Biological Physics

Pietro Cicuta and Diana Fusco

Experimental and Theoretical Physics
Part III
Michaelmas 2025

Notes version: v0.07
Release name: Giraffe Genes

Contents

Course aims and structure

Course aims

Possible questions

Is Biological Physics well defined?
A pragmatic answer is along the lines of what the late Sir Sam Edwards (former Cavendish professor, and pioneer of polymer physics) was fond of saying: "Physics is what physicists do". Sir Sam was not the first to hold this view, and in the Cavendish there has always been a strong tradition of applying physics to new areas, regardless of traditional disciplines. Often this process requires developing new physics. The rationale behind this is that Science is one enterprise, and so are the challenges that are posed by the "real world". You need to find areas where good progress can be made and where it is worth putting your effort. In this sense, if a physicist sees an opportunity to contribute in a unique way to biology, this can/should be pursued. Is is pretty obvious that "biology" is itself very broad (consider how many aspects, spatial and time scales, there are to living systems, reflected for example in a mosaic of departments and institutes that is not unique to Cambridge). Important questions, relevant to health and technology, and full of fundamental scientific value, can be posed at *many length-scales* from the molecular through cellular, organ/tissue, up to organisms, populations and ecology. In terms of *time-scales, biology also spans many orders of magnitude*, one can pose questions all across molecular dynamics, to organism development and maintenance of tissues, to evolution or ecological equilibria.

So, clearly, there will be many ways to apply physics (and develop new physics), and many different types of models that can be deployed or invented based on physical intuition. Unexpected and typically "physics" insight often comes under the form of understanding how an emergent property or behaviour arises non-trivially out of simpler underlying rules that did not encode anything resembling this property. If you think about it, this is not so different from how we treat condensed matter systems: despite the fact that it is easier to dig down with reductionist approaches when dealing with, say, materials, as opposed to a living thing, even in those simpler cases we never expect to have a unified model to provide quantitative predictions, at the same time, for all the material's behaviour. X-ray diffraction, melting temperat-

ures, properties of density, conductivity or elasticity... all rely on quite different descriptions and models. We are used, and ready to accept, in this context, the idea that we can abstract the important elements that underlie a certain phenomenon, and we find justification in a separation of time- or length-scales, or in finding the appropriate degrees of freedom. This leads us to come up with quite different 'physics' (stat mech, or continuum mechanics, or electrodynamics, etc.). When this is done well, we are capturing the "correct" mechanism, which entails many things, but mainly that: (a) we are indeed working with the most relevant mechanism, and hence are able to show how modifying some 'physically motivated' parameters changes properties or outcomes, often in non-trivial ways; (b) we are able to link to a set of data and often to make unexpected (and ideally verifiable) quantitative predictions. These are properties that somehow identify and distinguish the way a physicist defines what is 'understanding', as opposed to other quantitative approaches more typical of engineers, statisticians or mathematicians. These properties also lead us to make 'physical models' - be careful that the concept of 'model', which is so crucial for physics, has completely different meanings in each of biology, maths, engineering and of course outside science.

Is there new physics in biology?
This is a corollary to the statements above. But yes, particularly at the present time! Data now exists in very quantitative (and reproducible) form for a large number of biological systems and processes. Developments in the last decade, not just in genetics but also in imaging and other forms of measurement of the concentrations, dynamics and localisations of the key biological agents, have revolutionised many areas of biology with truly quantitative data of unprecedented resolution (time, space) and extensiveness (repeats, conditions). These lend themselves like never before to applying and developing physical models, in exactly the same spirit as in studying condensed matter systems, or other complex systems (nonlinear optics, strongly interacting cold atoms, etc.). There are also many biological systems where the data is not yet in a form that a physicist would find acceptable: This poses another family of challenges that physicists might want to take on, typically on the experimental (and in some cases computational or data analysis) frontier, developing new bespoke experiments and techniques.

What can be achieved in a course of 24 hours? The main aims of this course are:
(a) through good examples, and with a storyline as coherent as possible, mostly from the molecular to the cell level, show how physics (particularly statistical mechanics, soft matter, networks and nonlinear dynamics approaches) has been developed and ap-

plied in recent years to address both existing challenges, and even to define new categories in biological systems.
(b) provide (through (a)) an exposure and an education such that interested students will be able to make informed decisions on fields of further study.

What is this course not? What is not in this course?
This course is not a traditional 'biophysics' course, that term is usually meant to emphasise the molecular aspect (e.g. protein folding, molecular structures, biochemical interactions); we touch only some aspects. Another community (medical) defines biophysics as biomechanics, focusing on issues to do with circulation, pressure, etc. - this course has none of this 'macro-physiology' side. It is not an instrumentation course, and we only describe a few interesting techniques (instruments and protocols quickly improve and become obsolete - an exception is medical instrumentation, which due to the degree of certification involved changes slowly, but we do not cover that here).

There are also a lot more topics that would be closer to the spirit of this course, but that we cannot cover due to lack of time and personal expertise. The same type of thinking and modeling presented here could take you further in understanding embryonic development (and tissue homeostasis), aspects of evolution (it is possible to study evolution in the lab, on fast growing species, exploring influence of stresses, species competition, etc), ecology (also amenable to lab experiments, with suitable choices of model organisms). Interested students will find many colleagues in Cambridge and beyond working on these questions, and we hope this course will provide good 'transferable skills'.

Structure of the course

Given the preamble above, our challenge is to provide a coherent 'story', covering various concepts and examples that we think are useful. Whilst not wanting to overburden anyone with fact collections, a minimum of context is necessary and will be useful in any future interactions you will have with the world of Life Sciences.

The course is structured into seven modules (A-G). After the introduction, modules (B-G) each take between 2 and 4 lectures, and have a single lecturer (Prof. P.Cicuta or Dr D.Fusco). We expect this year four 'guest lecturers' who will give a lecture each (details non examinable) on their pioneering discoveries of quantitative aspects in cell biology, and how they pursue physical modeling. They will appear at appropriate moments in the course, hopefully "on topic", and they will be able to explain their research/experimental approaches in more detail than what is pos-

sible in the rest of the course.

We are fortunate that a handful of good textbooks have been published in the last few years. You will see that many illustrations and question sheet problems come from (Phillips et al., 2013), which is a very 'reader friendly' source. The book does not cover everything (and we don't use the whole book), and in some places we wanted to go deeper, so other sources are also used, and referenced in appropriate places.

Module A: Introduction to the course, a primer of useful concepts and to science of networks. 3 lectures

Physical biology of the cell - information processing 'central dogma'; Model building in biology; Quantitative models and the power of idealisation; Special role of *E.coli* in quantitative biology; Transcription and translation numbers; Cells and structures within them; Networks - graph representation; Random graphs; Motifs, feedback, modularity; Construction plans for cells.

Module B: Dynamics in cells. 3 lectures

Bioenergetics - free energy transductions in the cell; Single molecule techniques; Models of molecular motors: ratchets and asymmetric hopping; Cytoskeletal dynamics; Rotary Motors.

Module C: Bioelectricity: Sensing and Neural Biophysics. 2 lectures

The electrical status of cells and their membranes; The Hodgkin-Huxley Model for the generation of action potentials; Information processing in neurons.

Module D: Spatial Patterns. 3 lectures

Reaction diffusion equation; Turing instabilities; Notch-Delta concept.

Module E: Protein production and regulation of gene expression. 5 lectures

ODE for protein production; Biochemical (small number) noise; Gillespie algorithm; The mechanics of transcriptional regulation: the example of the Lac operon; Statistics of regulation: transcriptional and post-transcriptional; Strategies for regulating noise in gene expression; Case study: phage lambda, the hydrogen atom of molecular biology.

Module F: Circuits and dynamical systems. 4 lectures

The Gillespie algorithm for simulations of reaction systems; Properties of dynamical systems, and intro to methods; Feedback circuits; Genetic circuits with switching and with oscillating properties.

Module G: Population growth, genetic and evolution 4 lectures

Examples of evolution across length and time scales; Concepts of genetic diversity, mutation, selection and genetic drift; Neutral evolutionary models; Fitness landscapes; Beyond well-mixed: the role of space in evolution.

Concepts in networks

1

adapted from: MIT OpenCourseWare, Kardar/Mirny 2011

If limited to one role per protein, the roughly 30,000 Human genes would have limited utility. The key to diversity of behavior is: (i) the combinatorial power from many genes acting in concert; (ii) the time profile of expressing and suppressing genes, (iii) localization/compartmentalization of proteins in different locations, and (iv) interactions with the resources and stimuli from the environment. Various forms of behavior can then emerge from a palette of few elements.

The primary elements of a network are its nodes. These can be a set of genes or proteins or metabolic products (sugars, lipids) in the cell, or the interconnected neurons of the brain, or organisms in an ecosystems. Links between nodes indicate a direct interaction, for example between proteins that bind, neurons connected by synapses, or organisms in a predator/prey relationship. In its most basic form, the network can be represented by nodes $i = 1, 2, \cdots, N$ as points of a graph, and links L_{ij} as edges between pairs of points. Excluding self-connections, the maximal number of possible links is $N(N-1)$ with directed connections (e.g. as in a predator/prey relation), and $N(N-1)/2$ for undirected links (as in binding proteins). A subgraph is a portion of the total network, say with n nodes and l links. Some types of subgraphs have specific names; e.g. a *cycle* is a path starting and ending at the same node, while a *tree* is a branching structure without cycles.

The transcription network of *E. coli* (Figure 1.1), or yeast (Figure 1.2, from (Sneppen and Zocchi, 2005)), are very complex. But buried in this information are interesting statistical properties. Particularly, one can look for patterns that appear more (or less) often than in a random graph of equivalent number of nodes and links. Then once can think of why from a biological function or evolutionary perspective an organism might be "wired-up" in these non-trivial ways. Patterns that appear more often than expected in a random network are called *network motifs*.

Note (following U.Alon) that a transcription network is quite delicate to maintain against random genetic mutations: a mutation changing a single letter in the DNA of a promoter can change dramatically the affinity of a transcription factor, and result in the loss of an edge in the network. To get an idea of the rate of these

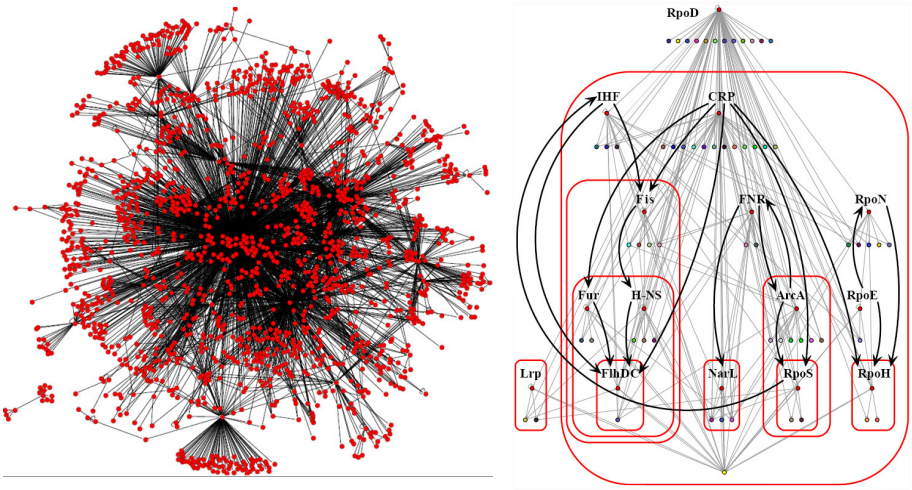


Fig. 1.1 Representations of data from RegulonDB, the database of E. coli regulation data (Salgado et al. 2006). On the left, the (known parts of) the *E. coli* transcriptional regulatory network. In this graphical representation, nodes are genes, and edges represent regulatory interactions. There is extreme complexity present in regulatory networks, but also biologically relevant organizational principles hidden in the architecture governing these networks. On the right, functional architecture of *E. coli* genetics as revealed by the natural decomposition approach. Red-labeled nodes represent global transcription factors. Genes composing modules were shrunk into a single colored node. Black arrows indicate regulatory interactions between global transcription factors. Red rounded-corner rectangles bound hierarchical layers. For the sake of clarity, RpoD (the housekeeping sigma factor) interactions are not shown, and the single yellow node at the bottom represents the megamodule whose submodules are held together only by intermodular genes. This analysis revealed that the functional architecture hierarchy exhibits feedback from well-defined independent modules devoted to particular cellular functions. The functions are globally coordinated by global transcription factors, and the disparate responses are integrated by intermodular genes. Images from: Freyre-Gonzalez, J. A. & Trevino-Quintanilla, L. G. (2010) Analyzing Regulatory Networks in Bacteria. Nature Education 3(9):24.

mutations: a single bacterium in 10 ml of culture will grow in 1 day to reach 10^{10} cells. So 10^{10} DNA replications. The mutation rate is about 10^{-9} per letter per replication. So the population at the end of the day will include for each letter in the genome 10 bacteria with a mutation in that letter. So a change of a DNA letter is achieved very rapidly in a bacteria population. Similar mutations can of course also add an edge to the network if they increase some affinity to bind a transcription factor. As a conclusion, edges in a network are under constant selection pressure in order to survive randomisation. So if some network motifs are found in transcription networks, there must be a selective advantage associated to them.

Analyzing biological data from the perspective of networks has gained interest recently. Much is known about the interplay of proteins that control expression of genes, the connections of the few hundred neurons in the roundworm *C. elegans*, and other ex-

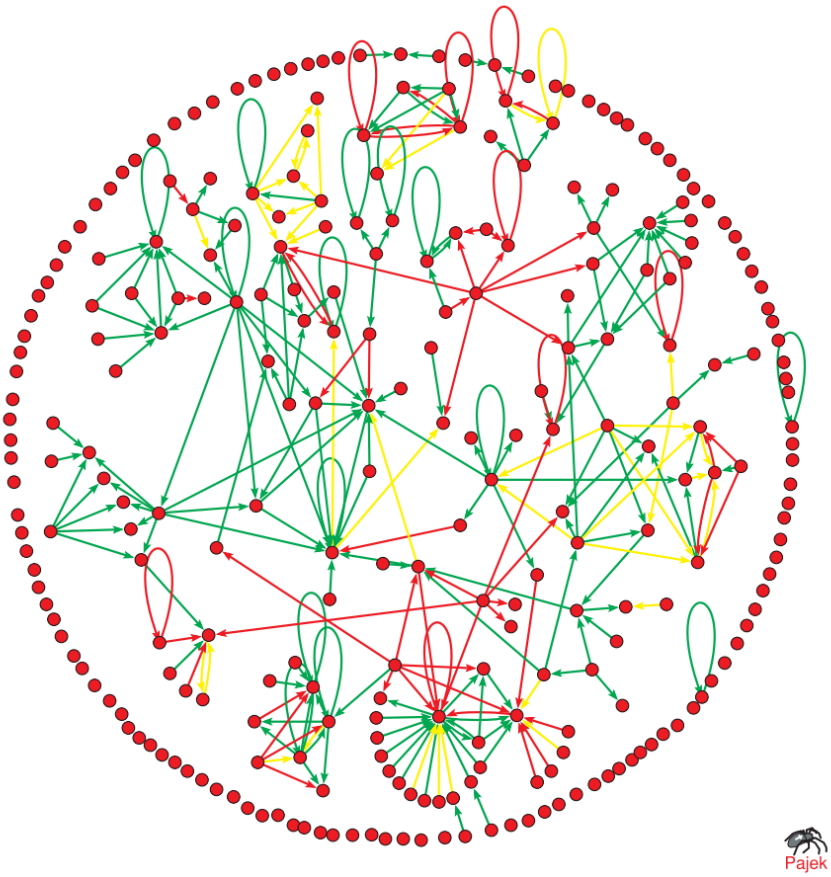


Fig. 1.2 Networks of transcriptional regulatory proteins in yeast. All proteins that are known to regulate at least one other protein are shown. Arrows indicate the direction of control, which may be either positive or negative. Functionally the network is roughly divided into an upper half that regulate metabolism, and a lower half that regulate cell growth and division. In addition there are a few cell stress response systems at the intersection between these two halves.

amples. One possible route to extracting information from such data is to look for specific motifs, subgroups of several nodes, that can cooperate in simple functions (e.g. a feedforward loop). A particular motif can be significant if it appears more (or less) frequently than expected. We thus need a simple model whose expectations can be compared with biological data. Random graphs, introduced by Erdös and Rényi, serve this purpose: The model consists of N nodes, with any pair connected at random and independently, with probability p.

