Lecture on

Modelling in Biology

Robert G. Endres[*]

Department of Life Sciences & Centre for Integrative Systems Biology and Bioinformatics,
Imperial College, London SW7 2AZ, United Kingdom

Systems biology aims to describe and predict the dynamics of molecules inside cells using experimental as well as theoretical and computational methods. In this introductory lecture, I describe a number of theoretical and computational methods, which can further be subdivided into deterministic and stochastic approaches. Specifically, I define what is meant by the dynamics of a molecular species and "noise". I also provide simple questions along with the text, so you can test your understanding.

Contents

I. INTRODUCTION

Biological cells can sense and respond to external and internal cues. Often, biochemical pathways transduce signals from cell-surface receptors to the cell interior, ultimately regulating cell shape, cell motility, metabolism, and/or gene expression. Unlike hard-wired microprocessors in computers, biological processes inside a living cell are based on random chemical reactions facilitated by the occasional encounter of the reacting molecular species by diffusion. Furthermore, the number of the molecules in a cell is relatively small, enhancing the effects of randomness in a cell. While molecules and proteins can exist in hundreds to tenthousands of copies in a cell, genes are present in only a very small number. As a result of these sources of stochasticity in a cell, the numbers of molecules in a cell can fluctuate significantly. Hence, noise in biology is an important aspect, which can deteriorate the precision of a specific response, but can also lead to beneficial cell-to-cell variability for increased survival fitness of a population.

In systems biology, we aim to quantitatively describe and predict the concentrations of molecules in a cell or a biochemical pathway in time. To mathematically describe the time evolution of copy numbers of molecular species, two different approaches are generally possible. First, there are deterministic approaches, such as differential equations, which predict the average concentration of a molecule in time. Second, there are stochastic approaches, which include the fluctuations of a molecular species in time, often requiring simulations instead of precise analytical or numerical solutions.

---

[*] E-mail: r.endres@imperial.ac.uk

II. DETERMINISTIC APPROACHES

A. Ordinary differential equations

When the number of a molecule in a cell is large, we can approximate the discrete copy number of a molecular species by a continuous variable, denoted e.g. by n. Its rate of change is most simply described by an ordinary differential equation of the form

$$\frac{dn}{dt} = \underbrace{k}_{\text{source}} - \underbrace{\gamma \cdot n}_{\text{decay}}, \tag{1}$$

where the positive first term on the right-hand side (RHS) describes the source of the molecule (here a constant rate of production), and the negative second term on the RHS describes the degradation or decay of the molecules. Here, $\gamma \cdot n$ means that each molecule has a constant rate of degradation (inverse of life time), so the total rate of degradation is the larger the more molecules there are. In other words, this term is equivalent to radioactive decay, well-known from physics. Note that while Eq. 1 is just the number of particles, we could also divide by the reaction volume V and write Eq. 1 in terms of concentration $[n]$, i.e. $d[n]/dt = k/V - \gamma \cdot [n] = \tilde{k} - \gamma[n]$, where $\tilde{k} = k/V$.

Questions:
(1) What are the units of the rate constants in Eq. 1?
(2) Why is it an "ordinary differential equation (ODE)"?
(3) What is the order of the reactions?
(4) Is it a linear or a nonlinear ODE?
(5) What does the time evolution look like for initial condition $n(0) = n(t = 0) = 0$?
(6) What is a partial differential equation (PDE)? Name a famous one.

