most change the numbers of molecules by one or two, matter. Low numbers are frequent *in vivo*: gene copy number is typically one or two, and transcription factors often number in the tens, at least in bacteria. There are now many reviews on biochemical stochasticity[1, 2, 3, 4].

Unambiguously measuring stochastic gene expression, however, can be challenging [5]. Naively, we could place Green Fluorescent Protein (GFP) on a bacterial chromosome downstream of a promoter that is activated by the system of interest. By measuring the variation in fluorescence across a population of cells, we could quantify stochasticity. Every biochemical reaction, however, is potentially stochastic. Fluorescence variation could be because of stochasticity in the process under study or could result from the general background 'hum' of stochasticity: stochastic effects in ribosome synthesis could lead to different numbers of ribosomes and so to differences in gene expression in each cell; stochastic effects in the cell cycle machinery may desynchronize the population; stochastic effects in signaling networks could cause each cell to respond uniquely, and so on.

Variation has then two classes: **intrinsic stochasticity**, the stochasticity inherent in the dynamics of the system and that arises from fluctuations in the timing of individual reactions, and **extrinsic stochasticity**, the stochasticity originating from reactions of the system of interest with other stochastic systems in the cell or its environment [6, 5]. In principle, intrinsic and extrinsic stochasticity can be measured by creating a copy of the network of interest in the same cellular environment as the original network [5]. We can define intrinsic and extrinsic variables for the system of interest, with fluctuations in these variables together generating intrinsic and extrinsic stochasticity [6]. The intrinsic variables of a system will typically specify the copy numbers of the molecular components of the system. For gene expression, the level of occupancy of the promoter by transcription factors, the numbers of mRNA molecules, and the number of proteins are all intrinsic variables. Imagining a second copy of the system – an identical gene and promoter elsewhere in the genome – then the instantaneous values of the intrinsic variables of this copy of the system will usually differ from those of the original system. At any point in time, for example, the number of mRNAs transcribed from the first copy of the gene will usually be different from the number of mRNAs transcribed from the second copy. Extrinsic variables, however, describe processes that equally affect each copy of the system. Their values are therefore the same for each copy. For example, the number of cytosolic RNA polymerases is an extrinsic variable because the rate of gene expression from both copies of the gene will increase if the number of cytosolic RNA polymerases increases and decrease if the number of cytosolic RNA polymerases decreases. In contrast, the number of transcribing RNA polymerases is an intrinsic variable because we expect the number of transcribing RNA polymerases to be different for each copy of the gene at any point in time.

Stochasticity is quantified by measuring an intrinsic variable for both copies of the system. For gene expression, the number of proteins is typically measured by using fluorescent proteins as markers [7, 5, 8, 9]. Imaging a population of cells then allows estimation of the distribution of protein levels at steady-state. Fluctuations of the intrinsic variable will *in vivo* have both intrinsic and extrinsic sources. The number of proteins will fluctuate because of intrinsic stochasticity generated during gene expression, but also because of stochasticity in, for example, the number of cytosolic RNA polymerases or ribosomes or proteosomes. We will use the term 'noise' to mean an empirical measure of stochasticity defined by the coefficient of variation (the standard deviation divided by the mean) of a stochastic process. An estimate of intrinsic stochasticity is the intrinsic noise which is defined as a measure of the difference between the value of an intrinsic variable for one copy of the system and its counterpart in the second copy. For gene expression, typically the intrinsic noise is the mean absolute difference (suitably normalized) at steady-state between the number of proteins expressed from one copy of the gene and the number of proteins expressed from the other copy [5]. Such a definition supports the intuition that intrinsic fluctuations cause variation in one copy of the system to be uncorrelated with variation in the other copy. Extrinsic noise is defined as the correlation coefficient between the intrinsic variable of one copy of the system and its counterpart for the other copy because extrinsic fluctuations equally affect both copies of the system and consequently cause correlations between variation in one copy and variation in the other. The intrinsic and extrinsic noise should be related to the coefficient of variation of the intrinsic variable of the original system of interest. This so-called total noise is given by the square root of the sum of the squares of the intrinsic and the extrinsic noise [6].

Such two-colour measurements of stochasticity have been applied to bacteria and yeast where gene expression has been characterized by using two copies of a promoter placed in the genome with each copy driving a distinguishable allele of Green Fluorescent Protein [5, 9]. Both intrinsic and extrinsic noise can be substantial giving, for example, a total noise of around 0.4, and so the standard deviation of protein numbers is 40% of the mean. Extrinsic noise is usually higher than intrinsic noise. There are some experimental caveats: both copies of the system should be placed 'equally' in the genome so that the probabilities of transcription and replication are equal. This 'equality' is perhaps best met by placing the two genes adjacent to each other [5]. Although conceptually there are no difficulties, practically problems arise with feedback. If the protein synthesized in one system can influence its own expression, the same protein will also influence expression in a second copy of the system. The two copies of the system have lost the (conditional) independence they require to be two simultaneous measurements of the same stochastic process.

A stochastic description of chemical reactions

For any network of chemical reactions, the lowest level of description commonly used in systems biology is the chemical master equation. This equation assumes that the system is well-stirred and so ignores spatial effects. It governs how the probability of the system being in any particular state changes with time. A system state is defined by the number of molecules present for each chemical species, and it will change every time a reaction occurs. From the master equation we can derive the deterministic approximation (a set of coupled differential equations) which is often used to describe system dynamics. The dynamics of the mean of each chemical species approximately obeys these deterministic equations as the numbers of molecules of all species increase [10, 11]. The master equation itself is usually only solvable analytically for linear systems: systems having only first-order chemical reactions.

Nevertheless, several approximations exist, all of which exploit the tendency of fluctuations to decrease as the numbers of molecules increase. The most systematic is the **linear noise** approach of van Kampen [12]. If the concentration of each chemical species is fixed, then changing the system volume, Ω, alters the number of molecules of every chemical species. The linear noise approximation is based on a systematic expansion of the master equation in the inverse of the system volume, Ω^{-1}. It leads to diffusion-like equations that accurately describe small fluctuations around any stable attractor of the system. For systems that tend to steady-state, a **Langevin** approach is also often used [13, 14, 15]. Here additive, white stochastic terms are included in the deterministic equations, with the magnitude of these terms being determined by the chemical reactions. At steady-state and for sufficiently high numbers of molecules, the Langevin and linear noise approaches are equivalent.

Unfortunately, all these methods become intractable, in general, once the number of chemical species in the system reaches more than three (we then need to analytically calculate the inverse of at least a 4×4 matrix or its eigenvalues). Rather than numerically solve the master equation, the **Gillespie algorithm** [16], a Monte Carlo method, is often used to simulate intrinsic fluctuations by generating one sample time course from the master equation. By doing many simulations and averaging, the mean and variance for each chemical species can be calculated as a function of time. Extrinsic fluctuations can be modelled as fluctuations in the parameters of the system, such as the kinetic rates [17, 18]. They can be included by a minor modification of the Gillespie algorithm that feeds in a pre-simulated time series of extrinsic fluctations and so generates both intrinsic and extrinsic fluctuations [18].

Here we will introduce the master equation and briefly discuss the Gillespie algorithm.

The master equation

Once molecules can react, the intrinsic stochsasticity destroys any certainty of the numbers and types of molecules present, and we must adopt a probabilistic description. For example, a model of gene expression is given by

$$\xrightarrow{k} \mathrm{C} \xrightarrow{d} 0$$

where protein C is synthesized on average every $1/k$ seconds and degrades on average every $1/d$ seconds. The reactions can be described by the probability

$$\mathcal{P}(n \text{ molecules of } C \text{ at time } t)$$

and how this probability evolves with time. Each reaction rate is interpreted as the probability per unit time of the appropriate reaction.