We shall explore a few features of Erdös-Rényi (ER) networks in the following sections. For the time being, we note that you can obtain the expected number of subgraphs of n nodes and l links as a product of the number of ways of picking n points and connecting them with l links, and a factor that accounts for the

number of ways of connecting the points into the desired graph:

$$\mathfrak{N}(n,l) \;=\; \binom{N}{n} p^l \times \frac{n!}{(symmetry\,factors)}.$$

For example, there are $n!/2$ ways to string n points along a straight line with $l = (n-1)$, and the expected number of such linear pathways is:

$$\mathfrak{N}(n\,\text{in a line}) \;=\; \frac{N!}{(N-n)!}\frac{p^{n-1}}{2},$$

while there are $n!/(2n)$ ways to make a cycle of n nodes and $l = n$ links, such that

$$\mathfrak{N}(n\,\text{in a cycle}) \;=\; \frac{N!}{(N-n)!}\frac{p^n}{2n}.$$

There is also a single way to make a complete graph in which any pair of nodes is connected by a link, i.e. $l = n(n-1)/2$, and

$$\mathfrak{N}(n\,\text{in complete graph}) \;=\; \frac{N!}{(N-n)!n!} p^{n(n-1)/2}.$$

The autoregulation network motif

In *E. coli* transcription network there is an excess of self-edges, the vast majority of which are repressors that implement negative autoregulation. How can this conclusion be reached? We need a way to compare with the expected number of self-edges in a random network.

With N nodes, there are $N(N-1)/2$ possible pairs of nodes that can be connected by an edge. Each edge can point in one of two directions, for a total of $N(N-1)$ possible places to put a directed edge. An edge can also begin and end at the same node, so there are a total of N possible self-edges. Total number of edges is thus

$$N(N-1) + N \;=\; N^2.$$

In the ER model, the E edges are placed at random in the N^2 possible positions, so each possible edge position is occupied with probability $p = E/N^2$.

Let's calculate the probability $P(k)$ of having k self edges in an ER network: a self edge needs to choose its node of origin as a destination, out of the possible N destinations. So

$$p_{self} \;=\; 1/N.$$

Since the E edges are placed at random, the probability of having k self edges is approx binomial:

$$P(k) \;=\; \binom{E}{k} p_{self}^k (1-p_{self})^{E-k}.$$

The average number of self-edges is E times the probability of being a self edge, i.e.

$$\langle N_{self} \rangle_{rand} = E p_{self} = E/N,$$

with a standard deviation that is approximately (because binomial approx Poisson) the square root of the mean, so

$$\sigma_{rand} \simeq \sqrt{E/N}.$$

Alon considers data where N=424, and E=519, and in which there are 40 self edges (34 are repressors). The random graph expectation is

$$\langle N_{self} \rangle_{rand} = E/N = 1.2 \text{ with } \sigma_{rand} \simeq \sqrt{1.2} = 1.1.$$

Obviously there is a difference of many standard deviations. We will return later to negative autoregulation, to see some properties of this simple network motif, and hence why it is highly selected.

Percolation cluster in large networks

A network can display two types of global connectivity. With few connections amongst nodes, there will be many disjoint clusters, with their typical size (but not necessarily number) increasing with the number of connections. At high connectivity there will be one very large cluster, and potentially a number of smaller clusters. In the limit of $N \to \infty$, a well defined percolation transition separates the two regimes in the random graph, as the probability p is varied (remember from above: p is the probability that a given pair of nodes is linked). Above the percolation transition, the number of nodes M in the largest cluster also goes to infinity, proportionately to the number of nodes, such that there is a finite percolation probability $P(p) = \lim_{N\to\infty} \frac{M}{N}$ (this is the probability for a node to belong to the infinite cluster).

For the random graph, $P(p)$ can be calculated from a self-consistency argument: Take a particular site and consider the probability that it is not connected to the infinite cluster. This is the case if none exist of the $(N-1)$ edges emanating from this site potentially connecting it to the large cluster. A particular edge connects to the infinite cluster with probability $pP(p)$ (that the edge exists, and that the adjoining site is on the large cluster), and hence

$$\begin{aligned} 1 - P(p) &= (\text{prob of no connections to any edge})^{N-1} \\ &= (1 - pP)^{N-1}. \end{aligned}$$

It is possible to show that there is a phase transition, which is a percolation transition, in this probability. If the limit $N \to \infty$ is taken, but also at the same time $p \to 0$ such that we keep

$p(N-1) = \langle k \rangle$, where $\langle k \rangle$ is the (finite) average number of edges per node, then the equation above can be expressed as

$$\begin{aligned} 1 - P(p) &= e^{-\langle k \rangle P} \\ i.e.\ P(p) &= 1 - e^{-\langle k \rangle P} \end{aligned}$$

which can be solved e.g. graphically. We see that if $\langle k \rangle \leq 1$, there is only $P = 0$ as a solution, whereas if $\langle k \rangle > 1$ then there can be a $P \neq 0$ solution, indicating the appearance of an infinite cluster. Close to the percolation transition at $\langle k \rangle_c$, P is small and we can expand the last expression, to get

$$P \approx \frac{2(\langle k \rangle - 1)}{\langle k \rangle^2} \approx 2(\langle k \rangle - 1).$$

Distance, Diameter, & Degree Distribution

There are typically several ways to traverse from a node i to a node j. The distance between any pair of nodes is defined as the number of edges along the shortest path between the nodes. For the entire network, we can define a diameter as the largest of all distances between pairs of nodes.

Distances to a particular node can be obtained efficiently by the following simple (burn and move) algorithm. In the first step, label the nodes connected to the starting point ($d = 1$), and then remove it from the network. Consider a random graph with $\langle k \rangle \gg 1$, such that $P \approx 1$. (Distances cannot be defined to disconnected clusters.) In the random graph, the number of sites with $d = 1$ will be around $p(N-1) = \langle k \rangle$. In the second step identify all sites connected to the set labeled before (and thus at $d = 2$), and then remove all sites with $d = 1$ from the network. From each site with $d = 1$, there are of the order of $p(N - \langle k \rangle - 1) \approx \langle k \rangle$ accessible sites, since $\langle k \rangle \ll N$. There are thus around $\langle k \rangle^2$ sites labeled with $d = 2$. This burn and move process can be repeated, with $N_p \lessapprox \langle k \rangle^p$ sites tagged at distance $d = p$. (Note that each step we have overestimated the number of sites by ignoring connections leading to sites already removed.) The procedure has to be stopped when all sites belonging to the cluster have been removed, i.e. for

$$\langle k \rangle^D \lessapprox N, \ \Rightarrow D \lessapprox \frac{\ln N}{\ln \langle k \rangle},$$

where D is a rough measure of the diameter of the network. Note that the diameter of a random network is quite small, justifying the popular lore of "six degrees of separation". In a population of a few billion, with each individual knowing a few thousand, the last equation in fact predicts a distance of three or four between any two. Clearly segregation by geographical and social barriers increases this distance. The model of "small world networks" considers mostly segregated communities, but shows that even a small

fraction of random links is sufficient to reintroduce a logarithmic behavior like in the expression above.

For $\langle k \rangle < 1$, the typical situation is of disjoint clusters. We can then inquire about the probability p_k that there are exactly k links emanating from a site. Since there are a total of $(N-1)$ potential connections from a site, in a random graph the probability that k such links are active is given by the binomial probability

$$P_k = \binom{N-1}{k} p^k (1-p)^{N-1-k}.$$

Taking the limits $N \to \infty$ and $p \to 0$ with $pN = \langle k \rangle$ as before, we obtain

$$p_k = \frac{N^k}{k!} \frac{p^k}{(1-p)^k} (1-p)^{N-1} = \frac{\langle k \rangle^k}{k!} e^{-\langle k \rangle},$$

i.e. a Poisson distribution with mean $\langle k \rangle$.

Looking at information across organisms, as exemplified in the data gathered in (Sneppen and Zocchi, 2005) for Figure 1.3, can also be very informative: here, it is shown that there is strong regularity (a linear dependence) between the *fraction* of proteins that regulate other proteins, and the size of the genome. One notices that those with a very small genome hardly use transcriptional regulation. More strikingly, it appears that the number of regulators, N_{reg}, grows much faster than the number of genes, N, it regulates. If life was just a bunch of independent switches, this would not be the case. That is, if living cells could be understood as composed of a number of modules (genes regulated together) each, for example, associated with a response to a corresponding external situation, then the fraction of regulators would be independent of the number of genes N. Networks are not just modular, they show strong features of an integrated circuitry, even on the largest scale. A question sheet exercise explores further the implications of these data on the connectivity of these regulatory networks.

Beyond the completely random network

A common feature of molecular networks is the wide distribution of directed links from individual proteins. There are many proteins that control only a few other proteins, but also there exist some proteins that control the expression level of many other proteins. It is not only proteins in the regulatory networks, but also metabolic networks and protein signaling networks. The distribution of proteins with a given number of neighbors (connectivity) K can often (if very crudely) be approximated by a power law

$$N(K) \sim 1/K^{\gamma}$$

with exponent $\gamma \simeq 2.5 \pm 0.5$ for protein-protein binding networks, and exponent $\gamma \simeq 1.5 \pm 0.5$ for "out-degree" distribution of tran-

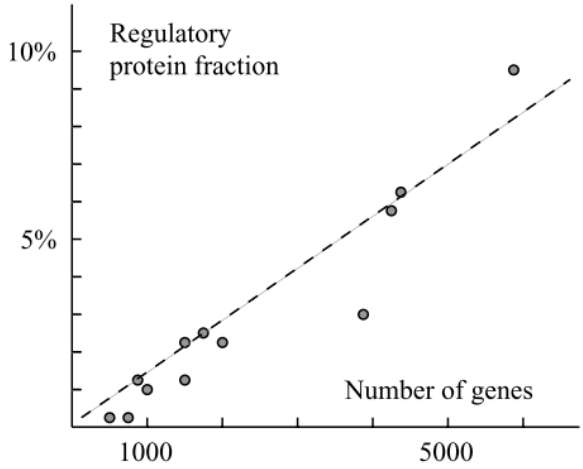


Fig. 1.3 Fraction of proteins that regulate other proteins, as a function of size of the organisms' gene pool. These data are for prokaryotes: smallest genome is *M. Genitalium* (480 genes); the largest genome is *P. Aeruginosa* (5570 genes). The linear relation demonstrates that each added gene should be regulated with respect to all previously added genes. Eukaryotes scale differently.

scription regulators. (Note that the broad distribution of the number of proteins regulated by a given protein, the "out-degree", differs from the much narrower distribution of "in-degrees".)

Models and results for random graphs built with various 'rules' are useful because they can be used as potential models for assessing significance of putative anomalies in the degree distributions biological and social networks.

Membrane potential and neurons: structures for information transfer and processing

2

2.1 Membrane potential

The physical chemistry of electric fields in acqueous ionic solutions has been introduced in the Part II Soft Condensed Matter course. In the presence of an electric field (which could be an external field, or a field due to surface-bound charges) ions in a solution drift to an equilibrium concentration distribution which minimises the free energy, i.e. a distribution that balances the electrostatic field energy together with the entropy. The details of this are in general extremely complex, and simple solutions exist only for simple geometries and weak fields where it's possible to linearise the equations.

Biological membranes composed of lipid bilayers are impermeable to both ions and macromolecules. Transport across membranes only occurs through specific protein complexes, which can be passive stable pores (channels, sometimes specific for particular ionic species) or active pumps (consuming chemical energy to transport ions in a directed fashion). The hydrophobic core of the lipids is a very non-polar chemical species with very low dielectric constant. So the bilayer as well as preventing ionic flow is also a very good isolator. Strong electric fields can be generated by confining ions on opposite sides of a membrane, and these conditions are at the heart of many cellular processes. One example is the generation of ATP, the chemical 'fuel' for most cellular activity, from glucose in mitochondria. This ATP synthesis is powered by the proton gradient across the inner membrane of the mitochondria. Another example is the propagation of electrical signals along the axons of neural cells (see Figures 2.1 and 2.2).

The combination of lipid bilayer, channels and pumps in the membrane can be thought of as resistors (voltage dependent), batteries (dependent on ionic concentrations) and capacitors in parallel. By modulating the number or activity of the channel molecules, the cell can regulate the gradient of different ionic species, and thus tune the membrane electrical potential.

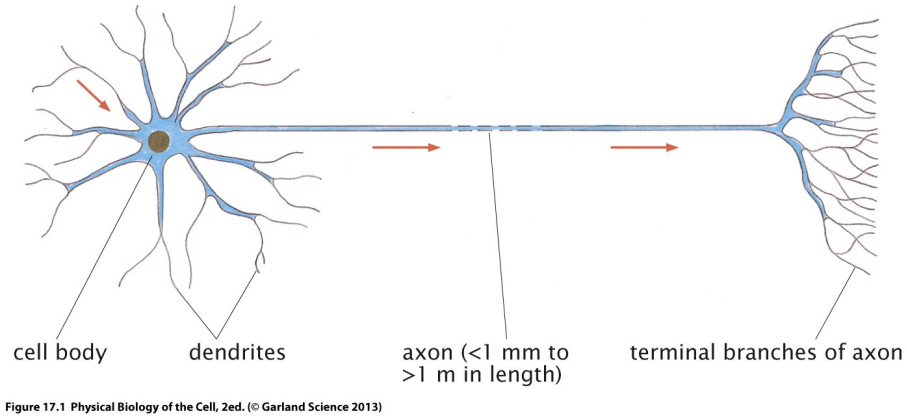


Fig. 2.1 Illustration of a nerve cell. The action potential carries a signal along the axon.

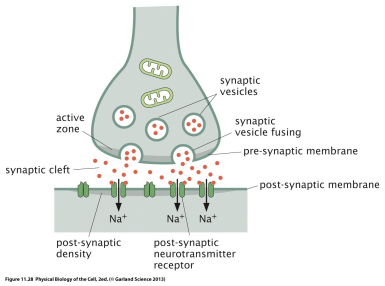


Fig. 2.2 Illustration of a synapse. When the action potential reaches the end of the axon, the electrical signal is converted to a chemical signal thanks to the presence of vesicles loaded with neurotransmitters, the "synaptic vesicles". These vesicles release the cargo into the cleft, and the neurotransmitters are sensed at the membrane of another cell. These neurotransmitters have the effect of lowering the polarisation state of the receiving cell membrane, and thus a chemical signal is turned back into an electrical signal.

The Nernst equation relates the voltages across a septum that allows only positive ions to cross:

$$V_2 - V_1 = \frac{k_B T}{ze} \ln \frac{c_1}{c_2}, \tag{2.1}$$

where c_1, c_2 are the concentrations at either side, ze is the charge on the ion. $k_B T/e \simeq 25\,\mathrm{mV}$. Voltages across membranes of neurons are in the range $10 - 100\,\mathrm{mV}$, see Table 2.1, so thermal effects can be important.