B. Analytical solution

To solve the example of Eq. 1 analytically, we use the method of *separation of variables*. We start with writing Eq. 1 differently

$$\frac{dn}{dt} = \frac{\bar{n} - n}{\tau} \tag{2}$$

with the old parameters given by $k = \bar{n}/\tau$ and $\gamma = \tau^{-1}$. Now, we separate the variables n and t via

$$\frac{dn}{\bar{n} - n} = \frac{dt}{\tau} \tag{3}$$

$$\implies \int^{n} \frac{d\tilde{n}}{\bar{n} - \tilde{n}} = \int^{t} \frac{d\tilde{t}}{\tau} \tag{4}$$

$$\implies -\ln(\bar{n} - n) = \frac{t}{\tau} + C, \tag{5}$$

where the constant of integration C is determined by the boundary condition. For $n(0) = 0$, we obtain $C = -\ln(\bar{n})$. Step-by-step, we solve for n and obtain the final solution

$$\ln(\bar{n} - n) = \ln(\bar{n}) - \frac{t}{\tau} \tag{6}$$

$$\implies \bar{n} - n = e^{\ln(\bar{n})} \cdot e^{-t/\tau} = \bar{n}e^{-t/\tau} \tag{7}$$

$$\implies n(t) = \bar{n}(1 - e^{-t/\tau}). \tag{8}$$

For small times t, $e^{-t/\tau} \approx 1 - t/\tau$ (by Taylor expansion) and we obtain an approximate linear relationship in time

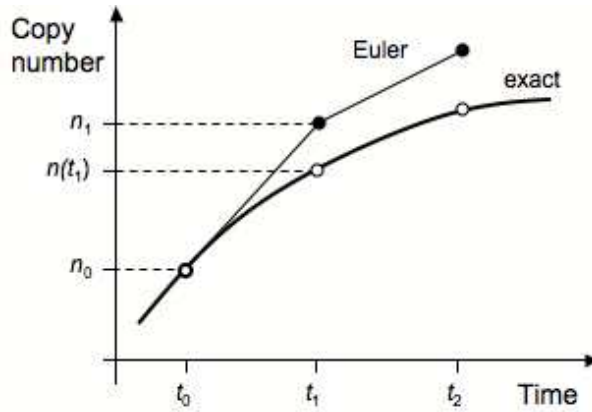$$n \approx \bar{n}[1 - (1 - t/\tau)] = \frac{\bar{n}}{\tau} \cdot t = kt. \tag{9}$$

FIG. 1: Illustration of the numerical Euler method.

For large t, we regain the steady-state level $\bar{n} = k/\gamma$.

Questions:
(1) Rederive Eq. 8 starting from Eq. 2.
(2) Plot $n(t)$, as well as the asymptotic bahaviour at small and very large times.
(3) Solve $dn/dt = k$ by direct integration.
(4) Solve $dn/dt = -\gamma n$ by separation of variables.

C. Numerical solution

Here, we use Euler's method to numerically solve Eq. 1, now written as

$$\frac{dn}{dt} = f(n), \tag{10}$$

where the RHS of Eq. 1 is given in terms of the n-dependent function $f(n)$. Using condition $n = n_0$ at $t = t_0$, we approximate solution in a systematic way using a vector-field interpretation: at n_0, quantity $f(n_0)$ is like the velocity of a ficticious particle at position n_0 in one spatial dimension. In a small time interval Δt, such a particle moves by $f(n_0)\Delta t$ (distance=rate$\times$ time). As a consequence, the new position of the particle is

$$n(t_0 + \Delta t) = n_1 \approx n_0 + f(n_0)\Delta t. \tag{11}$$

Now we iterate, using n_1 as the old position and $f(n_1)$ as the velocity. This leads to

$$n_2 = n_1 + f(n_1)\Delta t, \tag{12}$$

and more generally to

$$n_{i+1} = n_i + f(n_i)\Delta t, \tag{13}$$

where subscript i counts the iterations.

As illustrated in Fig. 1, Euler's method has severe disadvantages. To avoid large errors, time step Δt needs to be very small. However, if the time step is very small, one requires many iterations, and consequently, the solution progresses only very slowly and suffers from round-off errors. Additionally, derivative dn/dt is only calculated at the left-hand side of time interval Δt.