We will write $P_n(t)$ for the probability that n proteins exist at time t and consider the reactions that might have occurred just prior to having n molecules of protein. Let δt be a time interval small enough so that at most only one reaction can occur. If there are n proteins at time $t + \delta t$, then if a protein was synthesized during the interval δt, there must have been $n - 1$ proteins at time t. The probability of synthesis is

$$\mathcal{P}(\text{synthesis}) = k\delta t \tag{1}$$

which is independent of the number of proteins present. If we have n proteins at time $t + \delta t$ and a protein was degraded during the interval δt, however, there must have been $n + 1$ proteins at time t. The probability of degradation is

$$\mathcal{P}(\text{degradation}) = (n + 1)d\delta t. \tag{2}$$

Neither synthesis nor degradation may have occurred during δt. The number of proteins will be unchanged, which occurs with probability

$$\mathcal{P}(\text{no reaction}) = 1 - k\delta t - nd\delta t. \tag{3}$$

Notice that the probability of a protein degrading is $nd\delta t$ because n proteins must have existed at time t.

Putting these probabilities together, we can the master equation describing the time evolution of $P_n(t)$. Writing

$$P_n(t + \delta t) = P_{n-1}(t)k\delta t + P_{n+1}(t)d(n + 1)\delta t + P_n(t)(1 - k\delta t - nd\delta t). \tag{4}$$

dividing through by δt and taking the limit $\delta t \to 0$ gives

$$\frac{\partial}{\partial t} P_n = k\Big[P_{n-1} - P_n\Big] - d\Big[nP_n - (n + 1)P_{n+1}\Big] \tag{5}$$

Eq. 5 is an example of a master equation: all the moments of the probability distribution $P_n(t)$ can be derived from it.

Consider now a binary reaction:

$$\mathrm{A} + \mathrm{B} \xrightarrow{f} \mathrm{C} \tag{6}$$

where A and B bind irreversibly to form complex C with probability f per unit time. Suppose further that individual C molecules degrade with probability d per unit time

$$\mathrm{C} \xrightarrow{d} 0$$

The state of the system is then described by

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

which we will write as $P_{n_A,n_B,n_C}(t)$. We again consider a time interval δt small enough so that at most only one reaction can occur. If the system at time $t + \delta t$ has n_A, n_B, and n_C molecules of A, B, and C, 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. The probability of this reaction is

$$\mathcal{P}(f \text{ reaction}) = f(n_A + 1)(n_B + 1)\delta t. \tag{7}$$

Alternatively, reaction d could have occurred during δt and so the system then must have been in the state n_A, n_B, and $n_C + 1$ at time t. Its probability is

$$\mathcal{P}(d \text{ reaction}) = d(n_C + 1)\delta t. \tag{8}$$

Finally, no reaction may have occurred at all, and so the system would be unchanged at t (in the state n_A, n_B, and n_C):

$$\mathcal{P}(\text{no reaction}) = 1 - f n_A n_B \delta t - d n_C \delta t. \tag{9}$$

Thus we can find the master equation by writing

$$\begin{aligned} &P_{n_A,n_B,n_C}(t + \delta t) = \\ &\quad P_{n_A+1,n_B+1,n_C-1}(t)(n_A + 1)(n_B + 1) f\delta t + P_{n_A,n_B,n_C+1}(t)(n_C + 1) d\delta t \\ &\qquad\qquad + P_{n_A,n_B,n_C}(t)\Big[1 - n_A n_B f \delta t - n_C d \delta t\Big] \end{aligned} \tag{10}$$

or

$$\begin{aligned} \frac{\partial}{\partial t} P_{n_A,n_B,n_C} &= f\Big[(n_A + 1)(n_B + 1) P_{n_A+1,n_B+1,n_C-1} - n_A n_B P_{n_A,n_B,n_C}\Big] \\ &\quad - d\Big[n_C P_{n_A,n_B,n_C} - (n_C + 1) P_{n_A,n_B,n_C+1}\Big] \end{aligned} \tag{11}$$

in the limit of $\delta t \to 0$.

The definition of noise

Noise is typically defined as the coefficient of variation: the ratio of the standard deviation of a distribution to its mean. We will denote noise by η:

$$\eta = \frac{\sqrt{\langle N^2 \rangle - \langle N \rangle^2}}{\langle N \rangle} \tag{12}$$

for a random variable N. The noise is dimensionless and measures the magnitude of a typical fluctuation as a fraction of the mean.

Example: A birth-and-death processes

The model of gene expression

$$\xrightarrow{k} \mathrm{C} \xrightarrow{d} 0 \tag{13}$$

is a birth-and-death process. Proteins can only be synthesized (born) or degrade (die). We will solve the master equation for this system, Eq. 5, using a moment generating function.

The moment generating function for a probability distribution $P_n(t)$ is defined as

$$F(z,t) = \sum_{n=0}^{\infty} z^n P_n(t) \tag{14}$$

and can be thought of as a discrete transform. Differentiating the moment generating function with respect to z gives

$$\begin{aligned} \frac{\partial F}{\partial z} &= \sum_{n=0}^{\infty} n z^{n-1} P_n & (15) \\ \frac{\partial^2 F}{\partial z^2} &= \sum_{n=0}^{\infty} n(n-1) z^{n-2} P_n. & (16) \end{aligned}$$

The generating function and its derivatives have useful properties because of their dependence on the probability distribution $P_n(t)$:

$$\begin{aligned} F(z=1,t) &= \sum_{n=0}^{\infty} P_n(t) = 1 & (17) \\ \frac{\partial F}{\partial z}(z=1,t) &= \sum_{n=0}^{\infty} n P_n(t) = \langle n(t) \rangle & (18) \\ \frac{\partial^2 F}{\partial z^2}(z=1,t) &= \sum_{n=0}^{\infty} n(n-1) P_n(t) = \langle n^2(t) \rangle - \langle n(t) \rangle. & (19) \end{aligned}$$

Finding $F(z,t)$ therefore allows us to calculate all the moments of $P_n(t)$: $F(z,t)$ is called the moment generating function.

The master equation can be converted into a partial differential equation for the moment generating function. Multiplying (5) by z^n and summing over all n gives

$$\begin{aligned} \frac{\partial F}{\partial t} &= k \sum_n z^n P_{n-1} - kF - d \sum_n n z^n P_{n-1} + d \sum_n (n+1) z^n P_{n+1} \\ &= kz \sum_n z^{n-1} P_{n-1} - kF - dz \sum_n n z^{n-1} P_n + d \sum_n (n+1) z^n P_{n+1} & (20) \end{aligned}$$

where we have factored z out of some of the sums so that we can use (14) and (15). With these results and setting $P_n = 0$ if $n < 0$, we can write

$$\frac{\partial F}{\partial t} = kzF - F - dz\frac{\partial F}{\partial z} + d\frac{\partial F}{\partial z} \tag{21}$$

or

$$\frac{\partial F}{\partial t} = (z-1)\left(kF - d\frac{\partial F}{\partial z}\right). \tag{22}$$

This first order partial differential equation can be solved in general using the method of characteristics [12].