Looking at the ions, there is an excess of positive charge inside the cell. This is partly balanced by the negative charges of the nucleic acids and of many proteins. In a typical neural cell, the potential difference is -90 mV. Other cells will typically have smaller gradients. In thermodynamic equilibrium, one would expect eq.2.1 to hold separately for each ionic species, with the same Nernst potential, but clearly this is not consistent with the data in Table 2.1. The cell is not in equilibrium. There are voltage dependent, ion-selective channels, and energy consuming ion pumps. The most important pumps in the cell membrane are of two types: to pump calcium ions out of the cell, and to exchange sodium with potassium ions. Several hundred *distinct* ion channels are coded for in the human genome.

By evolving the distribution and properties of channels, axons have emerged as structures where a voltage can propagate as a signal. These signals are typically pulses of around 50 mV in amplitude, and of width between 1 and 500 ms depending on the organism and cell type. This pulse travels down the axon without changing shape, at speeds between 10 and 100 m/s, see Figure 2.3. A threshold value of voltage is required to trigger the action potential.

Hodgkin and Huxley worked between the 30s and the 50s measuring nerve signals, and managed to produce a model that was

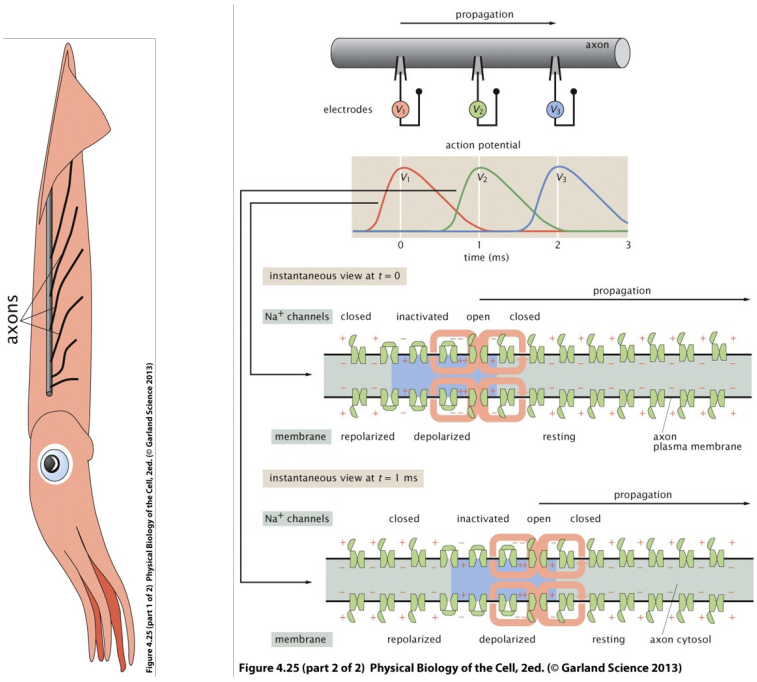


Fig. 2.3 The giant squid axon was the structure studied in original experiments on neural signals. These are for a mammalian skeletal muscle cell, for which the resting potential is around $V_{mem} = -90\,\mathrm{mV}$.

phenomenological (at the time a lot of the molecular structures were unknown) but yet had the correct mechanisms. They were awarded the 1962 Nobel prize in Physiology.

It is the transient (due to opening and closing of the channels) permeability to K^+ and Na^+ ions that lies at the heart of the action potential. The opening and closing are regulated by voltage (in other context, they can also be regulated by mechanical forces, membrane tension, ligand binding, phosphorylation). Membrane voltage keeps the channels closed, and their 'preferred' configuration (in absence of voltage) is open, see Figure 2.4.

Ion species	Intracellular concentration (mM)	Extracellular concentration (mM)	Nernst potential (mV)
K^+	155	4	−98
Na^+	12	145	67
Ca^{2+}	10^{-4}	1.5	130
Cl^-	4	120	−90

Table 17.1 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Table 2.1 Typical concentrations and voltages. These are for a mammalian skeletal muscle cell, for which the resting potential is around $V_{mem} = -90\,\mathrm{mV}$ (i.e., the inside of the cell is 90 mV lower than the exterior).

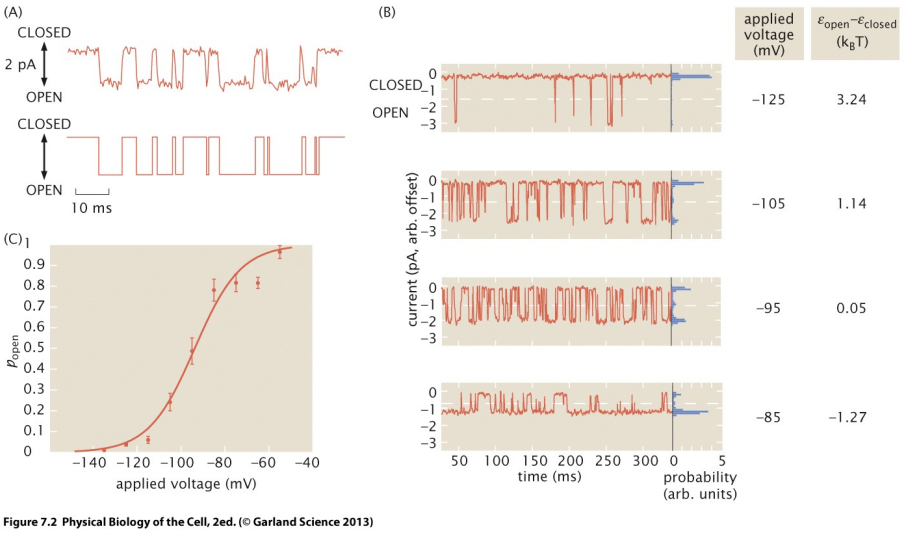


Fig. 2.4 Patch clamp experiments can measure the current through membranes. These data are for sodium channels. A two state approximation for the channel is appropriate. In (B), single channel data.

In the approximation of a two state model with energy differences $\Delta\epsilon$, the probability of the channel being open is:

$$p_{open} = \frac{e^{-\beta\Delta\epsilon}}{1 + e^{-\beta\Delta\epsilon}}. \tag{2.2}$$

The energy depends on the membrane potential:

$$\Delta\epsilon = \Delta\epsilon_{conf} - QfV_{mem}, \tag{2.3}$$

where $\Delta\epsilon_{conf}$ is the energy difference in the protein channel conformational states, and the second term accounts for the motion of charges Q in the potential. The quantity f is a fraction. If we have the field across the thickness d as $E = V_{mem}/d$, then $fV_{mem} = fdE$ is the drop in potential experienced by the 'gating charge' when it switches from the open to the closed position.

One has $p_{open} = 1/2$ at $\Delta\epsilon = 0$, which defines:

$$V^* = \frac{\Delta\epsilon_{conf}}{Qf}. \tag{2.4}$$

So we can rewrite the channel being open probability as:

$$p_{open} = \frac{1}{1 + e^{\beta q(V^* - V_{mem})}}, \tag{2.5}$$

with $q = Qf$. This sigmoidal shape is a key ingredient in the action potential.

2.2 Biological electricity and the Hodgkin-Huxley model

Two channel proteins are important in explaining propagating pulses:

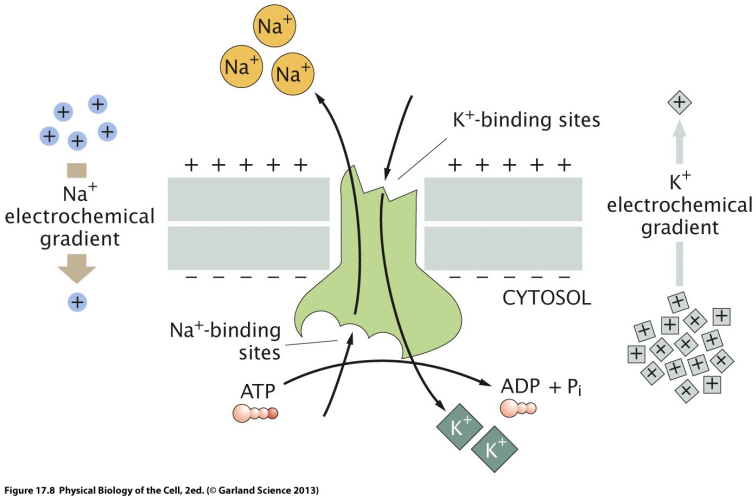


Fig. 2.5 The sodium-potassium pump.

- First, the sodium-potassium pump, see Figure 2.5. This pump consumes the energy of an ATP molecule ($20\,k_BT$ for hydrolysis of ATP) and uses that to pump 3 Na^+ ions out of the cell, and 2 K^+ ions inside.
- Second, potassium (and sodium) "leek" channel, which selectively allows potassium (or sodium) to transport.

If a potassium 'leek channel' is open, K^+ escape the cell following their concentration gradient. However, this change reduces the magnitude of the membrane potential. Potassium will reach a balance close to its Nernst potential. Now if a sodium channel opens, sodium will move both in response to its own concentration gradient, and also to re-equilibrate the charge lost by the potassium. Sodium will reach a balance close to its Nernst potential. The sodium channels are normally closed. They open transiently and locally, if a neighboring patch of membrane has been depolarised. These ideas can be made into a quantitative model.

The capacitance of a patch of the membrane is defined as

$$C_{patch} = \frac{Q_{patch}}{V_{mem}}, \tag{2.6}$$

where Q_{patch} is the excess charge on either side of the membrane patch. In terms of charge density σ, and patch area, we have $Q_{patch} = \sigma A_{patch}$. In a parallel plate capacitor the field is uniform and given by $\sigma/\epsilon_0\epsilon_r$, so with membrane thickness d we have

$$V_{mem} = \frac{\sigma d}{\epsilon_0\epsilon_r}, \tag{2.7}$$

hence with eq. 2.6

$$C_{patch} = \frac{\epsilon_0\epsilon_r A_{patch}}{d}. \tag{2.8}$$

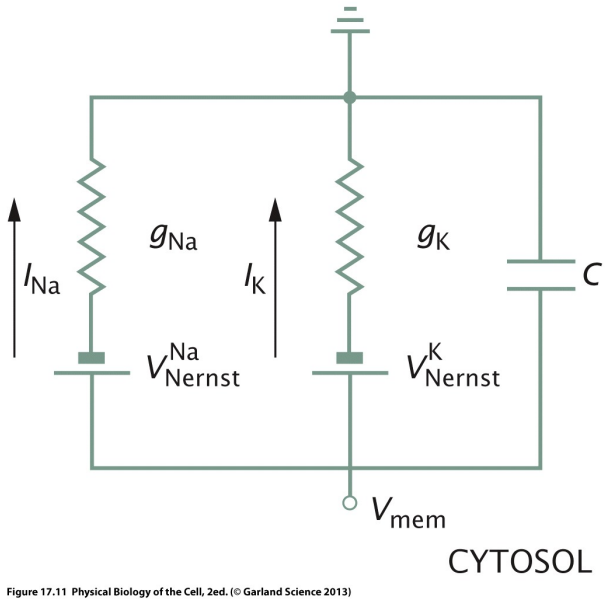


Fig. 2.6 Mapping of an excitable membrane into an electric circuit. The g are conductances.

With $\epsilon_r \simeq 2$ and $d \simeq 5\,\mathrm{nm}$, the capacitance per unit area is expected as $0.4\,\mu\mathrm{F/cm^2}$. Cell membranes have $1\,\mu\mathrm{F/cm^2}$.

It is possible to verify building on eq. 2.8 that a significant $\Delta V_{mem} \simeq 100\,\mathrm{mV}$ depolarisation can be carried out by moving a vanishingly small fraction (of order 10^{-5} of the ions in a cell) across the membrane.

The ionic current across the membrane is:

$$I = g\left(\frac{k_B T}{ze}\ln\frac{c_{in}}{c_{out}} + V_{in} - V_{out}\right), \tag{2.9}$$

where g is the conductance per unit area. Here positive current is a flow of positive ions out of the cell.

With eq. 2.1, we can write $I = g(V_{mem} - V_{Nernst})$, the form of an Ohm law.

If G_1 is the conductance of a single ionic channel, then $g_{patch} = N_{patch} p_{open} G_1$. Recall that p_{open} is not linear in the voltages, see eq. 2.5.

If p_{open} were *always* =1, i.e. the channel always open, then we'd expect Ohm conduction, and a linear $I-V$ dependence from eq. 2.9. This is what happens for the potassium ions in Figure 2.7. If p_{open} were a step change, opening at voltage V^*, and allowing g to jump to a higher value, then current I would drop suddenly as the voltage increases across V^*. The voltage would grow linearly both below and above V^*, but with a discontinuous jump.
The realistic sigmoidal form of p_{open} is quite similar to a jump, just smoothed out. Voltage drops, as in Figure 2.7 for the sodium ions.

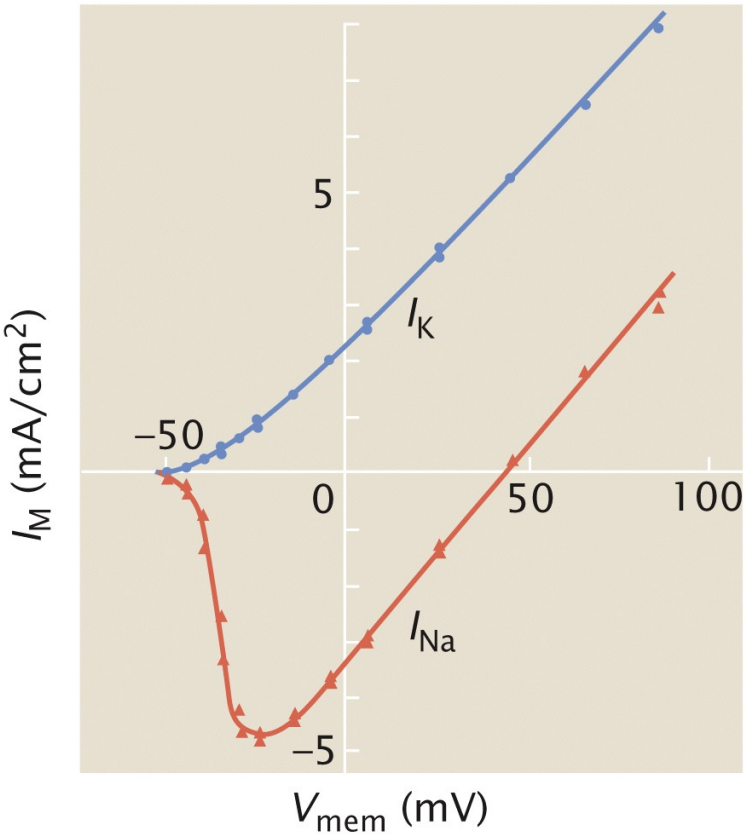


Figure 17.13 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Fig. 2.7 Data of current through a membrane patch, as a function of membrane voltage. Note the linear I_V characteristics of the potassium ion current, and the extremely non-linear response of the sodium ions.

2.2.1 Bistable behaviour of membrane voltage

The charge across the membrane can be written as $\Delta Q = -(I_K + I_{Na})\Delta t$ and also as $\Delta Q = C\Delta V_{mem}$, so putting the two together we obtain the differential equation valid in each small patch of membrane:

$$C\frac{dV_{mem}}{dt} = g_K(V^K_{Nernst} - V_{mem}) + g_{Na}(V^{Na}_{Nernst} - V_{mem}). \quad (2.10)$$

An interesting property of this is that it is bi-stable. Setting the time derivative =0,

$$V_{mem} = \frac{g_K V^K_{Nernst} + g_{Na} V^{Na}_{Nernst}}{g_K + g_{Na}}. \quad (2.11)$$

If $g_K >> g_{Na}$ then $V_{mem} \simeq V^K_{Nernst}$. If instead $g_{Na} >> g_K$ then $V_{mem} \simeq V^{Na}_{Nernst}$. The ion species with the highest conductance through the membrane sets the membrane potential to its Nernst potential. The conductances are of course related to the open/closed balance of the channels, and we have seen that this is Voltage regulated.