The improved Euler's method estimates the derivative also at the right side of the time interval, and uses the average derivate across the interval. To get the derivate at the right side of the time interval, we use estimate $f(\tilde{n}_{i+1})$,

calculated from $\tilde{n}_{i+1} = n_i + f(n_i)\Delta t$. The new position is now calculated from the average derivate

$$n_{i+1} = n_i + \frac{1}{2}[f(n_i) + f(\tilde{n}_{i+1})]\Delta t. \tag{14}$$

The error $E|n(t_i) - n_i| \propto (\Delta t)^2$, whereas the original Euler's method has the larger error $E|n(t_i) - n_i| \propto \Delta t$. The even further improved Forth-order Runge-Kutta method has an error as smll as $(\Delta t)^5$, but requires four function evaluations and hence is significantly slower.

Question:
Check that the error for the original Euler method is of order Δt, as well as determine the exact prefactor. Hint: first, use Taylor expansion to expand exact result. Then calculate integrated error, assuming we integrate over time $T = m\Delta t$ with m calculation steps.

D.   Kinetic laws

Law of mass action. The law states that the rate of an elementary reaction (a reaction that proceeds through only one transition state, i.e. one mechanistic step) is simply proportional to the product of the concentrations of the participating molecules. For instance, the reaction between an enzyme of concentration E and a substrate of concentration S, which produces a product of concentration P, is assumed to be reversible and describable by

$$E \; + \; S \; \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \; E \; + \; P. \tag{15}$$

By definition, the enzyme is not modified in this reaction. In this reaction, the forward and backwards rates are given by the products of the concentrations of the reacting species, multiplied by the rate constant, via

$$r_+ \; = \; k_1 E \cdot S \tag{16}$$
$$r_- \; = \; k_{-1} E \cdot P. \tag{17}$$

Using these rates, the rate of change of the product can be described by the ordinary differential equation

$$\frac{dP}{dt} \; = \; k_1 E \cdot S \; - \; k_{-1} E \cdot P. \tag{18}$$

Michaelis-Menten kinetics. This type of kinetics was shown to describe a large number of different enzymes reasonably well, and is relevant for situations where very simple kinetics can be assumed (, i.e. there is no intermediate or product inhibition, and there is no allostericity or cooperativity). More complex models exist for the cases where the assumptions of Michaelis-Menten kinetics are no longer appropriate.

  The validity of the following derivation rests on the reaction scheme given below and two key assumptions: (1) the total enzyme concentration and the concentration of the intermediate complex do not change over time. (2) The enzymatic reaction is assumed to be irreversible (small P)

$$E \; + \; S \; \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \; ES \; \overset{k_2}{\longrightarrow} \; E \; + \; P. \tag{19}$$

In this model, what does the important rate of change of the product, $dP/dt = k_2\, ES$, look like? While the product is the measurable quantity, we would like to remove the intermediate ES from the resulting formula, as it is normally not measured. Furthermore, we make the quasi-steady state assumption that the concentration of the intermediate complex ES changes much more slowly than the concentrations of the product and substrate (valid for large S and large k_2). Hence,

$$\frac{dES}{dt} = 0 \; = \; k_1 E \cdot S - (k_{-1} + k_2)ES \tag{20}$$

$$\implies ES \; = \; \frac{k_1 E \cdot S}{k_{-1} + k_2} = \frac{E \cdot S}{K_M}, \tag{21}$$

where $K_M = (k_{-1} + k_2)/k_1$ is the so-called Michaelis-Menten constant. For very small k_2, $K_M \approx K_D = k_{-1}/k_1$ is
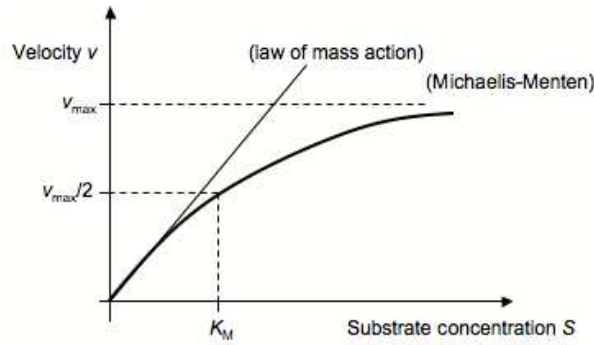
FIG. 2: Reaction velocity with saturation using Michaelis-Menten kinetics.

equilibrium dissociation constant.