We will solve (22) to find the steady-state probability distribution of protein numbers. At steady-state, $P_n(t)$ is independent of time and so $\frac{\partial F}{\partial t} = 0$ from (14). Consequently, (22) becomes

$$\frac{\partial F}{\partial z} = \frac{k}{d}F \tag{23}$$

which is an ordinary differential equation. This equation has a solution

$$F(z) = C\mathrm{e}^{\frac{k}{d}z} \tag{24}$$

for some constant C. This constant can be determined from (17), implying

$$F(z) = \mathrm{e}^{\frac{k}{d}(z-1)}. \tag{25}$$

By differentiation (25) with respect to z and using (18) and (19), the moments of n can be calculated. For this case, we can Taylor expand (25) and find the probability distribution P_n by comparing the expansion with (14). Expanding gives

$$F(z) = \mathrm{e}^{-\frac{k}{d}} \sum_{n=0}^{\infty} \frac{(k/d)^n}{n!} z^n \tag{26}$$

implying that the steady-state probability of having n proteins is

$$P_n = \mathrm{e}^{-k/d}\frac{(k/d)^n}{n!} \tag{27}$$

which is a Poisson distribution. The first two moments are

$$\begin{aligned} \langle n \rangle &= k/d \\ \langle n^2 \rangle - \langle n \rangle^2 &= k/d \;=\; \langle n \rangle \end{aligned} \tag{28}$$

and consequently the noise is

$$\eta = 1/\sqrt{\langle n \rangle} \tag{29}$$

from (12).

Eq. (29) demonstrates a ‘rule-of-thumb’: stochasticity generally become more significant as the number of molecules in the system decrease (Fig. 1). Approximate expression for the distribution of proteins now exist for more realistic models of gene expression [19, 20].

Recovering the deterministic equations

Solving the master equation is possible for linear systems, i.e. those with only first-order chemical reactions, but often only at steady-state [12, 21]. Solving for the moments of a master equation is often easier.

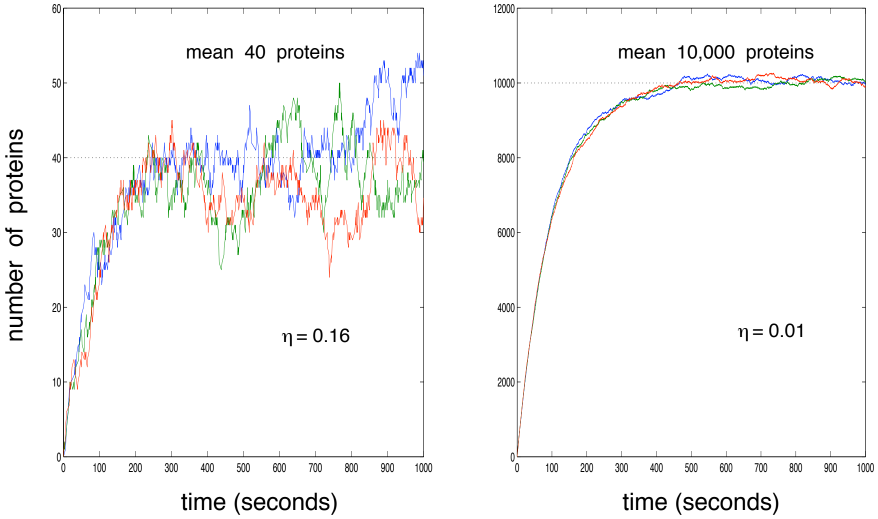

Figure 1: Three simulation runs of two birth-and-death models of gene expression (Eq. 13). Each model has different rate constants leading to different mean protein levels.

For the non-linear system of Eq. 6, we will use the master equation, (11), to derive the equation of motion for the mean of C. The mean of C is defined as

$$\langle C(t) \rangle = \sum_{n_A, n_B, n_C} n_C P_{n_A, n_B, n_C}(t) \tag{30}$$

and is a function of time.

Multiplying (11) by n_C and summing over n_A, n_B, and n_C gives

$$\begin{aligned} \frac{\partial}{\partial t} \langle C \rangle &= f \sum (n_C - 1 + 1)(n_A + 1)(n_B + 1) P_{n_A+1, n_B+1, n_C-1} \\ &\quad - f \sum n_A n_B n_C P_{n_A, n_B, n_C} - d \sum n_C^2 P_{n_A, n_B, n_C} \\ &\quad + d \sum (n_C + 1 - 1)(n_C + 1) P_{n_A, n_B, n_C+1} \end{aligned} \tag{31}$$

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

$$\begin{aligned} \langle ABC \rangle &= \sum_{n_A, n_B, n_C = 0}^{\infty} n_A n_B n_C P_{n_A, n_B, n_C} \\ &= \sum_{n_A, n_B, n_C = 0}^{\infty} (n_A + 1)(n_B + 1)(n_C - 1) P_{n_A+1, n_B+1, n_C-1} \end{aligned} \tag{32}$$

as $P_{n_A, n_B, n_C}(t)$ is zero if any of n_A, n_B, or n_C are negative, we have

$$\begin{aligned} \frac{\partial}{\partial t} \langle C \rangle &= f \Big[\langle ABC \rangle + \langle AB \rangle \Big] - f \langle ABC \rangle - d \langle C^2 \rangle + d \Big[\langle C^2 \rangle - \langle C \rangle \Big] \\ &= f \langle AB \rangle - d \langle C \rangle \end{aligned} \tag{33}$$

which is the microscope equation for the dynamics of the mean of C.

We can also consider the deterministic equation for the dynamics. Applying the law of mass action to this system, the concentration of C, $[C]$, obeys

$$\frac{d}{dt}[C] = \tilde{f}[A][B] - \tilde{d}[C] \tag{34}$$

where $\tilde{f}$ and $\tilde{d}$ are the macroscopic (deterministic) rate constants. The macroscopic concentration is related to the mean number of molecules by

$$[C] = \frac{\langle C \rangle}{V} \tag{35}$$

and so the deterministic equations are equations for the rate of change of the means of the different chemical species: using (35), (34) becomes

$$\frac{d}{dt}\langle C \rangle = \frac{\tilde{f}}{V}\langle A \rangle \langle B \rangle - \tilde{d}\langle C \rangle. \tag{36}$$

By comparing the deterministic equation, (36), with the microscopic equation, (33), we can relate the stochastic probabilities of reaction per unit time and the deterministic kinetic rates:

$$\begin{aligned} \tilde{f} &= \frac{V\langle AB \rangle}{\langle A \rangle \langle B \rangle} \cdot f \\ \tilde{d} &= d \end{aligned} \tag{37}$$

For first-order reactions both the kinetic rate and the probability are the same. The macroscopic rate $\tilde{f}$ is usually measured under conditions where the deterministic approximation holds and numbers of molecules are large. We can write

$$\begin{aligned} \tilde{f} &= \frac{V\Big(\langle A \rangle \langle B \rangle + \langle AB \rangle - \langle A \rangle \langle B \rangle\Big)}{\langle A \rangle \langle B \rangle} \cdot f \\ &= Vf \cdot \left(1 + \frac{\langle AB \rangle - \langle A \rangle \langle B \rangle}{\langle A \rangle \langle B \rangle}\right) \\ &\simeq Vf \end{aligned} \tag{38}$$

where the fluctuation term becomes negligible as the numbers of molecules increase because its numerator, the co-variance $\langle AB \rangle - \langle A \rangle \langle B \rangle$, is expected to be proportional to the mean number of molecules, while its denominator is proportional to the square of the mean number of molecules. Eq. (28) is an explicit example of this statement. Eq. (38) is almost always used to relate the macroscopic rate and the probability of reaction for second-order reactions.