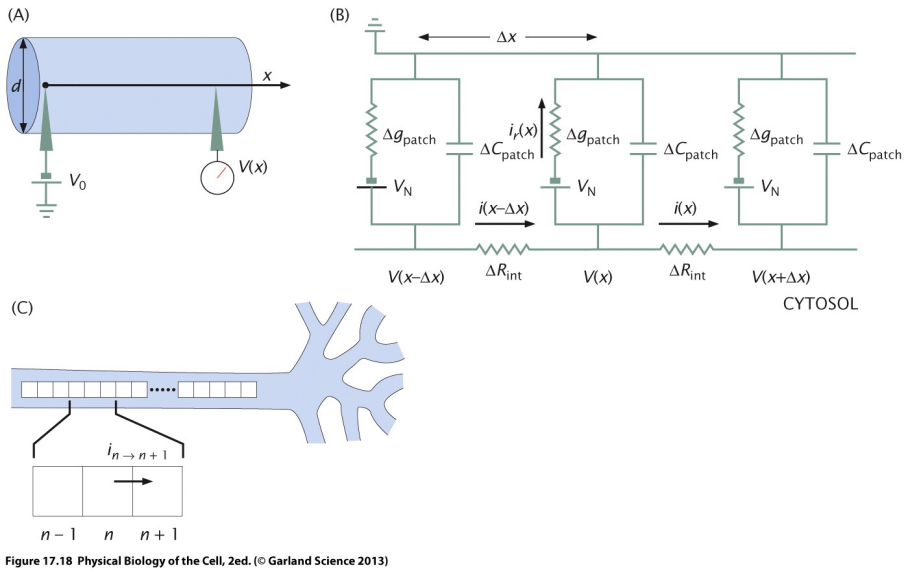


Fig. 2.8 We model the axon as a series of elements. Each element is a cylinder of diameter d and length Δx.

An approximation is to consider the potassium conductance as fixed; this is justified because its dynamics is slower than the dynamics of the sodium channel. Then, we have that the sodium conductance is proportional to its own p_{open} from eq. 2.5. For $V_{mem} < V^*$ this gives a low conductance, and therefore the membrane will take the Nernst potential of potassium. If $V_{mem} > V^*$ the sodium channels open, sodium conductance dominates. The membrane potential goes to the Nernst voltage of potassium, a positive value.

2.2.2 Cable equation

We can model the spatial transmission of signal by considering a series of circuits as in Figure 2.8. The change of voltage along the axon is connected to the longitudinal current: $V(x+\Delta x)-V(x) = -i(x)\Delta R_{int}$. The current $i(x)$ can change along the axon due to currents in the radial direction $i_r(x)$. In each module there is also conservation of current, which gives $i(x-\Delta x) - i(x) = i_r(x) = \Delta g_{patch}\,(V(x) - V_{Nernst})$. Since $\Delta R_{int} = \rho\Delta x/(\pi d^2/4)$ and $\Delta g_{patch} = g\pi d\Delta x$, where d is the axon diameter, ρ the resistivity of the intracellular medium, and g the conductance per unit area, we can combine these to give:

$$\frac{d^2V(x)}{dx^2} = \frac{1}{\lambda^2}(V(x) - V_{Nernst}), \tag{2.12}$$

with

$$\lambda = \sqrt{\frac{d}{4\rho g}}. \tag{2.13}$$

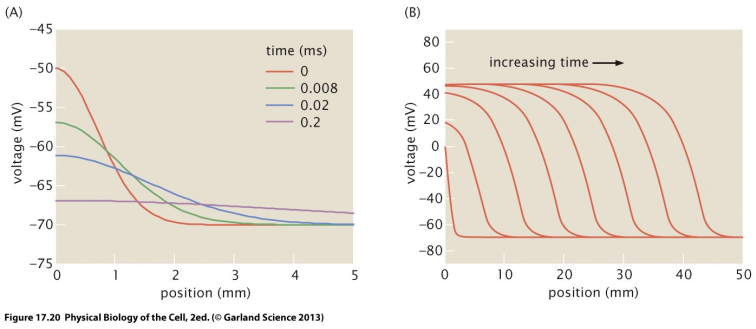


Fig. 2.9 Propagation for an input that overcomes the threshold.

The quantity λ gives the exponential decay of the voltage perturbation, along the axon. It is $\simeq 9\,\mathrm{mm}$ in the squid axon, where $d \simeq 0.5\,\mathrm{mm}$, $\rho \simeq 0.3\,\Omega\mathrm{m}$ and $g \simeq 5\,\Omega^{-1}\mathrm{m}^{-2}$.

2.2.3 Depolarisation Waves

To describe waves on the axon we have to combine the voltage bistability property with the cable equation.

Generalising the current equation to include two types of channels, and the capacitance current, gives:

$$\begin{aligned} i(x-\Delta x,t) \;-\; i(x,t) \;&=\; \Delta g^{Na}_{patch}\,(V(x,t)-V^{Na}_{Nernst}) \;+ \\ +\;\; \Delta g^{K}_{patch}\,(V(x,t)-V^{K}_{Nernst}) \;&+\; \Delta C_{patch}\frac{\partial V(x,t)}{\partial t}. \end{aligned} \tag{2.14}$$

We are again making the simplification that potassium conductance is a constant with voltage and low. The sodium conductance has the form described earlier, and values that switch from below the potassium to very high, increasing voltage.

The resulting equation, generalising the Cable equation to give time dependence, is:

$$\lambda^2\frac{\partial^2 V(x,t)}{\partial x^2}-\tau\frac{\partial V(x,t)}{\partial t} \;=\; (V(x,t)-V^{K}_{Nernst})+\frac{g_{Na}(V(x,t))}{g_K}(V(x,t)-V^{Na}_{Nernst}), \tag{2.15}$$

with λ as above, and $\tau = C/g$ with the conductances set by potassium (dynamics of the RC circuit).

The l.h.s. of Eq. 2.15 is the diffusion equation. If we imagine a localised current injected at $x = 0$ in a small region, then there can be two cases. (a) If the membrane voltage remains below the threshold to open the sodium channels. Then the r.h.s. of Eq. 2.15 remains small (negligible) and we simply have a diffusive relaxation of the voltage, see Figure 2.9(A). (b) If the voltage rises above the threshold to open the sodium channels, then the membrane potential near $x = 0$ changes to the Nernst potential of sodium. This then leads to a propagating front, see Figure 2.9(B).

The speed is λ/τ, and we had $\lambda \sim \sqrt{d}$, so large axons propagate signals faster.

2.3 A Hodgkin-Huxley model for spike propagation

A typical signal traveling in an axon has a conserved and local envelope - this is called a spike. To describe a spike, the missing concept from point (b) above is the inactivation of the sodium channels. In the membrane, sodium channels remain open for about 2 ms. After closing, they remain 'inactive' for few tens of ms. The voltage gated potassium channels are slower to open (several ms) but remain open as long as depolarisation is maintained. This time-asymmetry in the sodium response leads to uni-directional motion of the signal down the axon.

Information is coded as a frequency of spikes. i.e. a more intense input current is turned into a more rapid train of spikes compared to a low input current, with the shapes of the spikes being quite comparable. This process turns an analog input into a digital signal.

We now add inactivation of the sodium channels, with the concept of the inactive state, to complete the model of Eq. 2.15. There are three states for the sodium channel, and the transitions between them can be described by:

$$\begin{aligned} \frac{dp_C}{dt} &= -k_{open} p_C \\ \frac{dp_O}{dt} &= k_{open} p_C - k_{inactive} p_O \\ \frac{dp_I}{dt} &= k_{inactive} p_O . \end{aligned} \tag{2.16}$$

Note that we have not allowed in Eqs. 2.16 inactive→closed nor open→closed, so this simplification can only describe a single spike.

We take the rate of opening to be proportional to the probability of an open channel (same logic as elsewhere in the course, in regulating gene expression):

$$k_{open} = k_{open}^{max} \frac{1}{1 + e^{\beta q (V^* - V(x,t))}}, \tag{2.17}$$

and the inactivation rate as a constant.

The sodium conductance now depends on the state of the channels, and has the form:

$$g_{Na} = g_{Na}^{open} p_O + g_{Na}^{closed} p_C . \tag{2.18}$$

Crucially for the system to work, the parameters have evolved such that the open state conductivity of sodium channels is about

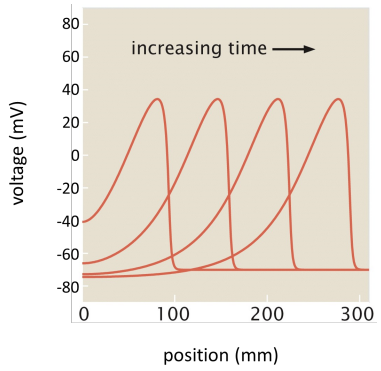


Fig. 2.10 Numerical solutions of the time dependent cable equation, with inactivation of the sodium channels.

20 times larger than g_K (potassium), and the closed state is about 20 times lower than potassium. This more complex sodium conductivity, function of the voltage, and hence space and time, needs to be considered in Eq. 2.15. The equation is readily solved numerically and the solutions are traveling spikes as in Figure 2.10.

Molecular Motors

3

3.1 Introduction

Living biological systems all share the same properties of being able to transform one kind of energy, normally chemical energy, into another in order to do work. For this purpose, over the course of evolution, cells have developed a large collection of molecular machinaries, each with different purposes. In this chapter, we will focus on one category of these: molecular motors.

We have mentioned at the beginning of the course that the cell is a very crowded environment and diffusion is often not sufficient to deliver molecules across the cells in a biological meaningful timescale. To bypass this problem and others, cells employ molecular motors that move *non randomly*. Non-random movement costs energy and these molecular motors are able to transform chemical energy, often in the form of ATP into kinetic energy.

Molecular motors are often classified as follows with some examples:

- Cytoskeletal motor proeins: myosin, kinesin, dynein
- Polymerization motors: actin, microtubules, RecA
- Ion pumps: Na-K pump
- Rotary motors: ATP-synthase, bacterial flagellar motors
- DNA motors: RNA polymerase, helicases

These motors perform many different functions, such as the contraction of stress fibers that give cells a certain shape and muscle contraction, cell motility, separation of the chromosomes in the two daughter cells, transport of cargo across the cell, etc...

The structure of these motors also come in different shape, however most of them exhibit head-domains, where ATP binds. Hydroplysis of ATP results in a conformational change, which is then guided and amplified into a larger structural change of the whole molecule leading to movement. For instance, kinesin and dynein move along the microtubules, which have a well-defined polarity, which determines the directionality of the movement of these motors. Many motors also have a cargo-binding domain where passive cargo can bind and then be actively transported (Fig. 3.1).

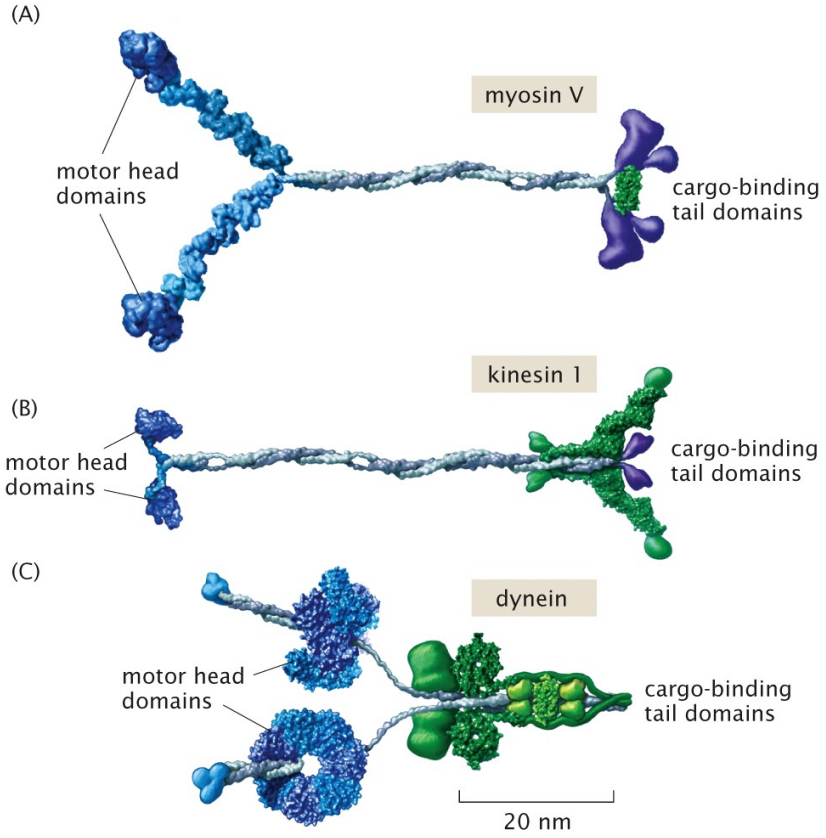


Figure 16.2 Physical Biology of the Cell, 2ed. (© Garland Science 2013)

Fig. 3.1 Key classes of translational motors

3.2 Energy source: ATP

Many molecular motors use hydrolysis of ATP as an energy source, which is then converted to motion. One hydrolysis event generates approximatly 20-30k_BT of energy, where T is temperature and k_B is the Boltzmann constant. Not all this energy will be transfered, as the process is not 100% efficient. The products of the chemical reaction, ADP and the orthophosphates, sit at a much lower energy. This is partially because the orthophosphates group is stabilized by multiple resonance structure and the electrostatic repulsion is reduced going from ATP to ADP.

ATP is obtained through the process of cellular respiration by glucose metabolization. One glucose molecule will produce of the order of 30 ATP molecules, by going through glycolysis (+7-9 ATP), the oxidative decarboxylation of pyruvate (+5 ATP), and the Krebs cycle (+20 ATP). In particular, at different points in this process, oxidative phosporilation occurs. It involves the electron transport chain that establishes a proton gradient across the boundary of the inner membrane by oxidizing NADH produced in the Krebs cycles. The resulting proton-motive force is a combination of an electrostatic term caused by the electrical potential

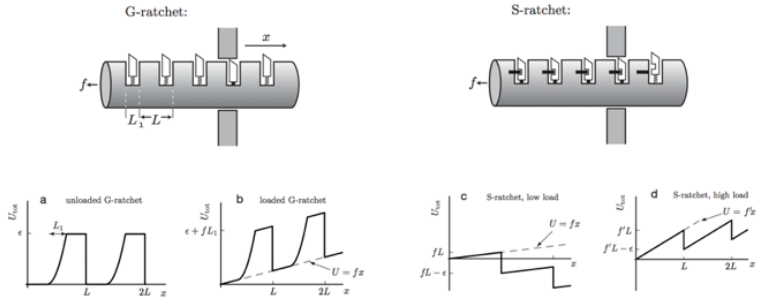


Fig. 3.2 Schematic of G-ratchet and S-ratchet.

gradient and a diffusion term caused by a difference in ion concentration.

$$pmf = V_m + \Delta\mu/e$$

3.3 Physical models of motors

To model the dynamics of molecular motors we will think of them as random walkers moving in an energy landscape, in which the motor can assume different internal states.

3.3.1 Thermal rachet

Molecular machines live in a world dominated by thermal fluctuations. These thermal fluctuations can be exploited to jump over energy barriers.

Imagine a machine that when it encounters an energy bump, only needs to wait some time for a thermal fluctuation to come and allow it to jump over the barrier. We can model this as a shaft with a series of beveled bolts mounted on springs that prevents the shaft to move to the left (G-ratchet in fig. 3.2). The shaft also pulls a load represented by a force f directed to the left. Occasionally, large enough thermal fluctuations will be able to kick the shaft with energy greater than $\epsilon + fL$ required to jump the next bolt and the shaft will move to the right. If you imagine to wrap the shaft around a circle, this machine would be able to constantly extract work from the surrounding thermal motion and thus violate the second law of thermodynamics.