Using the total concentration $E_0 = E + ES$ of the enzyme, where E is the free and ES is the bound enzyme, we obtain for the bound enzyme concentration

$$ES = \frac{(E_0 - ES)S}{K_M} \tag{22}$$

$$\implies ES = E_0 \frac{S}{S + K_M}. \tag{23}$$

Using this expression for ES, we are finally able to calculate the rate of change of product P, often also called the velocity v of the reaction

$$v = \frac{dP}{dt} = k_2 \cdot ES = k_2 E_0 \frac{S}{S + K_M} = v_{\max} \frac{S}{S + K_M}, \tag{24}$$

with $v_{\max} = k_2 E_0$ the maximal velocity of the reaction. For little available substrate, i.e. small S, we obtain mass-action type kinetics

$$v \approx \frac{v_{\max}}{K_M} \cdot S = \frac{k_2 E_0}{K_M} S. \tag{25}$$

Figure 2 shows the reaction velocity as a function of substrate concentration S, including the asumptotic behaviour for small and very large concentrations of the substrate. For large S, the enzyme saturates, and the velocity reaches a maximum. Note, however, that S changes normally, so often people just calculate the initial velocity for a given S.

Question:
Rederive Eq. 24.

III. STOCHASTIC APPROACHES

A. Master equation

As explained above, if the number n of a molecule is large, we can ignore the fact that n is an integer with its dynamics given by

$$\frac{dn}{dt} = k - \gamma \cdot n = f_n - g_n, \tag{26}$$

where we called the source term f_n and the decay term g_n. However, if the number is small, we cannot ignore the discreteness of n, and a more general approach based on the master equation is needed. The master equation describes the reaction kinetics probabilistically, such as for an ensemble of cells. What terms do we use to describe the rate of change of the probability $dp_n/dt = ...$ for n molecules? If we make the imaginary time interval small enough, the number n can change at most by one molecule. As an example, the number of reactions by rate r in time T is rT.
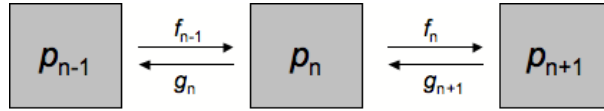
FIG. 3: Processes which affect probability p_n of observing n molecules.

We can now divide time T into N small bins (time intervals), leading to rT/N reactions per bin. If we make N large enough, no more than one reaction will occur per bin in line with the stoichiometry given. In this case, we can use simple one-step processes, which only change the number n by one

$$\frac{dp_n}{dt} = -(f_n + g_n)p_n + f_{n-1}p_{n-1} + g_{n+1}p_{n+1}, \qquad (27)$$

where p_n is the probability to have n molecules. The first, negative term describes the decay of the state with n molecules, either by adding one molecule with rate f_n or by removing one with rate g_n. The second and third positive terms describe the production of one molecule, either by creating a molecule from state $n-1$ (term 2) or by destroying one from state $n+1$ (term 3). Equation 27 is illustrated in Fig. 3. While Eq. 27 is linear and simple, it constitutes an infinite set of coupled differential equations, making it impossible to solve exactly in most cases.

Emergence of deterministic law. The deterministic approach Eq. 26 describes the average number of particles, while the master equation describes the probability distribution, characterised by the average (first moment), the variance (second moment), and higher moments. These can generally be obtained without solving the master equation explicitly. In the following, we derive the first moment and rediscover the determinstic law.