An exception: homo-dimerization reactions

A homo-dimerization reaction

$$\mathrm{A} + \mathrm{A} \xrightarrow{f} \mathrm{A}_2$$

occurs when two identical monomers combine to form a dimer. This reaction is common among transcription factors. The master equation is now

$$\frac{\partial P_{n_A}}{\partial t} = f\left[\left(\begin{array}{c} n_A + 2 \\ 2 \end{array}\right) P_{n_A+2} - \left(\begin{array}{c} n_A \\ 2 \end{array}\right) P_{n_A}\right] \tag{39}$$

where each coefficient is the number of ways of forming a dimer. Eq. (37) becomes

$$2\frac{\tilde{f}}{V}\langle A\rangle^2 = f\langle A(A-1)\rangle. \tag{40}$$

Assuming that $\tilde{f}$ is measured for large numbers of molecules, we can write

$$\langle A(A-1)\rangle \simeq \langle A\rangle^2 \tag{41}$$

and so to

$$\tilde{f} \simeq \frac{fV}{2} \tag{42}$$

which is the inter-conversion formula for dimerization reactions.

Simulating stochastic biochemical reactions

The Gillespie algorithm [16] is most commonly used to simulate intrinsic fluctuations in biochemical systems. The equivalent of two dice are rolled on the computer: one to choose which reaction will occur next and the other to choose when that reaction will occur. Assume that we have a system in which n different reactions are possible, then the probability that starting from time t a reaction only occurs between $t+\tau$ and $t+\tau+\delta\tau$ must be calculated for each reaction. Let this probability be $P_i(\tau)\delta\tau$ for reaction i, say.

For example, if reaction i corresponds to the second-order reaction of Eq. 6, then

$$\begin{aligned}\mathcal{P}(\text{reaction } i \text{ in time } \delta\tau) &= n_A n_B f \delta\tau \\ &= a_i \delta\tau\end{aligned} \tag{43}$$

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

$$\begin{aligned}P_i(\tau)\delta\tau &= \mathcal{P}(\text{no reaction for time } \tau) \\ &\quad \times\mathcal{P}(\text{reaction } i \text{ happens in time } \delta\tau) \\ &\equiv P_0(\tau)a_i\delta\tau\end{aligned} \tag{44}$$

with $P_0(\tau)$ the probability that no reaction occurs during the interval τ. This probability is the product of the probability of having no reactions at time τ and the probability of no reactions occurring in time $\delta\tau$:

$$P_0(\tau+\delta\tau) = P_0(\tau)\Big[1 - \sum_{j=1}^{n} a_j \delta\tau\Big] \tag{45}$$

which implies

$$\frac{dP_0}{d\tau} = -P_0 \sum_{j=1}^{n} a_j \tag{46}$$

and so

$$P_0(\tau) = \exp\Big(-\tau \sum a_j\Big). \tag{47}$$

Thus we have

$$P_i(\tau) = a_i \mathrm{e}^{-\tau \sum a_j} \tag{48}$$

from (47).

To choose which reaction to simulate, an n-sided die is rolled with each side corresponding to a reaction and weighted by the reaction's propensity. A second die is then used to determine the time when the reaction occurs by sampling from (47). All the chemical species and the time variable are updated to reflect the occurrence of the reaction, and the process is then repeated. See Gillespie (1977) [16] for more details.

Extrinsic fluctuations can be included by considering reaction rates that change with time [18]. A reaction rate is often a function of the concentration of another protein and so fluctuates because this protein concentration fluctuates. For example, v_0 in Fig. 2 is a function of the concentration of free RNA polymerases and v_1 is a function of the concentration of free ribosomes. By simulating extrinsic fluctuations with the desired properties before running the Gillepsie algorithm and then approximating this extrinsic time series by a sequence of linear changes over small time intervals, we can 'feed' the extrinsic fluctuations into the Gillepsie algorithm and so let a parameter, or many parameters, fluctuate extrinsically.

Langevin theory: an improved model of gene expression

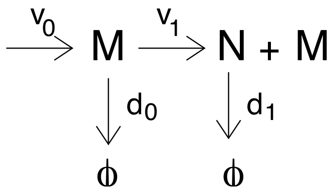

Figure 2: A model of gene expression that explicitly includes transcription (rate v_0) and translation (rate v_1) as first-order processes. mRNA is denoted by M and protein by N.

We can model transcription and translation as first-order reactions [22]. Both mRNA, M, and protein, N, are present, and each has their own half-life (determined by the inverse of their degradation rates).

The Langevin solution

Langevin theory gives an approximation to the solution of the master equation. It is strictly only valid when numbers of molecules are large. Stochastic terms are explicitly added to the deterministic equations of the system. For the model of Fig. 2, the deterministic equations are

$$\begin{aligned}\frac{dM}{dt} &= v_0 - d_0 M \\ \frac{dN}{dt} &= v_1 M - d_1 N. \qquad (49)\end{aligned}$$

A Langevin model adds a stochastic variable, $\xi(t)$, to each

$$\begin{aligned}\frac{dM}{dt} &= v_0 - d_0 M + \xi_1(t) \\ \frac{dN}{dt} &= v_1 M - d_1 N + \xi_2(t) \qquad (50)\end{aligned}$$

and is only fully specified when the probability distributions for the ξ_i are given. The ξ_i must be specified so that they mimic thermal fluctuations and model intrinsic fluctuations. 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).

To define ξ, we must give its mean and variance as functions of time and its autocorrelation.

Understanding stochasticity: autocorrelations

The autocorrelation time of a stochastic variable describes the average life-time of a typical fluctuation. We will denote it by τ. Fig. 3 shows typical behaviour of a stochastic variable obeying a Poisson distribution. Time has been rescaled by the autocorrelation time. On average, the number of molecules changes significantly only over a time τ (1 in these units).

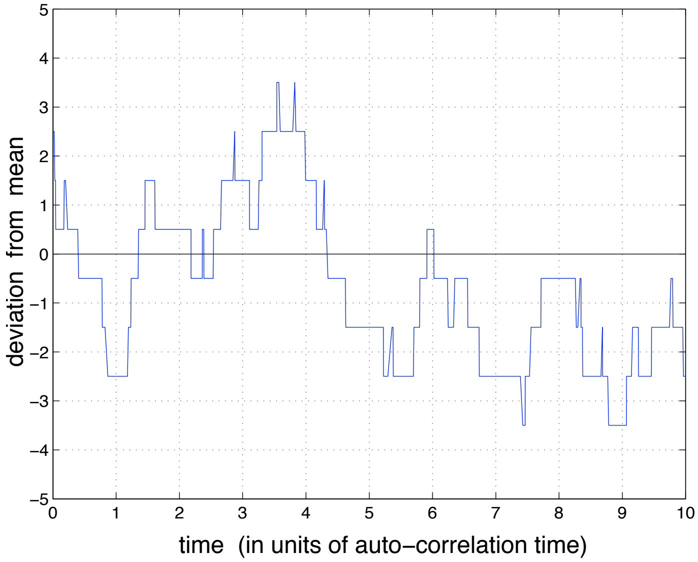

Figure 3: A time-series of a birth-death process. Time has been rescaled by the autocorrelation time. The deviation from the mean, $n - \langle n \rangle$, in numbers of molecules is plotted on the y-axis.