How can we resolve this paradox? As shown in fig 3.2 if k_BT is comparable to ϵ and sufficient to pull the weight to the right, then it will also be sufficient to allow the shaft to retract because the bolt springs will be subject to thermal fluctuations. So, effectively, the shaft will be pulled in the direction of the force.

However, let's modify the shaft so that there are latches that keep the bolts down if the bolt is to left of the wall, and releases it if it is on the right side (S-ratchet in fig. 3.2). In this case, it will be the potential energy stored in the compressed bolts that will be used to pull a cargo right as long as the springs are stiff enough. Let's formalize this mathematically.

Imagine that the shaft is wrapped around a circle and it has n bolts, $P(x)$ denotes the probability that a ratchet will at be position x at steady-state and we are tracking M ratchets. At each time step Δt a ratchet can get a thermal kick that can move it to the right or to the left. There will $MP(x)\Delta x$ ratchets between position $x - \Delta x/2$ and $x + \Delta x/2$ and half of them will step to the right in time Δt. Similarly, there will be $MP(x+\Delta x)\Delta x$ ratchets between $x + \Delta x/2$ and $x + 3\Delta x/2$ and half of them will step to left in time Δt. Then the number of ratchets that cross position $x + \Delta x$ from left to right is

$$1/2M[P(x) - P(x + \Delta x)]\Delta x \approx -1/2\Delta x^2 M\frac{dP}{dx} = -DM\frac{dP}{dx}\Delta t.$$

where $D = \Delta x^2/2\Delta t$ is the diffusion coefficient.

Let's define the $F = -dU_{tot}/dx$. In the absence of any diffusive motion, this force would impart a drift velocity given by $v_{drift} = F/\zeta$ where ζ is the friction coefficient. By using Einstein relation, we can rewrite this as

$$v_{drift} = -\frac{D}{k_b T}\frac{dU_{tot}}{dx}.$$

Then the total number of ratchets crossing x in time Δt from the left due to drift will be $MP(x)v_{drift}$. Therefore in total, we have to contributions and the total flux of ratchets from left to right is

$$j^{1D} = -MD(\frac{dP}{dx} + \frac{1}{k_B T}P\frac{dU_{tot}}{dx}).$$

Note that at equilibrium, the flux goes to zero. The equilibrium probability distribution of the ratchets in the external field will be the Boltzmann distribution

$$P_{eq}(x) = P_0 e^{-U_{tot}/k_B T}$$

Because the number of ratchets has to be conserved $\partial_t P = -\partial_x j/M$ and therefore

$$\partial_t P = D[\partial_x(P\partial_x U_{tot}/k_B T) + \partial_x^2 P]$$

called the Smoluchowski equation.

3.3.2 Polymerization ratchet

Polymerization motors can also be modeled as ratchets. The basic idea consists in the fact that a growing filament can result in a

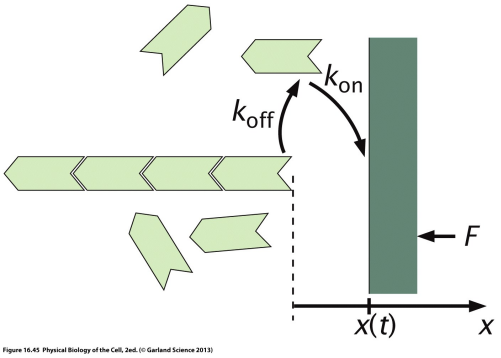


Fig. 3.3 Sketch of a filament reaching a wall

pushing force on a resisting barrier by virtue of fluctuations in the position of the barrier or the filament. Imagine a filament that has grown in order to reach a barrier (cell wall). If the barrier or the filament jiggle, then a new monomer may be able to squeeze in (fig. 3.3).

For now, let's neglect the action of the force. Let's assume that each monomer has length δ and can be inserted if the distance between the tip of the filament and the barrier reaches such size. This is a first passage type problem, which we are going to solve using an analogy. Imaging that you have a particle that is diffusing along a distance δ and you want to determine the first time it reached position δ if it starts at 0. The current due to diffusion will be

$$j_0 = -D\partial_x P$$

where P is the probability of finding the particle at position x. At steady-state, we know that

$$\partial_x^2 P = 0$$

meaning that $P = Ax + B$. To compute A and B, we exploit the fact that $P(x)$ has to be normalized:

$$\int_0^\delta P(x)dx = 1 = A\delta^2/2 + B\delta$$

and that $P(\delta) = 0$. With these two constraints, we find that $A = -2/\delta^2$ and $B = 2\delta$.

The current is then

$$j_0 = -DA = 2D/\delta^2$$

and the first passage time is the inverse of this quantity

$$\tau_1 = \delta^2/2D.$$

Then the rate of polimerization will be

$$v = \delta/\tau_1 = 2D/\delta.$$

Now, if we apply a constant force F, the current will become

$$j_0 = -\partial_x P - P\frac{F}{\zeta}$$

where the friction coefficient $\zeta = k_B T/D$. At steady-state the current is constant and the differential equation has general solution

$$P(x) = Ae^{-Fx/k_B T} - j_0\zeta/F.$$

Similarly to before we use the two constraints $P(\delta) = 0$ and $\int_P(x)dx = 1$ to find A and j_0. The result is that the steady-state current is

$$j_0 = [(k_B T\zeta/F^2)(e^{F\delta/k_B T} - 1) - \zeta\delta/F]^{-1}$$

and the growth rate of the filament will be

$$v = \delta j_0 = \frac{D}{\delta}\frac{(F\delta/k_B T)^2}{e^{F\delta/k_B T} - 1 - F\delta/k_B T}.$$

3.4 Linear motors

3.4.1 One state model

For translational motors, we will start by assuming that the motor has no internal states and simply jumps from one position to another with a forward rate k_+ and a backward rate k_- over a discretized one-dimensional filament. These rates are related to the energy potential the motor is moving in U, which is related to the force $U = fx$. For the motor to move, the rates k_- and k_+ will have to be different.

Let's define $p(n,t)$ the probability of finding the motor at position n at time t. Then the probability at time $t + dt$ will be

$$p(n,t+dt) = k_+ dtp(n-1,t)+k_- dtp(n+1,t)+(1-k_+ dt-k_- dt)p(n,t)$$

By sending dt to 0 and defining positions as $x = na$, we find the following master equation:

$$\partial_t p = (k_- - k_+)a\partial_x p + \frac{1}{2}(k_+ + k_-)a^2\partial_x^2 p$$

which corresponds to a diffusion equation with drift with diffusion coefficient $D = (k_+ + k_-)a^2/2$ and drift velocity $V = (k_- - k_+)a$.

The two rates are related to the difference in energy between going forward versus backwards. If ATP hydrolysis allows to take step forward, then

$$\frac{K_+}{k_-} = e^{-\beta\Delta G_h}$$

where ΔG_h corresponds to the free energy from hydrolysis. In the presence of a force that pulls the motors backwords, than the

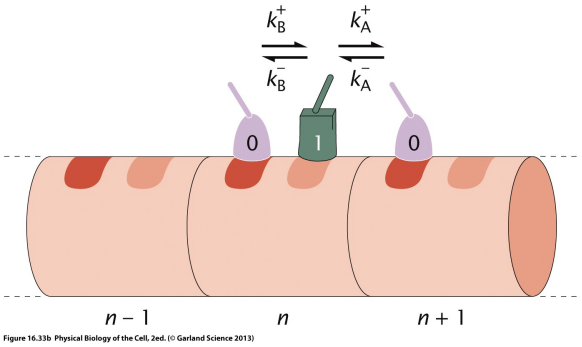


Fig. 3.4 Schematic of a two-state model

energy of the step forward will be raised by a factor Fa and the reaction coefficients will satisfy

$$\frac{K_+(F)}{k_-(F)} = e^{-\beta(\Delta G_h + Fa)}$$

The free energy released by hydrolysis is

$$\Delta G_h = \Delta G_0 + k_B T \ln\left(\frac{[ADP][P_i]}{[ATP]}\right)$$

where ΔG_0 is the energy released by breaking a chemical bond in the ATP molecule. This results in

$$\frac{k_+}{k_-} \propto [ATP]$$

which predicts either a velocity that saturates with $[ATP]$ if all the dependence is in the forward direction, or that is linear with $[ATP]$ if all the dependence is in the backward direction.

3.4.2 Two state model

Some motors do not seem to follow the simple one state model above and multiple intermediate states have to be considered. The simplest case is a two-state model where the motor, in addition to moving forward and backward, can be in two conformational states and these states have to alternate in order to provide movement.

In this case, the probability distribution $p_i(n,t)$ will refer to the distribution of the two internal states the motor can be in. Let's define k_A^+ the rate at which state 1 will convert to state 0 and move forward. Similarly, k_B^+ will be the rate at which state 0 will convert to state 1 and move forward. Analogous definitions will be for k_A^- and k_B^- (See fig. 3.4). Then the master equations are

$$\begin{cases} \frac{dp_0}{dt} = k_A^+ p_1 + k_B^- p_1 - k_A^- p_0 - k_B^+ p_0 \\ \frac{dp_1}{dt} = k_A^- p_0 + k_B^+ p_0 - k_A^+ p_1 - k_B^- p_1 \end{cases}$$

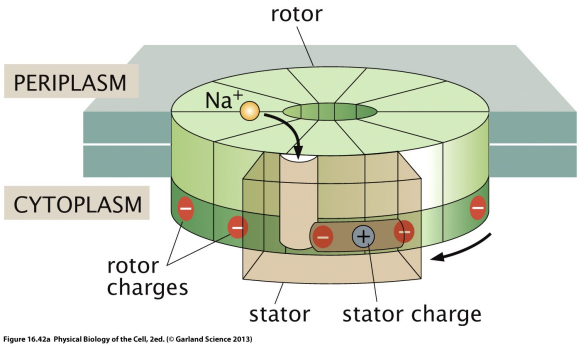


Fig. 3.5 Typical structure of a rotary motor

At steady state, we find that

$$(k_A^+ + k_B^-)p_1 = (k_A^- + k_B^+)p_0$$

Because $p_0 + p_1 = 1$ then

$$\begin{cases} p_0 = \frac{k_A^+ + k_B^-}{k_A^+ + k_A^- + k_B^+ + k_B^-} \\ p_1 = \frac{k_A^- + k_B^+}{k_A^+ + k_A^- + k_B^+ + k_B^-} \end{cases}$$

Let's assume that when the motor transitions from 0 to 1 it travels a distance δ and when going from 1 to 0 it travels a distance $a - \delta$. Then the net velocity will be

$$v = \delta(p_0 k_B^+ - p_1 k_B^-) + (a - \delta)(p_1 k_A^+ - p_0 k_A^-)$$

By replacing the solutions for p_0 and p_1, we find the average velocity

$$\langle v \rangle = a \frac{k_A^+ k_B^+ - k_A^- k_B^-}{k_A^+ + k_A^- + k_B^+ + k_B^-}$$

Note that the velocity does not depend on the distance δ traveled by the motor when transitioning between the two states.

3.5 Rotary Motors

Similar models to the ones developed in the previous section can be applied to rotary motors by imagining to wrap the track around a circle. Also the motors move through discrete steps that are coupled with energy-realising reactions, such as ATP hydrolysis or transport of ions down an electrical gradient. In rotatory motors, instead of talking of linear speed, we talk about angular speed, instead of force, we talk about torque.

The motor is driven by thermal fluctuations that result in rotational diffusion of the rotor. The diffusion is rectified, leading to directed motion by electrostatic between the charges on the rotor, the stator and the Na^+ ions. A rotor charge is captured by

the stator charge, which is opposite in sign. It will then diffuse away until it finds itself in a input channel, where it is exposed to a high concentration of Na^+. The driving force is then the free-energy difference experienced by the ion as it travels through the membrane (fig. 3.5).

One special example of rotary motors are bacteria flagella. The energy source of these motors is the proton-motive force. Most models that explain the behavior of these motors try to reproduce the relationship between torque and speed found in experiments. One example of such models by Xing et al describes the kynetic of the motor with four physical ingredients: (i) load and motor are connected through a soft elastic linkage, (ii) motor rotation is tightly coupled to ion flux, (iii) motor rotation is driven by proton driven conformation changes and (iv) the proton channel in the stator is gated by rotor movement.

Biopatterning

4

4.1 Introduction to Dynamical Systems

4.1.1 Elements of non-linear dynamical systems

We will focus on concrete examples in the context of gene expression, but let us first introduce some of the general framework and useful tools that have been developed in general for the study of non-linear dynamics. We follow here the monograph by Strogatz (Strogatz, 2014).

The dynamics of a general non-linear system can be described by a set of coupled differential equations

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, ...x_n) \\ &\vdots \\ \dot{x}_n &= f_n(x_1, ...x_n). \end{aligned}$$

For example, damped harmonic motion with the second order (linear) DE

$$m\ddot{x} + b\dot{x} + kx = 0$$

can be written a set of coupled first-order equations as

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= -\frac{k}{m}x_1 - \frac{b}{m}x_2. \end{aligned}$$

We examine first the one-variable system ("flow on the line"), then two-variable systems ("flow on the plane") where oscillations and limit cycles can exist. For reasons of time we do not touch here on three-variable system ("3-D flow") which is the minimal situation to exhibit chaotic dynamics. In general, an n-variable system requires n equations to represent it. Many very interesting biological systems can be simplified to two variables.

Flows on the line

We start with an examination of the possible *trajectories* of a system. That is, we plot the path in a 2n-dimensional space, where the dimensions are the n independent coordinates and their corresponding momenta. Here, we take a fairly loose view of this definition, and we will generally just use the independent coordinates and their time- derivatives. We begin by examining the one-dimensional flow, that is, the dynamics of a single first-order DE,

$$\dot{x} = f(x).$$

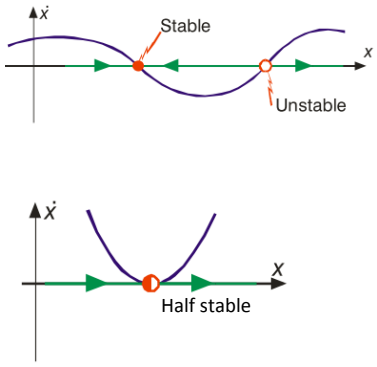


Fig. 4.1 Illustrations of the types of fixed points in 1-D systems. Note the notation: stable fixed points are denoted by filled circles; unstable fixed points by open circles, and half-stable points by half- filled circles, as shown in the examples. Note the notation: stable fixed points are denoted by filled circles; unstable fixed points by open circles, and half-stable points by half- filled circles, as shown in the examples.

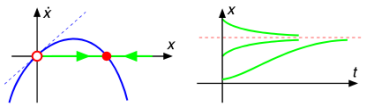


Fig. 4.2 Fixed points of the autocatalytic system.

Fixed points of a 1-D flow. The function f is single-valued for all x. The dynamics therefore take place along a line (the x axis). In the notation of Strogatz, the phase-plane plot represents a vector field on the line: the velocity vector $\dot{x}$ is shown for every x. The trajectory is a plot of $\dot{x}$ as a function of x. The time coordinate is thus implicit - we could, for example, mark off time ticks along the curve given any starting value of x, and hence $\dot{x}$, but the main properties of the system are apparent directly from the phase-plane plot.

We can immediately identify two types of *fixed point*. These are values of x for which x is zero, so that the system is, momentarily at least, at rest.