Using notation $\langle n \rangle = \sum_{n=0}^{\infty} np_n$ for the average number of molecules, we obtain from the master equation

$$\frac{d\langle n \rangle}{dt} = -k\sum_{n=0} np_n - \gamma \sum_{n=0} n^2 p_n + k\sum_{n=0} np_{n-1} + \gamma \sum_{n=0} n(n+1)p_{n+1}. \qquad (28)$$

The third and forth terms on the RHS of Eq. 28 can be expressed in terms of p_n by substituting

$$\sum_{n=0} np_{n-1} = \sum_{n=0} (n-1)p_{n-1} + \sum_{n=0} p_{n-1}$$

$$= \sum_{n=0} np_n + \sum_{n=0} p_n = 1 + \sum_{n=0} np_n \qquad (29)$$

$$\sum_{n=0} n(n+1)p_{n+1} = \sum_{n=0} (n+1)^2 p_{n+1} - \sum_{n=0} (n+1)p_{n+1} = \sum_{n=0} n^2 p_n - \sum_{n=0} np_n, \qquad (30)$$

where the last step only works if f_n and g_n are linear functions of n. As a result, we finally obtain

$$\frac{d\langle n \rangle}{dt} = k - \gamma \langle n \rangle. \qquad (31)$$

Note that in Eq. 31, we used $\langle n \rangle$ for the average number of molecules, while in Eq. 1 we simply used n. At steady-state, the time derivative is zero, and we obtain

$$0 = k - \langle n \rangle^* \gamma \implies \langle n \rangle^* = \bar{n} = k/\gamma. \qquad (32)$$

Question:
Rederive Eq. 31.

B. Poisson distribution

The Poisson distribution is a discrete probability distribution that describes the probability that a number of events occurs in a fixed period of time, if events occur with a known average rate independently of previous events. Since these assumptions are often fulfilled in biological systems, the distribution is very important in biology. The Poisson

distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

As an example, in the following we derive the steady-state probability distribution of Eq. 27 and show that it is Poissonian. At steady-state, all time derivatives are zero, and we have

$$\frac{dp_n}{dt} = 0 \ \Rightarrow \ 0 = -(f_n + g_n)p_n + f_{n-1}p_{n-1} + g_{n+1}p_{n+1}, \tag{33}$$

Dividing by γ and using steady-state value $\bar{n}$ from Eq. 32, we obtain

$$\implies \ \ -(\bar{n} + n)p_n + \bar{n}p_{n-1} + (n+1)p_{n+1} = 0 \tag{34}$$

$$\implies \ \ (n+1)p_{n+1} - \bar{n}p_n = np_n - \bar{n}p_{n-1} = 0. \tag{35}$$

The last step comes from Eq. 35 being valid for all n, including $n = 0$, since $p_{n-1} = 0$. Rearranging terms, we obtain the recursive formula

$$p_n = \frac{\bar{n}}{n}p_{n-1} = \frac{\bar{n}}{n} \cdot \frac{\bar{n}}{(n-1)}p_{n-2} = \cdots = \frac{\bar{n}^n}{n!}p_0. \tag{36}$$

As p_0 is not specified, we obtain it from normalization. Setting the sum over all p_n equal to one, we obtain

$$\sum_{n=0} p_n = p_0 \sum_{n=0} \frac{\bar{n}^n}{n!} = p_0 e^{\bar{n}} = 1 \ \to \ p_0 = e^{-\bar{n}}, \tag{37}$$

and for the final probability distribution

$$p_n = \frac{\bar{n}^n}{n!}e^{-\bar{n}}, \tag{38}$$

which only depends on average $\bar{n}$. The Poissonian result is not unexpected as the rates of production and degradation are simple, i.e. respectively constant and linear.

Note the following facts:

(1) For a Poisson distribution, the variance $\langle (\delta n)^2 \rangle = \langle (n - \langle n \rangle)^2 \rangle = \langle n^2 \rangle - \langle n \rangle^2$ is equal to the mean $\langle n \rangle$, i.e. $\langle (\delta n)^2 \rangle = \langle n \rangle$.