The autocorrelation time is found from the autocorrelation function. For a stochastic variable N, the autocorrelation function is

$$\begin{aligned} C_N(t_1, t_2) &= \Big\langle \Big[N(t_1) - \langle N(t_1) \rangle \Big] \Big[N(t_2) - \langle N(t_2) \rangle \Big] \Big\rangle \\ &= \Big\langle \Big\{ N(t_1)N(t_2) - \langle N(t_1) \rangle N(t_2) - N(t_1) \langle N(t_2) \rangle + \langle N(t_1) \rangle \langle N(t_2) \rangle \Big\} \Big\rangle \\ &= \langle N(t_1)N(t_2) \rangle - \langle N(t_1) \rangle \langle N(t_2) \rangle. \end{aligned} \tag{51}$$

It quantifies how a deviation of N away from its mean at time t_1 is correlated with the deviation from the mean at a later time t_2. It is determined by the typical life-time of a fluctuation. When $t_1 = t_2$, (51) is just the variance of $N(t)$.

Stationary processes are processes that are invariant under time translations and so are statistically identical at all time points. For a stationary process, such as the steady-state behaviour of a chemical system, the autocorrelation function obeys

$$C_N(t_1, t_2) = C_N(t_2 - t_1). \tag{52}$$

It is a function of one variable: the time difference between the two time points considered. Fig. 4 shows the steady-state autocorrelation function for the Poisson model of gene expression. It is normalized by the variance and is fit well by an exponential decay: $e^{-t/\tau}$. A typical fluctuation only persists for the timescale τ as enough new reaction events occur during τ to significantly change the dynamics and remove any memory the system may have had of earlier behaviour.

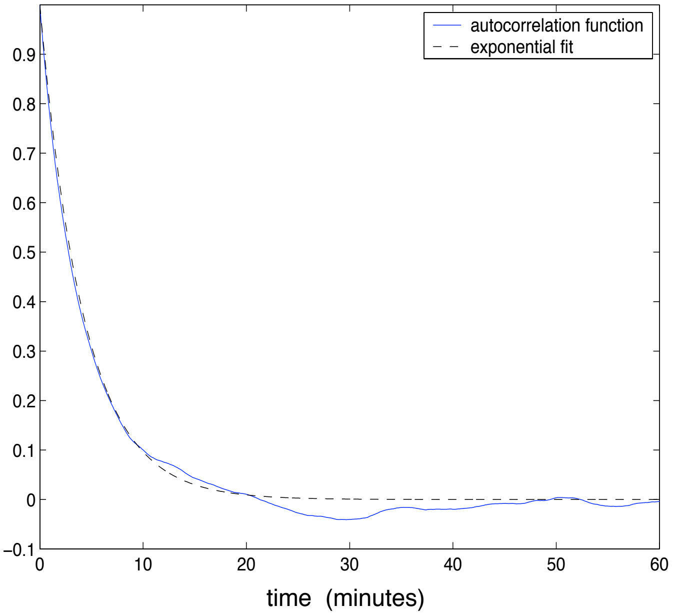

Figure 4: Auto-correlation function for a birth-death process. The dotted line is an exponential fit using an autocorrelation time of $1/d \simeq 4.2$ minutes.

For linear systems, the time-scale associated with degradation determines the steady-state autocorrelation time. Degradation provides the restoring force that keeps the number of proteins fluctuating around their mean steady-state value. The probability of degradation in time δt, $d \times n \times \delta t$, changes as the number of proteins n changes. It increases as the number of proteins rises above the mean value, increasing the probability of degradation and of return to mean levels; it decreases as the number of proteins falls below mean levels, decreasing the probability of degradation and increasing again the probability of returning to mean values. For a linear system with multiple time-scales, the autocorrelation function is a sum of terms, each exponentially decreasing with $t_1 - t_2$ at a time-scale set by the inverse of a degradation-like rate.

White noise

In Langevin theory, a stochastic variable, ξ, is added to each deterministic equation. This variable describes thermal fluctuations: those fluctuations that arise from collisions of the molecule of interest with surrounding molecules. Such collisions act to either increase or decrease the probability of reaction. *A priori*, there is no reason why thermal fluctuations would favour one effect over the other and so $\xi(t)$ is defined to have a mean of zero:

$$\langle \xi(t) \rangle = 0. \tag{53}$$

The time-scale associated with collisions is assumed to be much shorter than the time-scale of a typical reaction. The changes in internal energy and position of the molecule of interest because of collisions with solvent molecules are therefore uncorrelated at the reaction time-scale. Mathematically, the autocorrelation time, τ, of the autocorrelation function

$$C_\xi(t_2 - t_1) = \langle \xi(t_1)\xi(t_2) \rangle \tag{54}$$

is taken to zero. If Γ/τ is the variance of ξ at time t, the auto-correlation function is

$$C_\xi(t_2 - t_1) = \frac{\Gamma}{\tau} \mathrm{e}^{-(t_2-t_1)/\tau} \tag{55}$$

which becomes

$$\langle \xi(t_1)\xi(t_2) \rangle = \Gamma \delta(t_2 - t_1) \tag{56}$$

in the limit of $\tau \to 0$ where $\delta(t)$ is the Dirac delta function. A stochastic variable that obeys (53) and (56) is referred to as ‘white’. It is completely uncorrelated in time and has zero mean. Stochastic variables with zero mean and a finite auto-correlation time are considered ‘coloured’. The parameter Γ determines the magnitude of fluctuations and needs to be carefully specified (see [12] for a discussion of how Einstein famously chose Γ to appropriately model Brownian motion).

Langevin theory for stochastic gene expression

We now return to modelling the gene expression of Fig. 2. Eq. (50) is shown again below

$$\begin{aligned} \frac{dM}{dt} &= v_0 - d_0 M + \xi_1(t) \\ \frac{dN}{dt} &= v_1 M - d_1 N + \xi_2(t) \end{aligned} \tag{57}$$

and is the deterministic equations of Fig. 2 with additive, white stochastic variables.

Although we expect ξ_1 and ξ_2 to have zero mean and zero autocorrelation times, we can show that this assumptions are true explicitly by first considering the steady-state solution of (57) in the absence of the stochastic variables ξ_i:

$$M_s = \frac{v_0}{d_0} \quad ; \quad N_s = \frac{v_1}{d_1} M_s \tag{58}$$

If we assume that the system is at or very close to steady-state, and consider a time interval δt small enough such that at most only one reaction can occur, then ξ_1 and ξ_2 can only have the values

$$\xi_i \delta t = \begin{cases} +1 \\ 0 \\ -1 \end{cases} \tag{59}$$

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

Define

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

i.e. the probability that the number of mRNAs changes by an amount i and that the number of proteins changes by an amount j. Then the reaction scheme of Fig. 2 implies

$$\begin{aligned}
P(+1,0) &= v_0 \delta t \\
P(+1,-1) &= 0 \\
P(+1,+1) &= 0 \\
\\
P(-1,0) &= d_0 M_s \delta t \\
P(-1,+1) &= 0 \\
P(-1,-1) &= 0 \\
\\
P(0,+1) &= v_1 M_s \delta t \\
P(0,0) &= 1 - v_0 \delta t - v_1 M_s \delta t - d_0 M_s \delta t - d_1 N_s \delta t \\
P(0,-1) &= d_1 N_s \delta t
\end{aligned} \tag{60}$$

at steady-state.