- A *stable* fixed point results whenever $\dot{x}$ is zero and the slope of the $\dot{x}$ vs x curve $d\dot{x}/dx$ is negative. This ensures that for small fluctuations away from the fixed point, as shown in green arrows on the plot, the velocity $\dot{x}$ is in a sense to bring the system back to the fixed point. A stable fixed point is also known as a *sink* or an *attractor*.
- An *unstable* fixed point, on the other hand, has $d\dot{x}/dx > 0$, so that small fluctuations result in a motion directed away from the fixed point. Other names for an unstable fixed point include *source* or *repeller*.
- One other type of fixed point is possible, and is known as a half-stable point.

Example of Autocatalytic chemical reaction. Consider the reaction

$$\mathrm{A} + \mathrm{X} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} 2\mathrm{X}$$

which is a non-linear dynamical system. The presence of X stimulates further production of X hence the term "autocatalytic". (This is one model for the growth of amyloid plaques in the brain in diseases such as BSE and CJD: the presence of a small amount of plaque, X, catalyses the conversion of normal protein, A, to plaque.) There are two variables in the process: a, the concentration of reactant A, and x, the concentration of reactant X. If the concentration of A is always large, then it will be effectively constant. The problem then reduces to dynamics in one variable.

Given the rate constants for forward and reverse reactions, k_1 and k_2, the equation governing the dynamics is

$$\dot{x} = k_1 a x - k_2 x^2.$$

We can sketch the trajectory in the phase-plane, as shown. It is also straightforward to sketch the concentration vs time, as in the right hand panel. Since $\dot{x}$ is linearly proportional to x in the vicinity of the fixed points, the approach to equilibrium must be exponential.

Dynamic variables and control variables. In the example above, x and a are *dynamical variables*: that is, they are the

variables which change with time. The two other variables, k_1 and k_2, are control variables. In that particular case, varying the control variables did not change the general character of the dynamics, but only the details.

Consider now the system described by

$$\dot{x} = x^2 + a.$$

As a is increased from a negative value, the two equilibria one stable, and one unstable first approach each other, then merge to form a half-stable fixed point, and finally annihilate. The control parameter, or variable, a, thus determines the stability of the system.

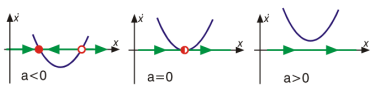


Fig. 4.3 In this example, a is control variable. Its value determines the stability of the system.

In general, complex dynamical systems have fewer control parameters than dynamical variables. We are interested in situations, such as that shown above, where a change in one or more of the control parameters leads to discontinuities i.e., qualitatively different dynamics, such as a change from stable to unstable behaviour. This is the basis of Catastrophe Theory. The key result from catastrophe theory is that the number of configurations of discontinuities depends on the number of control variables, and not on the number of dynamical variables.

In particular, if there are four or fewer control variables, there are only seven distinct types of catastrophe, and in none of these is more than two dynamical variables involved. In the next section we consider all cases up to two control parameters. For simplicity we restrict ourselves to a single dynamical variable, x, with little loss of generality.

Potential methods. The existence of stable, unstable and half-stable fixed points (i.e. equilibria) suggests another way of looking at the dynamics, in terms of an underlying potential, which we shall here denote by $V(x)$. Stable equilibria are local minima in $V(x)$, unstable equilibria are local maxima and half-stable fixed points are points of inflection.

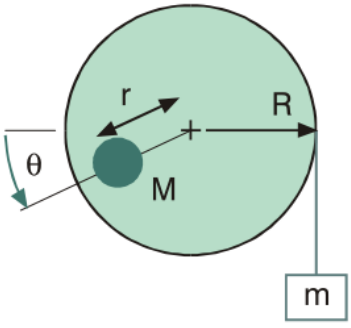


Fig. 4.4 Mechanical pulley example Here,a is control variable. Its value determines the stability of the system. In this weighted pulley, The gravitational potential is given by $V = mR\theta - Mr\sin\theta$, we can simplify notation $V = A\theta - B\sin\theta$. For small θ we can approximate this as $V \simeq (A-B)\theta + \frac{B}{6}\theta^3$. That has the same behavior as $V = \alpha\theta + \theta^3$, with $\alpha = 6(A-B)/B$. The system will thus be stable as long as $\alpha < 0$, i.e. $B > A$, i.e. $Mr > mR$.

In this course we are dealing with the evolution of arbitrary dynamical systems (as loosely interpreted), and hence there may not actually be a true potential energy. In mechanical systems there often is one. In terms of the equation $\dot{x} = f(x)$, we can define the potential to be

$$f(x) = -\frac{dV}{dx}.$$

For a first-order system (and hence one-dimensional motion) we have to imagine a particle with an inertia which is negligible in comparison with the damping force.

The negative sign implies that the force on a particle is always "downhill", towards lower potential. This can be shown simply by applying the chain rule to the time-derivative of the potential

and applying the definition of the potential:

$$\begin{aligned} \frac{dV}{dt} &= \frac{dV}{dx}\frac{dx}{dt} \\ &= -\left(\frac{dV}{dx}\right)^2 \leq 0. \end{aligned}$$

Thus $V(t)$ decreases along trajectories, and the particle always moves towards lower potential.

In summary, the potential has the following properties:

(1) $-dV/dx$ is force-like (i.e., is in the direction of motion).

(2) Equilibrium positions, x^* (fixed points) are given by $-dV/dx = 0$.

(3) The stability of the fixed point is determined by the sign of $-d^2V/dx^2|_{x^*}$.

Forms of the potential curve. The potential function can always be approximated by a Taylor series, so that

$$V(x) = a + bx + cx^2 + ...$$

We can ignore a, since it is just a constant and does not affect the dynamics. In the vicinity of a single fixed point (i.e. equilibrium) we can also eliminate b by shifting the coordinate system to put the fixed point at the origin (although b cannot be ignored for multiple fixed points). This leaves us with

$$V(x) = cx^2 + dx^3 + ex^4 + ...$$

We can now enumerate the possibilities.

(1) **Harmonic Potential.** This is the simplest possible form, and the only one possible for purely linear systems:

$$V(x) = \alpha x^2.$$

There is a single fixed point, $x^* = 0$, for all α. If $\alpha > 0$ then the fixed point is stable; if $\alpha < 0$ then it is unstable.

(2) **Asymmetric cubic potential: The saddle-node bifurcation.** The potential has the form

$$V(x) = \alpha x + x^3.$$

For $\alpha > 0$, no equilibrium position is possible. For $\alpha < 0$, then there is always one stable and one unstable equilibrium. Here we introduce the idea of *control space*. We can plot the location of the fixed point, x^* , as a function of the control parameter, α, as shown in Figure 4.5.
On the control space plot, the solid line denotes the location of the *stable* equilibrium, while the dashed line indicates the locus of the *unstable* equilibrium, both as a function of α.

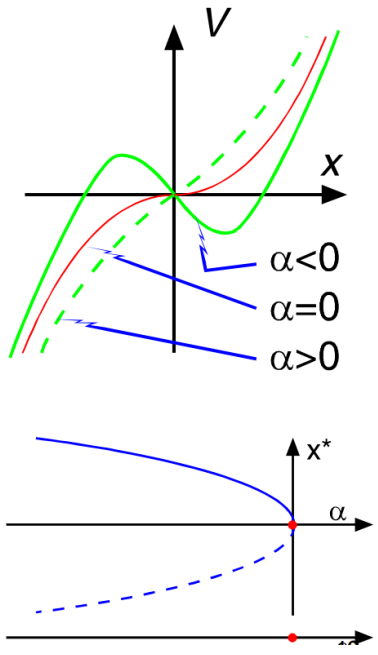


Fig. 4.5 cubic and linear potential

The form of the instability shown here is what Strogatz calls a saddle-node bifurcation, and sometimes known as a limit point instability or a fold.
The phase-plane trajectories for this system were shown earlier, for the system with $\dot{x} = x^2 + a$ (Figure 4.3). This is the origin of the term "saddle-node bifurcation" as a is decreased through zero the fixed point is first created, and then bifurcates into two: one stable and one unstable. See example in Figure 4.4.

(3) **Cubic potential with quadratic term: The transcritical bifurcation.** The potential this time includes a term in x^2 rather than a linear term as in the previous section.

$$V(x) = x^3 + \alpha x^2$$

The effect of this is to give a double root, and hence a fixed point, at the origin, regardless of the location of the third root.
The bifurcation diagram is shown in Figure 4.6. This is generally known as the transcritical bifurcation. One physical example of such a system is the laser.

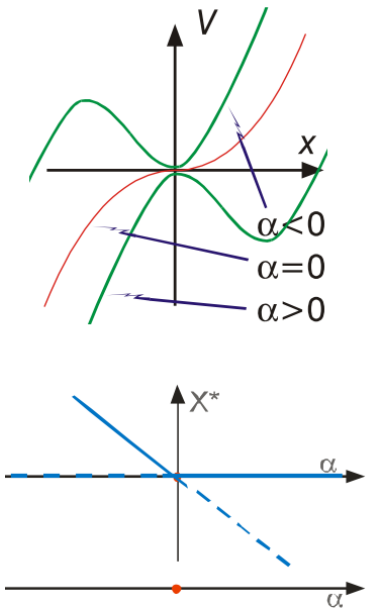


Fig. 4.6 cubic and quadratic potential

(4) **Symmetric quartic potential: The pitchfork bifurcation.** The potential is:

$$V(x) = x^4 + \alpha x^2.$$

Two cases:

- For $\alpha \geq 0$ there is just one stable equilibrium;
- For $\alpha < 0$ there is one unstable equilibrium and two stable equilibrium points.

Plotted in Figure 4.7 is the case of positive term on the 4^{th} power. In this case we refer to the Stable Symmetric Transition. It is also known as a Pitchfork Bifurcation (see Strogatz) from the shape of the bifurcation diagram, as shown at right. One example of this sort of potential is the Euler strut.
If we take the negative sign on the 4^{th} power, the additional quartic term may also act to destabilize the system, and the locus of the fixed points changes qualitatively (exercise).

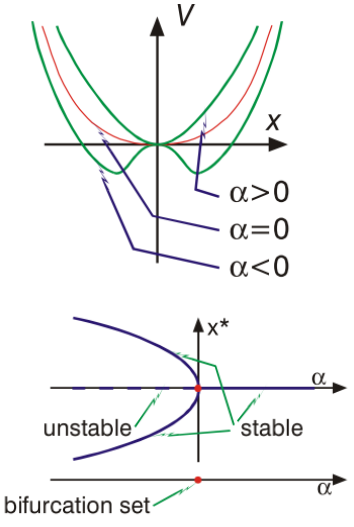


Fig. 4.7 symmetric quartic potential

(5) **Asymmetric quartic potential with two control parameters: the Cusp catastrophe.** We now consider an asymmetric potential, of the form

$$V(x) = \alpha x^2 + x^4 + \beta x$$

where the βx term introduces asymmetry to the symmetric quartic form of the previous case. We now have two control parameters, α and β. Depending on the sign of α, then, we get two different sorts of behaviour.

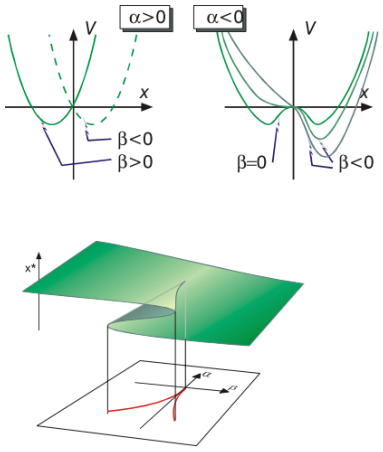


Fig. 4.8 asymmetric quartic potential

If $\alpha > 0$ then the linear term merely shifts the position of the fixed point, but does not qualitatively change the dynamics from that of a simple harmonic potential. If $\alpha < 0$ however, the linear term can eliminate the unstable fixed points and one of the stable fixed points as well.
The control space diagram the bifurcation set is now two dimensional. Consider the *equilibrium surface*, or a plot of the location of x^* against α and β. The bifurcation set is the set of points in the (α, β) plane dividing the plane into different regions of stability, and has a characteristic cusp shape, see Figure 4.9.
As we move from the shaded to the non-shaded region (i.e. across the bifurcation set), there is a sudden change in behaviour, with marked hysteresis when the path is reversed.

4.1.2 Two dimensional systems

Oscillations are not possible in one-variable systems, so that all the "fixed points" are static. In two variables, we have the possibility of periodic ("closed") trajectories, which are known as limit cycles, as we would predict from our knowledge of the harmonic oscillator a classic two-variable system. Near a fixed point, the general nonlinear differential equations can be approximated by a set of linear equations, so we can examine the behaviour near fixed points of any two-variable system by generalizing the harmonic oscillator equations.

Concepts for 2-variable systems: Phase space and nullclines

We have seen fixed points in one variable. This concept generalises to two variables, and their stability can be determined by linearising the system and looking at eigenvalues (real parts) of the matrix of partial derivatives. Two other important concepts in nonlinear systems are phasespace and nullclines.

For 2 coordinates, the phaseplane is a sketch of the system time evolution in $(x; y)$ coordinates. This is different from a sketch against time which is the way you often see data. Phaseplanes can provide much more useful info such as a global way of looking at your system.

Nullclines are curves that enable one to break the plane into regions of different qualitative behaviour. If 2 variables $\dot{x} = f(x, y)$, $\dot{y} = g(x, y)$, then nullclines are $f(x, y) = 0$ and $g(x, y) = 0$, and their crossing points are the fixed points.

Nature of stable solutions in 2-variable systems

Non-linear systems can, in the vicinity of a fixed point, be linearised. As shown in the lecture slides, linear systems have only

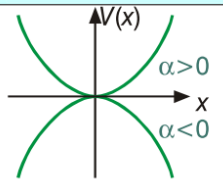
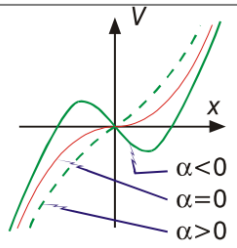
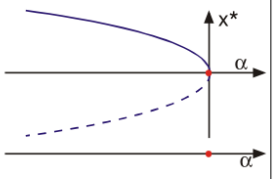
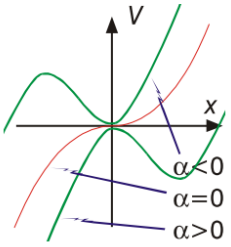
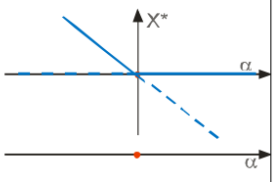
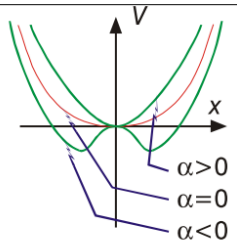
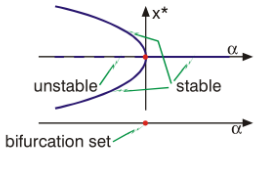
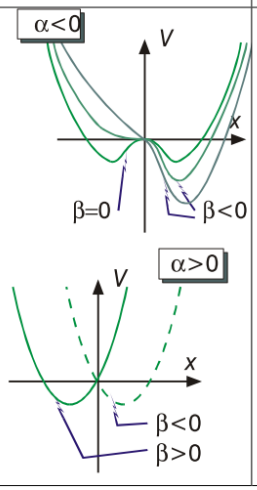
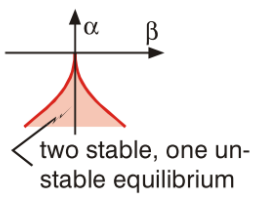
Name	f(x)	V(x)	Potential	Bifurcation diagram
Harmonic potential	$-2\alpha x$	αx^2		
Limit-point instability Saddle-node bifurcation Fold	$-\alpha - 3x^2$	$\alpha x + x^3$		
Transcritical Bifurcation	$-2\alpha x - 3x^2$	$\alpha x^2 + x^3$		
Pitchfork Bifurcation Stable symmetric transition	$-2\alpha x - 3x^2$	$\alpha x^2 \pm x^4$		
Cusp catastrophe	$2\alpha x + 4x^3 + \beta$	$\alpha x^2 + x^4 + \beta x$		

Fig. 4.9 summary of instabilities in one dimensional systems

five types of solution: spiral (stable or unstable; node (stable or unstable), and saddle point.
A non-linear system can also exhibit a different solution, known as 'limit cycle'. The "limit cycle is defined as an isolated, closed trajectory. That is, neighbouring trajectories are not closed. It is an attractor if neighbouring trajectories spiral towards it, or a repeller if they spiral away.