(2) While the Poisson distribution is only characterised by the average, the Gaussian distribution depends on the average and variance. The binomial distribution, describing the outcome of N yes/no trials (coin flips), depends not only on the probability p for an event to occur, but also $q = 1-p$ that no event occurs. Its variance is Npq.

(3) The Poisson distribution can be derived as a limiting case to the binomial distribution as the number of trials N goes to infinity and the expected number of successes Np remains fixed. Therefore it can be used as an approximation of the binomial distribution if N is sufficiently large and p is sufficiently small, so $q \approx 1$.

(4) For sufficiently large average values, the Gaussian distribution with same average and variance is an excellent approximation to the Poisson distribution.

Limit of large numbers. We said earlier that when noise is small, ordinary differential equations can be used to describe the kinetics of the average number of molecules. Using the Poisson distribution, we can be more specific about what we mean by small noise or fluctuations in molecule numbers. Since for a Poisson distribution $\langle (\delta n)^2 \rangle = \langle n \rangle$, and hence the absolute size of the fluctuations in the number increases with the average number, the relative fluctuations, given by the standard deviation over the mean,

$$\frac{\sqrt{\langle (\delta n)^2 \rangle}}{\langle n \rangle} = \frac{1}{\sqrt{\langle n \rangle}} \tag{39}$$

decrease with increasing average number. Consequently, for a very large numbers of molecules, the relative fluctuations can be neglected, and the kinetics of the average is an excellent approximation for the time evolution of the system. As an example, if the average is 20 molecules, then the relative fluctuations, Eq. 39, are 22%. If instead the average
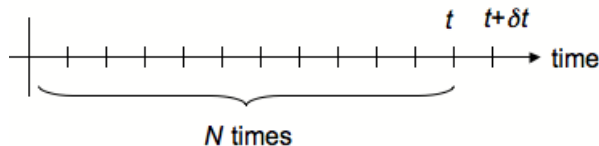
FIG. 4: Illustration of discretisation of time for deriving waiting-time distribution.

is 500, then the relative fluctuations are only 4%.

C. Gillespie simulations

Waiting-time distribution. What about the time intervals between successive occurances of a reaction and how are they distributed? Suppose a reaction occurs with rate (probability per time) r. Then the probability that the reaction occurs in time interval δt is $P(\text{event in } \delta t) = r\delta t$ and that it does not occur is $P(\text{no event in } \delta t) = 1 - r\delta t$. This can be used to calculate the distribution that the event occurs in time interval $[t, t + \delta t]$, but not earlier. For this purpose, we imagine the time to be divided into N small bins of size δt as shown in Fig. 4. In this case, we can apply the probabilites already given. Hence, we obtain

$$P(t) = P(\text{event in } [t, t + \delta t]) = \underbrace{(1 - r\delta t)\ldots(1 - r\delta t)}_{N \text{ times}} \cdot r\delta t = \left(1 - r\frac{t}{N}\right)^N \cdot r\delta t \approx re^{-rt}\delta t. \tag{40}$$

As a result we obtain that the waiting times between successive events are exponentially distributed. In fact, knowing $P(t)$ provides a simple numerical procedure to simulate stochastic systems, called Gillespie simulation. Specifically, $P(t)$ tells us when reactions occur given the total rate of all possible reactions, and once a reaction occurs, how to update the simulation time.

Question:
Calculate and check expressions for average $\langle t \rangle = \int_0^\infty t P(t) dt = 1/r$ and variance $\langle (\delta t)^2 \rangle = 1/r^2$ using integration by parts.

Gillespie simulations. A Gillespie simulation provides the same probability distribution as the corresponding master equation, and hence is exact (not an approximation). However, lots of simulations might be required to provide good statistics. In addition, there are other kinetic Monte Carlo simulations possible, often implementing faster, approximate schemes.