We can use these probabilities to calculate the moments of the ξ_i. First,

$$\begin{aligned}
\langle \xi_1 \delta t \rangle &= (+1) \times v_0 \delta t + (-1) \times d_0 M_s \delta t + (0) \times (1 - v_0 \delta t - d_0 M_s \delta t) \\
&= (v_0 - d_0 M_s) \delta t \\
&= 0
\end{aligned} \tag{61}$$

and

$$\begin{aligned}
\langle \xi_2 \delta t \rangle &= (+1) \times v_1 M_s \delta t + (-1) \times d_1 N_s \delta t \\
&= (v_1 M_s - d_1 N_s) \delta t \\
&= 0
\end{aligned} \tag{62}$$

using (58). 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

$$\begin{aligned}
\langle \xi_1^2 \delta t^2 \rangle &= (+1)^2 \times v_0 \delta t + (-1)^2 \times d_0 M_s \delta t \\
&= (v_0 + d_0 M_s) \delta t \\
&= 2 d_0 M_s \delta t
\end{aligned} \tag{63}$$

or

$$\langle \xi_1^2 \rangle = \frac{2 d_0 M_s}{\delta t} \tag{64}$$

and, similarly,

$$\begin{aligned}
\langle \xi_2^2 \rangle &= \frac{2 d_1 N_s}{\delta t} \\
\langle \xi_1 \xi_2 \rangle &= 0
\end{aligned} \tag{65}$$

If the system is close to steady-state and the steady-state values of M_s and N_s are large enough such that

$$|M - M_s| \ll M_s \quad ; \quad |N - N_s| \ll N_s \tag{66}$$

then we can assume that (60) is valid for all times. Consequently, ξ_1 at time t_2, say, is completely uncorrelated with ξ_1 at time t_1, where $|t_2 - t_1| > \delta t$ (just as the throws of a die whose outcomes are also given by fixed probabilities and are uncorrelated). Thus, we define as white stochastic terms

$$\begin{aligned}
\langle \xi_1(t_1)\xi_1(t_2)\rangle &= 2d_0 M_s \delta(t_2 - t_1) \\
\langle \xi_2(t_1)\xi_2(t_2)\rangle &= 2d_1 N_s \delta(t_2 - t_1) \\
\langle \xi_1(t_1)\xi_2(t_2)\rangle &= 0
\end{aligned} \tag{67}$$

with their magnitudes coming from (63) and (65).

This definition of ξ_1 and ξ_2 implies that the steady-state solution of (57) will have the true mean and variance of N and M obtained from the master equation, providing (66) is obeyed.

A further simplification

Although it is possible to directly solve the two coupled differential equations of (57), we can also take advantage of the different time-scales associated with mRNA and protein. Typically, mRNA life-time is of order minutes while protein life-time is of order hours in bacteria. Fig. 5 shows a simulated time series of protein and mRNA: protein has a longer autocorrelation time of $1/d_1$ compared to the mRNA autocorrelation time of $1/d_0$.

Many mRNA fluctuations occur during one protein fluctuation, and so the mean level of mRNA reaches steady-state relatively quickly. Therefore, we can set

$$\frac{dM}{dt} \simeq 0 \tag{68}$$

which implies that

$$\begin{aligned}
M &= \frac{v_0}{d_0} + \frac{\xi_1}{d_0} \\
&= M_s + \frac{\xi_1}{d_0}
\end{aligned} \tag{69}$$

Consequently, the equation for protein, (57), becomes

$$\frac{dN}{dt} = v_1 M_s - d_1 N + \frac{v_1}{d_0}\xi_1 + \xi_2 \tag{70}$$

and so is a function of the two stochastic variables ξ_1 and ξ_2. To simplify (70), we define a new stochastic variable

$$\Psi = \frac{v_1}{d_0}\xi_1 + \xi_2 \tag{71}$$

which has mean

$$\langle \Psi \rangle = \frac{v_1}{d_0}\langle \xi_1 \rangle + \langle \xi_2 \rangle = 0 \tag{72}$$

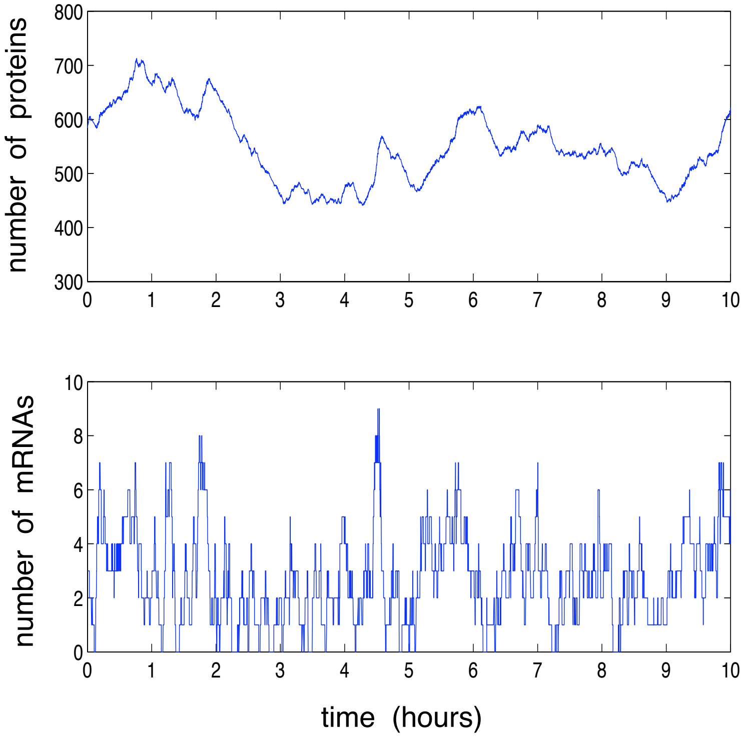

Figure 5: Protein and mRNA numbers from a simulation of the scheme of Fig. 2. Protein half-life is approximately 1 hour while that of mRNA is only 3 minutes.

from (61) and (62), and mean square

$$\begin{aligned}\langle \Psi(t_1)\Psi(t_2)\rangle &= \left(\frac{v_1}{d_0}\right)^2 \langle \xi_1(t_1)\xi_1(t_2)\rangle + 2\left(\frac{v_1}{d_0}\right)\langle \xi_1(t_1)\xi_2(t_2)\rangle \\ &\quad + \langle \xi_2(t_1)\xi_2(t_2)\rangle \end{aligned} \tag{73}$$

From Eqs. (67), this result simplifies

$$\begin{aligned}\langle \Psi(t_1)\Psi(t_2)\rangle &= \left(\frac{v_1}{d_0}\right)^2 2d_0 M_s \delta(t_2-t_1) + 2d_1 N_s \delta(t_2-t_1) \\ &= 2\left[\frac{v_1^2}{d_0}M_s + d_1 N_s\right]\delta(t_2-t_1) \\ &= 2d_1\left[\frac{v_1}{d_1}M_s\frac{v_1}{d_0} + N_s\right]\delta(t_2-t_1) \\ &= 2d_1 N_s\left[1+\frac{v_1}{d_0}\right]\delta(t_2-t_1) \end{aligned} \tag{74}$$

and so we need only consider one equation:

$$\frac{dN}{dt} = v_1 M_s - d_1 N + \Psi(t) \tag{75}$$

The effects of the mRNA fluctuations have been absorbed into the protein fluctuations and their magnitude has increased: compare (67) and (74).