4.2 Biopatterning

In the following, we will explore how self-organizing patterns at a macroscopic scale emerge from the microscopic dynamics of the components of a biological system. How do zebra stripes emerge? Or the symmetric geometry of flowers? Or the ordering of lens-making cells in the eye? Biology exhibits an incredible number of examples of regular patterns that can occur in both space and time.

Generally speaking, pattern formation can be treated as a question in communication of spatial information. We can think of four different mechanisms that can lead to patterning: (i) Turing patterns, (ii) morphogen gradients, (iii) lateral inhibition and (iv) clock and wavefront mechanisms. Here, we will focus on patterns in space generated by the first three types of mechanisms.

4.3 Turing Patterns

Patterning can arise spontaneously in a system due to the presence of instabilities in the system. An example of this are Turing patterns. The occurrence of these patterns was hypothesized by Turing as a mechanism to explain the spontaneous emergence of morphogen patterns in development. Although it turned out that biological systems seem to rely on different mechanisms during morphogensis, it remains true that many patterns of gene expression can be driven by reaction-diffusion equations as those describing Turing pattern.

Let's imagine to have to molecules, X and Y, distributed in space, which can react with each other and diffuse. Their dynamics can be mathematically formalized with a system of partial differential equations:

$$\partial_t X = D_X \partial_x^2 X + f(X,Y) \tag{4.1}$$
$$\partial_t Y = D_Y \partial_x^2 Y + g(X,Y) \tag{4.2}$$

with initial conditions

$$X(x,0) = X_0(x) \tag{4.3}$$
$$Y(x,0) = Y_0(x) \tag{4.4}$$

and fixed boundary conditions.

For the sake of brevity, we will introduce a vector notation where

$$\vec{X} = \begin{bmatrix} X \\ Y \end{bmatrix}, \; F(\vec{X}) = \begin{bmatrix} F(X,Y) \\ G(X,Y) \end{bmatrix}, \; D = \begin{bmatrix} D_X & 0 \\ 0 & D_Y \end{bmatrix}$$

Let's assume that in the absence of diffusion the system has steady-state (X^*, Y^*) and let's examine the stability of the steady-state by moving slightly away. If

$$\vec{Z} = \begin{bmatrix} X - X^* \\ Y - Y^* \end{bmatrix}$$

then the time derivateve of $\vec{Z}$ is

$$\dot{\vec{Z}} = \begin{bmatrix} f(X - X^*, Y - Y^*) \\ g(X - X^*, Y - Y^*) \end{bmatrix}$$

We can linearize these equations (first term of the Taylor expansion) and we get:

$$\begin{bmatrix} f(X - X^*, Y - Y^*) \\ g(X - X^*, Y - Y^*) \end{bmatrix} = \begin{bmatrix} f(X^*, Y^*) \\ g(X^*, Y^*) \end{bmatrix} + A\vec{Z} = A\vec{Z},$$

where we exploied the fact that (X^*, Y^*) is steady-state and

$$A = \begin{bmatrix} \partial_X f & \partial_Y f \\ \partial_X g & \partial_Y g \end{bmatrix}$$

is the Jacobian of F. We then get

$$\dot{\vec{Z}} = A\vec{Z}.$$

The solution of this equation will of of the form

$$\vec{Z} = \sum_{n=1}^{2} w_i \vec{v}_i \exp(\lambda_i t)$$

where $\vec{v}_i$ are the eigenvectors of A with eigenvalues:

$$\lambda_i = \frac{Tr(A) \pm \sqrt{Tr(A)^2 - 4Det(A)}}{2}$$

and w_i depend on the initial conditions.

If at long time scales, $\vec{Z}$ tends to zero, then the steady-state will be stable. This happens if both eigenvalues are negative, which is equivalent to $Tr(A) < 0$ and $Det(A) > 0$.

If we add diffusion, the equation we want to solve is

$$\dot{\vec{Z}} = A\vec{Z} + D\partial_x^2 \vec{Z},$$

which admits solutions of the form:

$$\vec{Z} = Z_0 \exp(\lambda t) \exp(iqx).$$

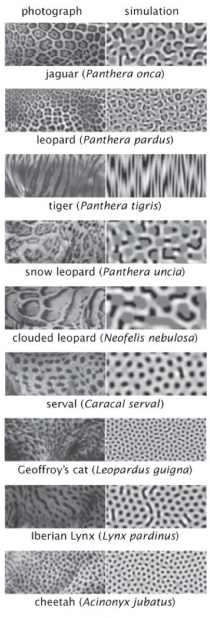


Fig. 4.10 Coat patterns in a variety of species (left) and corresponding simulated pattern (right).

These represent spatial waves with wave number q whose amplitude is either growing or shrinking with time.

By substituting the solution in the equation above, one obtains:

$$(A - q^2 D - \lambda I) Z_0 = 0,$$

where I is the identity matrix. The system is unstable if at least one of the eigenvalues in $A - q^2 D$ has a real positive part, which happens if $Det(A - q^2 D) < 0$.

For a diffusion driven instability to occur, we then have the condition

$$Det(A - q^2 D) = D_X D_Y q^4 - q^2 (D_Y \partial_X f + D_X \partial_Y g) + Det(A) < 0.$$

Both $D_X D_Y q^4$ and $Det(A)$ are positive. Therefore, we need to have

$$D_Y \partial_X f + D_X \partial_Y g > 0$$

and the critical case of pattern emerges when

$$D_X D_Y q^4 - q^2 (D_Y \partial_X f + D_X \partial_Y g) + Det(A) = 0.$$

We can determine the value of q corresponding to the minimum of $Det(A - q^2 D)$. This corresponds to the wave number that grows at the onset of instability. If we define $k = q^2$, then

$$\frac{d(Det(A - kD)}{dk} = 2 D_X D_Y k - (D_Y \partial_X f + D_X \partial_Y g) = 0$$

We find that

$$k_{min} = q^2_{min} = \frac{D_Y \partial_X f + D_X \partial_Y g}{2 D_X D_Y}$$

The second derivative of $Det(A - kD) = 2 D_X D_Y > 0$ proving that this is a minimum.

If we replace this solution back in the condition for instability, we get:

$$\frac{-(D_Y \partial_X f + D_X \partial_Y g)^2}{4 D_X D_Y} + Det(A) < 0$$

and therefore

$$4 D_X D_Y Det(A) - (D_Y \partial_X f + D_X \partial_Y g)^2 < 0.$$

If we define the relative diffusion $\delta = D_Y / D_X$ then the expression above can be written as

$$4 \delta Det(A) - (\delta \partial_X f + \partial_Y g)^2 < 0$$

The root of this equation defines the critical diffusion ratio for the onset of instability (pattern formation).

In the case of finite systems with fixed boundary conditions, only some modes are allowed. For instance, if the system is defined

over $[0, L]$, then the solution can be combinations of only wave numbers $q_n = \frac{2\pi n}{L}$ such that

$$\vec{Z} = \sum_n w_n \vec{v}_n \exp(\lambda_n t) \exp(i q_n x).$$

Any mode will be unstable if

$$D_X D_Y q_N^4 - q_n^2 (D_Y \partial_X f + D_X \partial_Y g) + Det(A) < 0.$$

There are several reasons why to molecules can have different diffusion coefficient: their size, their binding rate to other surrounding molecules, the ability to penetrate through certain compartments. Note that Turing patterns are not limited to small molecules diffusing around. Cell populations that migrate and interact can also give rise to Turing patterns.

4.4 Morphogen gradients

As we mentioned earlier, Turing patterns do not seem to be the mechanism by which pattern formation arises in development. In this case, cells decide their fate using positional information regarding the concentration of a morphogen (see Fig. 4.11 for a simple French flag pattern). The morphogen gradient that cells sense normally has a maternal origin. Interestingly, the resulting patterns are extremely robust to noise in morphogen concentration.

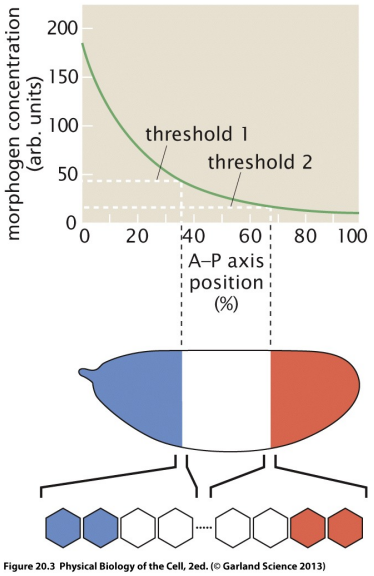


Fig. 4.11 Example of patterning driven by a morphogen gradient

To start, let's consider a morphogen M that diffuses at rate D and degrades at rate α. The morphogen is introduced at time t at position $x = 0$ in concentration M_0. The equation describing its dynamic will be

$$\partial_t M = D \partial_X^2 M - \alpha M,$$

which has steady-state solution $M(x) = M_0 \exp(-x/\lambda)$, where $\lambda = \sqrt{D/\alpha}$.

If cell fate and patterning is determined by the concentration of the morphogen, similarly to the french flag model, than changes in initial concentrations would in this case affect the resulting patterning. For instance, if the boundary between two regions is determined by the position x_0 at which $M(x_0) = T$, then this position would shift by $\delta = \lambda \log \frac{M_0'}{M_0}$, if the initial concentration shifts from M_0 to M_0'.

Let's consider an alternative model where the morphogen promotes its own degradation:

$$\partial_t M = D \partial_x^2 M - \alpha M^2.$$

The steady-state solution of this equation is

$$M(x) = \frac{6D}{\alpha(x + \epsilon)^2},$$

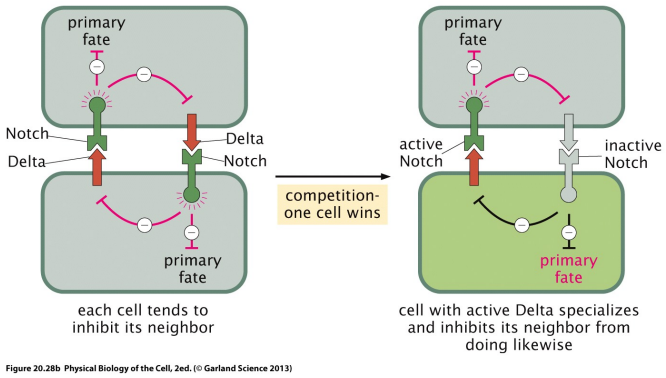


Fig. 4.12 Sketch of the feedback loop between Delta and Notch of two neighboring cells

where $\epsilon = \sqrt{6D/\alpha M_0}$. In this case, as long as M_0 is very large, ϵ will be very small, and $M(x)$ will be almost independent of M_0.

4.5 Lateral inhibition: the Notch-Delta concept

Some biological tissues exhibit specific patterns where neighboring cells have "opposite" morphologies creating a checkboard pattern. Neither reaction-diffusion models nor morphogen gradients can give rise to such phenomenon. However, one can think of a mechanisms by which neighboring cells inhibit each other's gene expression program leading to opposite cell fate. Here, we will consider a well-studied example of such mechanism: the Notch-Delta system.

Notch is a mebrane-bound receptor that is activated by the ligand Delta supplied by a neighboring cell (Fig. 4.12). When the two interact, the activated Notch inhibits the production of the Delta ligand in its own cell. Therefore, the neighboring cell that initially supplied Delta will not have its own Notch receptors interacting with the Delta of the first cell, which will promote the production of Delta creating a positive feedback loop where the different gene expression in the two neighboring cells will be amplified.

To understand this mathematically, let's define N_1 and N_2 the Notch activity in cell 1 and 2, and D_1 and D_2 the Delta activity in cell 1 and 2. Then, the dynamic of the system can be described by the following set of differential equations:

$$\dot{N}_1 = F(D_2) - \gamma_N N_1 \tag{4.5}$$
$$\dot{D}_1 = G(N_1) - \gamma_D D_1 \tag{4.6}$$
$$\dot{N}_2 = F(D_1) - \gamma_N N_2 \tag{4.7}$$
$$\dot{D}_2 = G(N_2) - \gamma_D D_2 \tag{4.8}$$

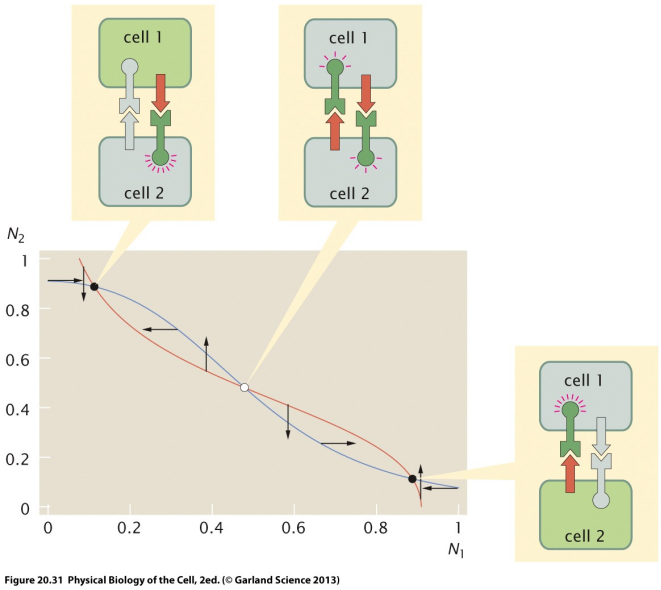


Fig. 4.13 Phase portrait of the Delta-Notch system

where γ_N and γ_D represent the degradation rate of Notch and Delta, respectively, $F(D_i)$ represents the activation of Notch by the Delta of the neighboring cell, and $G(N_i)$ represents the suppression of Delta by the Notch of the same cell.

We can change render these equations dimensionless by dividing all the equations by γ_N:

$$\dot{N}_1 = f(D_2) - N_1 \tag{4.9}$$
$$\dot{D}_1 = v[g(N_1) - D_1] \tag{4.10}$$
$$\dot{N}_2 = f(D_1) - N_2 \tag{4.11}$$
$$\dot{D}_2 = v[g(N_2) - D_2] \tag{4.12}$$

where $v = \gamma_D/\gamma_N$, $f = F/\gamma_N$ and $g = G/\gamma_D$.

Let's consider the limit when $v \gg 1$ (Delta decays very fast compared to Notch). This effectively means that we can separate time-scale so that the Delta time scale is much faster than that of Notch. Then one can approximate $D_1 \approx g(N_1)$ and $D_2 \approx g(N2)$ and get:

$$\dot{N}_1 = f(g(N_2)) - N1 \tag{4.13}$$
$$\dot{N}_2 = f(g(N_1)) - N2. \tag{4.14}$$

To understand the dynamic, let's draw a phage protrait (fig. 4.13). The nullclines, defined as $N1 = f(g(N_2))$ and $N2 = f(g(N_1))$ are shown in red and blue. The points at which the nullclines intersect are the fixed points where the system is at equilibrium. The nullclines divide the phage protrait in different areas where we can determine whether N_1 or N_2 will decrease or increase (arrows in fig. 4.13). This shows that the stable equilibria are in the two extreme cases, where the two cells have opposite behavior.