Suppose there are different chemical species with certain copy numbers, as well as $\mu = 1, 2, \ldots$ possible chemical reactions with rates r_μ at a particular point in time. The total rate that something happens is $\Gamma = \sum_{\mu=1} r_\mu$. Running a Gillespie simulation on this system requires repeated drawing of two random numbers, using the following interative scheme:

Step 1: Determine next occuring reaction by drawing random number R_1 between 0 and 1, and picking reaction μ accordingly. Choose reaction $\mu = 1$, if $0 \leq R_1 < r_1/\Gamma$, choose reaction $\mu = 2$, if $r_1/\Gamma \leq R_1 < (r_1 + r_2)/\Gamma$, and so on.

Step 2: Update molecule numbers according reaction stoichiometry.

Step 3: Increment simulation time by $\tau = -\ln(R_2)/\Gamma$, where R_2 is a second random number between 0 and 1. This time increment τ is obtained by inverting waiting-time distribution $R_2 = e^{-\Gamma\tau}$ (see also Fig. 5 for a graphical derivation). Then go to step 1 and iterate.

Example: Reaction $\mu = 1$ is dimer formation from species P_1 and P_2 via

$$P_1 + P_2 \xrightarrow{k} Z, \tag{41}$$

where Z is the dimer species. Assuming molecular numbers N_1 and N_2 for species P_1 and P_2, respectively, the reaction
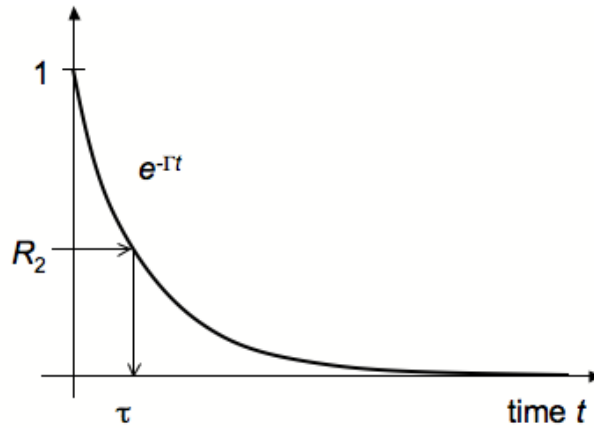
FIG. 5: Updating time in Gillespie simulations with R_2 a random number and Γ the total rate that something happens.

rate is given by

$$r = \begin{cases} c\,N_1 \cdot N_2, & \text{if } P_1 \text{ and } P_2 \text{ are different species (heterodimer)} \\[2mm] c\,\frac{N_1 \cdot (N_1 - 1)}{2}, & \text{if } P_1 \text{ and } P_2 \text{ are same species (homodimer),} \end{cases} \tag{42}$$

with c a prefactor. The difference between the first and second version of rate r in Eq. 42 is that N_1 and N_2 are distinguishable in the former, and indistinguishable in the latter, reducing the number of possible reactions. Furthermore, c is related to the macroscopic rate constant k. Writing in terms of concentrations (using small letters)

$$\frac{dz}{dt} = k \cdot p_1 \cdot p_2, \tag{43}$$

it is apparent that k has units of volume per time. Hence,

$$c = \begin{cases} \frac{k}{V}, & \text{(heterodimer)} \\[2mm] \frac{2k}{V}, & \text{(homodimer),} \end{cases} \tag{44}$$

with the two versions of dimerisation from Eq. 42.

---

[1] Endres RG, Physical principles in sensing and signaling (Oxford, 2013).
[2] Strogatz ST, Nonlinear dynamics and chaos (Westview, Cambridge, 1994).
[3] Raj A, van Oudenaarden A (2008) Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell* **135**: 216-226.
[4] Rao CV, Wolf DM, Arkin AP (2002) Control, exploitation and tolerance of intracellular noise. *Nature* **420**: 231-237.
[5] Van Kampen NG, Stochastic processes in physics and chemistry (Elsevier, 2007).
[6] Berg HC, Random walk in biology (Princeton University Press, 1993).