Solving the model

Eq. (75) can be written as

$$\frac{d}{dt}\left(N\mathrm{e}^{d_1 t}\right) = v_1 M_s \mathrm{e}^{d_1 t} + \Psi \mathrm{e}^{d_1 t} \tag{76}$$

and so integrated

$$N(t)\mathrm{e}^{d_1 t} - N_s = \frac{v_1 M_s}{d_1}\left(\mathrm{e}^{d_1 t} - 1\right) + \int_0^t \Psi(t')\mathrm{e}^{d_1 t'} dt' \tag{77}$$

where we have assumed that $N = N_s$ when $t = 0$. Thus

$$N(t) = N_s + \mathrm{e}^{-d_1 t}\int_0^t \Psi(t')\mathrm{e}^{d_1 t'} dt' \tag{78}$$

Using the properties of $\Psi(t)$, (72) and (74), as well as (78), the mean protein number satisfies

$$\begin{aligned} \langle N(t)\rangle &= N_s + \mathrm{e}^{-d_1 t}\int_0^t \langle\Psi(t')\rangle \mathrm{e}^{d_1 t'} dt' \\ &= N_s \end{aligned} \tag{79}$$

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

We can also use (78) to find the autocorrelation function of the protein number:

$$\begin{aligned} &\langle N(t_1)N(t_2)\rangle \\ &\quad = \left\langle \left[N_s + \mathrm{e}^{-d_1 t_1}\int_0^{t_1} \Psi(t')\mathrm{e}^{d_1 t'} dt'\right] \times \left[N_s + \mathrm{e}^{-d_1 t_2}\int_0^{t_2} \Psi(t'')\mathrm{e}^{d_1 t''} dt''\right]\right\rangle \\ &\quad = N_s^2 + \mathrm{e}^{-d_1(t_1+t_2)}\int_0^{t_1} \mathrm{e}^{d_1 t'} dt' \int_0^{t_2} \mathrm{e}^{d_1 t''} dt'' \langle \Psi(t')\Psi(t'')\rangle \end{aligned} \tag{80}$$

as $\langle\Psi\rangle = 0$. From (74), we then have

$$\langle N(t_1)N(t_2)\rangle - N_s^2 = 2d_1 N_s\left(1 + \frac{v_1}{d_0}\right)\mathrm{e}^{-d_1(t_1+t_2)}\int_0^{t_1} dt' \int_0^{t_2} dt'' \mathrm{e}^{d_1(t'+t'')}\delta(t'-t'') \tag{81}$$

To evaluate the double integral, we need to determine when t' is equal to t''. If $t_2 \geq t_1$, then the integral can be decomposed into

$$\begin{aligned} \int_0^{t_2} dt' \int_0^{t_1} dt'' &= \left(\int_{t_1}^{t_2} dt' + \int_0^{t_1} dt'\right)\int_0^{t_1} dt'' \\ &= \int_{t_1}^{t_2} dt' \int_0^{t_1} dt'' + \int_0^{t_1} dt' \int_0^{t_1} dt'' \end{aligned} \tag{82}$$

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

$$\begin{aligned}
&\int_0^{t_2} dt' \int_0^{t_1} dt'' \mathrm{e}^{d_1(t'+t'')}\delta(t'-t'') \\
&\quad = \int_{t_1}^{t_2} dt' \int_0^{t_1} dt'' \mathrm{e}^{d_1(t'+t'')}\delta(t'-t'') + \int_0^{t_1} dt' \int_0^{t_1} dt'' \mathrm{e}^{d_1(t'+t'')}\delta(t'-t'') \\
&\quad = \int_0^{t_1} \mathrm{e}^{2d_1 t'} dt' \\
&\quad = \frac{1}{2d_1}\left(\mathrm{e}^{2d_1 t_1} - 1\right) \qquad (83)
\end{aligned}$$

because the first integral evaluates to zero.

Consequently, (81) becomes

$$\begin{aligned}
\langle N(t_1)N(t_2)\rangle - N_s^2 &= 2d_1 N_s \left(1+\frac{v_1}{d_0}\right)\mathrm{e}^{-d_1(t_1+t_2)}\frac{1}{2d_1}\left(\mathrm{e}^{2d_1 t_1}-1\right) \\
&= N_s\left(1+\frac{v_1}{d_0}\right)\left(\mathrm{e}^{-d_1(t_2-t_1)} - \mathrm{e}^{-d_1(t_1+t_2)}\right) \qquad (84)
\end{aligned}$$

and we finally have

$$\langle N(t_1)N(t_2)\rangle - \langle N(t_1)\rangle\langle N(t_2)\rangle = N_s\left(1+\frac{v_1}{d_0}\right)\left(\mathrm{e}^{-d_1(t_2-t_1)} - \mathrm{e}^{-d_1(t_1+t_2)}\right) \qquad (85)$$

as $\langle N(t)\rangle = N_s$. Eq. (85) is the autocorrelation function for protein number and becomes

$$C_N = N_s\left(1+\frac{v_1}{d_0}\right)\mathrm{e}^{-d_1(t_2-t_1)} \qquad (86)$$

after long times $t_2 > t_1 \gg 1$. The protein autocorrelation time is $1/d_1$.

We can also find similar expressions for mRNA. Eq. (75) has the same structure as the equation for mRNA

$$\frac{dM}{dt} = v_0 - d_0 M + \xi_1(t) \qquad (87)$$

with a constant rate of production and first-order degradation. The solution of (87) will therefore be of the same form as (86), but with d_1 replaced by d_0 and the magnitude of the stochastic term coming from (67) rather than (74). This substitution gives

$$C_M = M_s \mathrm{e}^{-d_0(t_2-t_1)} \qquad (88)$$

so that the autocorrelation time of the mRNA is $1/d_0$.

We can calculate the noise in mRNA when $t_1 = t_2$ because then the autocorrelation becomes the variance:

$$\begin{aligned}
\eta_M^2 &= \frac{\langle M(t)^2\rangle - \langle M(t)\rangle^2}{\langle M(t)\rangle^2} \\
&= \frac{M_s}{M_s^2} \\
&= \frac{1}{\langle M\rangle} \qquad (89)
\end{aligned}$$

Eqs. (88) and (89) are the solutions to any birth-and-death model and correspond to the expressions given in (28) and (29).

The protein noise is a little more complicated. It satisfies

$$\begin{aligned} \eta_N^2 &= \frac{1}{N_s} + \frac{v_1}{d_0}\frac{1}{N_s} \\ &= \frac{1}{N_s} + \frac{d_1}{d_0}\frac{1}{M_s} \\ &= \frac{1}{\langle N \rangle} + \frac{d_1}{d_0}\frac{1}{\langle M \rangle} \end{aligned} \tag{90}$$

which should be compared with (29) for a birth-death process. The mRNA acts as a fluctuating source of proteins and increases the noise above the Poisson value. Eq. (90) can be described as

$$(\text{protein noise})^2 = (\text{Poisson noise})^2 + \frac{\text{mRNA lifetime}}{\text{protein lifetime}} \times (\text{mRNA noise})^2 \tag{91}$$

The Poisson noise is augmented by a time average of the mRNA noise. As the protein life-time increases compared to the mRNA life-time, each protein averages over more mRNA fluctuations and the overall protein noise decreases. Ultimately, η_N approaches the Poisson result as $d_1/d_0 \to 0$.

More generally, we should include active and inactive states of the promoter. With this extension, the model of gene expression appears valid for bacteria [23], yeast [9], slime moulds [24], and mammalian cells [25, 26]. Physically, the two states of the promoter could reflect changes in the structure of chromatin, the binding of transcription factors, or stalling of RNA polymerases during transcription.

Typical numbers for constitutive expression

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

$$\begin{aligned} d_1 = 1/\text{hour} \quad &; \quad d_0 = 1/3 \text{ minutes} \\ \langle N \rangle = 10^3 \quad &; \quad \langle M \rangle = 5 \end{aligned} \tag{92}$$

and so (90) becomes

$$\begin{aligned} \eta_N^2 &= 1/1000 + 3/60 \times 1/5 \\ &= 0.001 + 0.01 \end{aligned} \tag{93}$$

The mRNA term determines the overall magnitude of the noise.

Appendix 1: Dirac delta function

The Dirac delta function can be considered the limit of a zero mean normal distribution as its standard deviation tends to zero:

$$\delta(x) =_{n\to\infty}^{\lim} \frac{n}{\sqrt{\pi}} \mathrm{e}^{-n^2 x^2} \tag{A1}$$

This limit gives a function whose integral over all x is one, but that becomes increasingly more and more spiked at zero (Fig. 6). Ultimately

$$\delta(x) = 0 \text{ for all } x \neq 0 \tag{A2}$$

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

$$\int_{-\infty}^{\infty} \delta(x) dx = 1. \tag{A3}$$

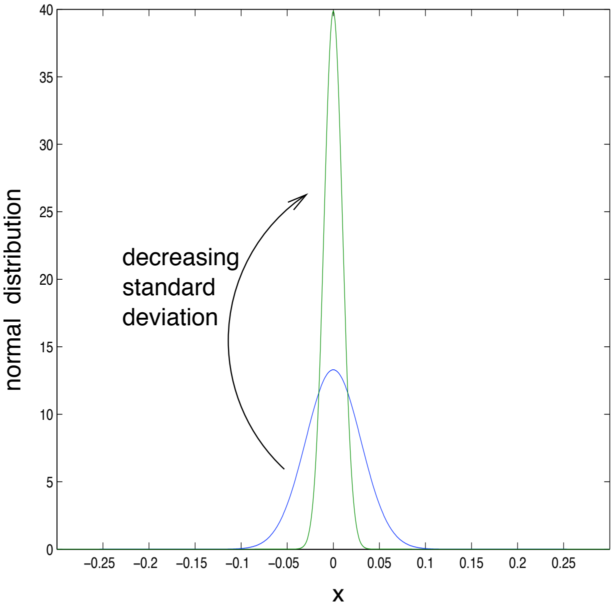

Figure 6: The Dirac delta function is the ‘spike’ limit of a normal distribution as its standard deviation tends to zero.

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

$$\int_{-\infty}^{\infty} f(x)\delta(x) dx = f(0) \tag{A4}$$

or more generally

$$\int_{-\infty}^{\infty} f(x)\delta(x-y) dx = f(y). \tag{A5}$$

Appendix 2: Sampling from a probability distribution

Often we wish to sample from a particular probability distribution, $P(x)$, say. The cumulative distribution of $P(x)$ is

$$F(x) = \int_{x_{\min}}^{x} P(x') dx' \tag{A6}$$

and

$$\begin{aligned}\mathcal{P}(x \leq x_0) &= \int_{x_{\min}}^{x_0} P(x')dx' \\ &= F(x_0) \qquad \text{(A7)}\end{aligned}$$

A sketch of the typical behaviour of $F(x)$ is shown in Fig. 7. If $x \leq x_0$, then $F(x) \leq F(x_0)$ because $F(x)$ is a monotonic increasing function (by definition).

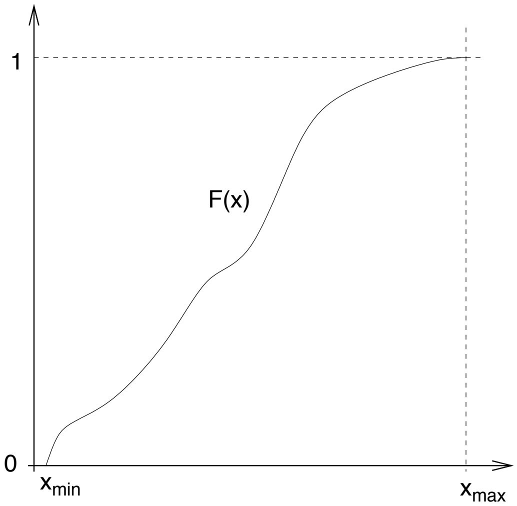

Figure 7: A typical plot of cumulative frequency versus x.

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

$$\mathcal{P}(y \leq y_0) = \int_0^{y_0} dy' = y_0 \qquad \text{(A8)}$$

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

$$x = F^{-1}(y) \qquad \text{(A9)}$$

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

$$\begin{aligned}\mathcal{P}(x \leq x_0) &= \mathcal{P}(F^{-1}(y) \leq x_0) \\ &= \mathcal{P}(F.F^{-1}(y) \leq F(x_0)) \qquad \text{(A10)}\end{aligned}$$

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

$$\begin{aligned}\mathcal{P}(x \leq x_0) &= \mathcal{P}(y \leq F(x_0)) \\ &= F(x_0) \qquad \text{(A11)}\end{aligned}$$

as y is a sample between 0 and 1 from the uniform distribution: see (A8). Thus the x of (A9) obeys (A7) and so is a sample from $P(x)$.

If we can calculate the inverse function of the cumulative frequency of a distribution $P(x)$, then applying this inverse function to a sample from the uniform distribution gives a sample from $P(x)$.

References

[1] Kaern M, Elston TC, Blake WJ, Collins JJ (2005) Stochasticity in gene expression: from theories to phenotypes. *Nat Rev Genet* 6:451–464.

[2] Shahrezaei V, Swain PS (2008) The stochastic nature of biochemical networks. *Curr Opin Biotechnol* 19:369–374.

[3] Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* 135:216–226.

[4] Eldar A, Elowitz MB (2010) Functional roles for noise in genetic circuits. *Nature* 467:167–173.

[5] Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186.

[6] Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* 99:12795–12800.

[7] Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nat Genet* 31:69–73.

[8] Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* 422:633–637.

[9] Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304:1811–1814.

[10] Samoilov MS, Arkin AP (2006) Deviant effects in molecular reaction pathways. *Nat Biotechnol* 24:1235–1240.

[11] Grima R (2010) An effective rate equation approach to reaction kinetics in small volumes: theory and application to biochemical reactions in nonequilibrium steady-state conditions. *J Chem Phys* 133:035101.

[12] Van Kampen NG (1981) *Stochastic processes in physics and chemistry* (North-Holland, Amsterdam, The Netherlands).

[13] Gillespie DT (2000) The chemical Langevin equation. *J Chem Phys* 113:297.

[14] Hasty J, Pradines J, Dolnik M, Collins JJ (2000) Noise-based switches and amplifiers for gene expression. *Proc Natl Acad Sci USA* 97:2075–2080.

[15] Swain PS (2004) Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *J Mol Biol* 344:965–976.

[16] Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340–2361.

[17] Paulsson J (2004) Summing up the noise in gene networks. *Nature* 427:415–418.

[18] Shahrezaei V, Ollivier JF, Swain PS (2008) Colored extrinsic fluctuations and stochastic gene expression. *Mol Syst Biol* 4:196.

[19] Friedman N, Cai L, Xie XS (2006) Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* 97:168302.

[20] Shahrezaei V, Swain PS (2008) Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences* 105:17256–17261.

[21] Gardiner CW (1990) *Handbook of stochastic methods* (Springer, Berlin, Germany).

[22] Thattai M, van Oudenaarden A (2001) Intrinsic noise in gene regulatory networks. *Proc Natl Acad Sci USA* 98:8614–8619.

[23] Golding I, Paulsson J, Zawilski SM, Cox EC (2005) Real-time kinetics of gene activity in individual bacteria. *Cell* 123:1025–1036.

[24] Chubb JR, Trcek T, Shenoy SM, Singer RH (2006) Transcriptional pulsing of a developmental gene. *Curr Biol* 16:1018–1025.

[25] Raj A, Peskin CS, Tranchina D, Vargas DY, Tyagi S (2006) Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* 4:e309.

[26] Sigal A, *et al.* (2006) Variability and memory of protein levels in human cells. *Nature* 444:643–646.