Physics of regulation: Reactions and Stat Mech of promoters

5

5.1 Modeling protein production with ODE

Ordinary differential equations can be used to describe chemical reactions inside the cell.[1] Molecular interactions between protein molecules, other small molecules, and DNA binding sites can turn on and off the activities of proteins and genes. These regulatory interactions combine into complex regulatory networks that ultimately control how cells behave. Here, we will use ordinary differential equations (ODEs) to describe how these regulatory networks work. ODEs provide a powerful tool for predicting how a regulatory network that is wired in a particular way will behave inside the cell. We will consider in this section two rather simple but very important examples (an unregulated gene and a negatively autoregulated gene), but the same methods are used to analyse much more complicated networks with many tens of genes and proteins.

The interior of cells is complicated: eukaryotic cells contain different cell compartments (e.g. the nucleus), and the contents of these compartments can also be organised in complicated ways. Prokaryotic cells, such as bacteria, don't have compartments but they are highly packed with proteins and DNA, and some proteins tend to occupy specific regions of the cell.

Although this spatial structure probably plays an important role in the ways in which cells function, we can understand many aspects of cell regulation without taking it into account. Here, we will make the important assumption that the interior of the cell (or a particular cellular compartment) is "well mixed" (this will not always be the case!)

[1] credit for lectures on ODE and noise modeling: Dr Rosalind Allen, Univ. Edinburgh, through IoP Biological Physics online teaching material

General intro to reaction ODE

Suppose that when an A molecule collides with a B molecule, the two can react to produce a molecule of type C. Starting from a mixture of A and B, we would like to know how many C molecules will have been produced at time t. We suppose that in a small

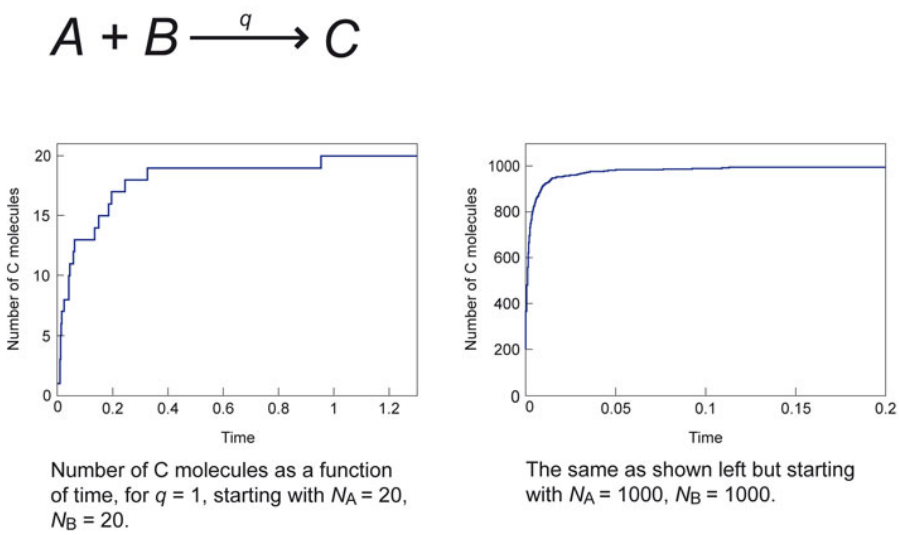


Fig. 5.1 Simulations of simple chemical reaction.

interval of time dt, the probability of a C molecule being produced is qN_AN_B/V, where V is the volume of the system, N_A is the number of A molecules and N_B is the number of B molecules (the probability scales with $1/V$ since a pair of A and B molecules will be less likely to meet each other in a larger volume). We can then write

$$\text{source} \xrightarrow{q} \text{P}$$

The constant q is called the rate constant.[2]

Figure 5.1 shows the output of a numerical simulation of the reaction

$$\text{A} + \text{B} \longrightarrow \text{C}$$

In these simulations, we have assumed that the volume, V, is set to 1. In the left-hand plot, we can see that the number of C molecules (N_C) increases over time, and that the C molecules are produced at random points in time, whenever an A and a B molecule happen to collide. This randomness can be important if there are only a small number of A and B molecules, and we will return to this later. The right-hand plot shows the same reaction, but with many more A and B molecules. In this case, many collisions happen in a small time interval and the plot for the number of C molecules versus time is much smoother. In fact, we can assume that the number of A, B and C molecules (per unit volume) change continuously with time. This is an important assumption because it allows us to write ODEs to describe how the system changes with time. The variables in these ODEs are the concentrations (number per unit volume) of the chemical species,

[2]The usual symbol for a rate constant is k, but there are several rate constants in this lecture, so we are using q for this one to avoid having several different constants all called k.

in this case A, B and C, which we denote c_A, c_B and c_C (e.g. $c_A = N_A/V$). For example, the set of ODEs that represents the reaction of A and B to produce C
$\mathrm{A} + \mathrm{B} \xrightarrow{\mathrm{q}} \mathrm{C}$
is

$$\begin{aligned} \frac{dc_A}{dt} &= \frac{dc_B}{dt} = -qc_a c_B \\ \frac{dc_C}{dt} &= qc_a c_B. \end{aligned}$$

It's important to note that because this is a second order or bimolecular reaction (it involves two reacting molecules), the dimensions of the rate constant are (concentration^{-1})(time^{-1}). We also need to specify initial conditions, e.g. $c_A(0) = c_B(0) = c_0$ and $c_C(0) = 0$.

Application to protein production

We can use the same ordinary differential equation methods to understand how cells control the production of protein molecules from their genes. Here, we are interested in how the concentration, c_{P}, of a specific protein molecule, P, changes with time inside the cell. Protein P is produced from its gene, gP, by transcription (to make messenger RNA) followed by translation (to make an amino-acid chain) and protein folding. We could model all of these processes in detail but for now let's just suppose that protein P is produced at a constant rate, k, as long as the gene, gP, is active. This reaction is zeroth order: the protein P is created at a constant rate that does not depend on any other variables in the model. The dimensions of the rate constant for this reaction are therefore (concentration)(time^{-1}).

We write this as a chemical reaction,
$\mathrm{source} \xrightarrow{\mathrm{k}} \mathrm{P}$
In this reaction, the "source" is actually the gene, gP, plus the whole machinery of transcription and translation. Here we just put this into a 'black box' and assume that protein P is produced at a constant rate.

Protein molecules are also removed from the cell; This could be because another protein molecule actively degrades them or because the cell is growing and dividing into daughter cells (and every time the cell divides, a given protein molecule has a chance of being lost). For now, let's just assume that there is a fixed probability per unit time, μ, that any given molecule of P is removed. We can also write this as a chemical reaction,
$\mathrm{P} \xrightarrow{\mu} \mathrm{sink}$
This is a first-order or unimolecular reaction: a single molecule of P reacts. For unimolecular reactions, the dimensions of the rate constant are (time)$^{-1}$. The "sink" here is another black box; P

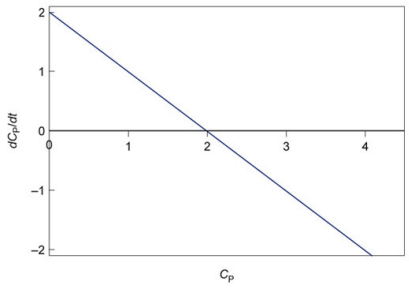


Fig. 5.2 Rate of change of protein concentration.

might have been removed into a daughter cell or it might have degraded into unspecified products.

Combining the constant rate of production, k, and the constant rate, per molecule, of loss, μ, we can write a differential equation for the rate of change of the concentration c_{P} of P molecules:

$$\frac{dc_P(t)}{dt} = k - \mu c_P(t). \tag{5.1}$$

We can tell a lot about the system without actually solving this equation. Figure 5.2 shows the rate of change, dc_{P}/dt, plotted for different concentrations of protein, c_{P}, for parameter values $k = 2$ and $\mu = 1$. When the concentration of protein is small ($c_{\mathrm{P}} < k/\mu$), the rate is positive. This means that there will be net protein production (c_{P} will increase). However, when the concentration of protein is large ($c_{\mathrm{P}} > k/\mu$), dc_{P}/dt is negative. This means there will be a net loss of protein. We can also see that for $c_{\mathrm{P}} = k/\mu$, dc_{P}/dt is zero. When the protein concentration reaches this value, there will be no net change: production balances removal. This is the steady-state protein concentration, $c_{\mathrm{P}}^{(ss)}$.

Steady-state concentrations are a very important property of regulatory networks, and quite often this is all that people focus on when they study a model for a particular regulatory network.

The value of $c_{\mathrm{P}}^{(ss)}$ depends on both k and μ. If protein removal is due to cell division and if the average time between cell divisions (the cell cycle time) is τ, then

$$\mu = \frac{\ln 2}{\tau}. \tag{5.2}$$

For the bacterium *E. coli* on a good food source, τ is about 30 min, so μ is about 0.02/min. Protein production rates, k ,vary greatly, from virtually zero to about 50/min. So the number of protein molecules in a cell (assuming that there is only one copy of the gene) can vary from zero to several thousand.

For the simple model discussed here, we also solve the model for the time-dependent protein concentration, $c_{\mathrm{P}}(t)$. This is important because genes can be turned on or off in response to signals, and we'd like to know how fast the cell can respond to a given signal. The time-dependent solution for protein concentration in this model can be found by simple integration,

$$c_{\mathrm{P}}(t) = \frac{k}{\mu}(1 - \mathrm{e}^{-\mu t}) + c_{\mathrm{P}}(0)\mathrm{e}^{-\mu t}. \tag{5.3}$$

To work out how fast the cell can respond to a signal, let's suppose that protein P is an enzyme that allows the cell to metabolise lactose. Initially, the gene, gP, is repressed because a repressor protein is bound to its promoter. We assume that initially no protein P is present: $c_{\mathrm{P}}(0) = 0$. At time zero, the cell detects some lactose and the repressor leaves the promoter, so the gene

becomes activated. How quickly can the cell produce protein P and start metabolising lactose? If $c_{\mathrm{P}}(0) = 0$, then the dynamics is given by

$$c_{\mathrm{P}}(t) \;=\; \frac{k}{\mu}(1 - \mathrm{e}^{-\mu t}). \tag{5.4}$$

We define the rise time, t_{rise}, as the time it takes for protein P to reach half of its steady-state value. Setting $c_{\mathrm{P}}(t)$ to $c_{\mathrm{P}}^{(ss)}/2$ and solving for t_{rise}, we obtain

$$t_{rise} \;=\; -\frac{1}{\mu} \ln \left[1 - \frac{\mu c_{\mathrm{P}}^{(ss)}}{2k} \right]. \tag{5.5}$$

which becomes, when we substitute in $c_{\mathrm{P}}^{(ss)} = k/\mu$,

$$t_{rise} \;=\; \ln(2)/\mu. \tag{5.6}$$

This result tells us that the response time of this simple network is determined only by the protein-removal rate. For bacteria, protein removal is usually due to cell growth and division. As we saw earlier, the removal rate, μ, is typically $\ln(2)/\tau$, where τ is the cell cycle time. So the response time for bacterial gene networks is typically of the order of the cell cycle time, which is at least 30 min.

ODE for negatively autoregulated gene

Genes can be turned on and off by the binding of specific proteins to the DNA in the promoter region. In many cases, proteins actually turn off their own production (i.e. the protein product of a gene is a repressor that binds to its own gene and turns off protein production). This is an example of negative feedback and is called negative autoregulation. It turns out that for *E.coli*, and probably for other organisms too, negative autoregulation happens much more often than one would expect if the regulatory "connections" between genes were chosen at random. Why has negative autoregulation been selected by evolution as a favoured regulatory motif? To try to understand this, let's write down the equivalent differential equation model for a protein that represses its own production. We recall that for a protein binding to a DNA binding site, the probability that the binding site is occupied is:

$$p_{bound} \;=\; \frac{\left(\frac{c}{c_0}\right) \exp -\beta\Delta\epsilon}{1 \;+\; \left(\frac{c}{c_0}\right) \exp -\beta\Delta\epsilon}, \tag{5.7}$$

where c/c_0 is the concentration of protein (relative to some standard value, c_0) and $\Delta\epsilon$ is the change in energy when the protein binds. We can define a dissociation constant, K_d, as

$$K_d \;=\; c_0 \mathrm{e}^{\beta\Delta\epsilon}. \tag{5.8}$$

For low concentrations (where c/c_0 is very small), we can see that the probability p_{bound} that the binding site is bound becomes proportional to the inverse of the dissociation constant: $p_{bound} \to c/K_d$. This shows us that K_d is actually just the equilibrium constant for the dissociation of the protein from its binding site. The reason why this proportionality does not hold at higher concentrations is that the binding site becomes saturated with protein.

The more strongly the protein binds to its DNA binding site, the more negative $\Delta\epsilon$ will be. Strong negative autoregulation (large negative $\Delta\epsilon$ therefore corresponds to a small value of K_d.

Combining the equations above, we get

$$p_{bound} = \frac{\left(\frac{c_P}{K_d}\right)}{1 + \left(\frac{c_P}{K_d}\right)}, \tag{5.9}$$

and the probability that the binding site is unoccupied is given by

$$p_{unbound} = 1 - p_{bound} = \frac{1}{1 + \left(\frac{c_P}{K_d}\right)}. \tag{5.10}$$

Returning to our differential equation for the production and degradation of protein, the production rate is now proportional to the probability that the promoter binding site is not occupied by protein:

$$\frac{dc_P(t)}{dt} = kp_{unbound} - \mu c_P = \frac{k}{1 + \left(\frac{c_P}{K_d}\right)} - \mu c_P. \tag{5.11}$$

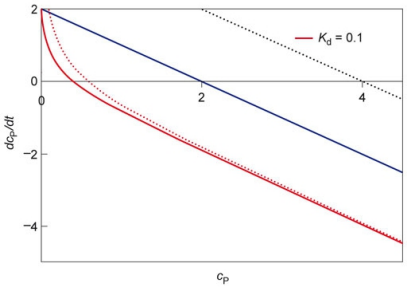


Fig. 5.3 Rate of change of protein concentration with negative autoregulation. The solid lines are for $k = 2$ and $m = 1$, and the dotted lines are for $k = 4$ and $m = 1$. The blue lines show the result without negative feedback (for the same k and m).

We now have a nonlinear differential equation for the concentration of protein, $c_P(t)$. Let's find out what the steady-state protein concentration is. Figure 5.3 shows a plot of the rate of change of c_P versus c_P, for two values of the production rate k. Also plotted are the results for a gene without negative autoregulation. We see that as in the non-regulated case, when the protein concentration c_P is low production dominates, while when the protein concentration is high protein degradation dominates over production. Again for one particular value of protein concentration production and degradation are balanced ($dc_P/dt = 0$), and this is the steady-state protein concentration.

We can see from Figure 5.3 that negative autoregulation affects the steady-state protein concentration in two important ways. First, the steady-state protein concentration is lower for the negatively autoregulated gene (shown in red) than for the unregulated gene (shown in blue). Second, when we compare the results for two different values of the production rate, k (solid and dotted lines), we can see that for the unregulated gene the steady-state protein concentration depends strongly on k (in fact, we know from our calculations above that it is proportional to k); while for

the negatively autoregulated gene, $c_{\mathrm{P}}^{(ss)}$ changes only a little when k is changed by a factor of two. Both of these effects have important implications for the performance of the gene, as we shall see.

To get an expression for the steady-state protein concentration $c_{\mathrm{P}}^{(ss)}$ for the negatively autoregulated gene, we set the rate of change of $c_{\mathrm{P}}(t)$ to zero:

$$\frac{dc_{\mathrm{P}}(t)}{dt} = \frac{k}{1 + \left(\frac{c_{\mathrm{P}}}{K_d}\right)} - \mu c_{\mathrm{P}} = 0, \tag{5.12}$$

obtaining

$$c_{\mathrm{P}}^{(ss)} = \frac{K_d}{2}\left[-1 + \sqrt{1 + \frac{k}{\mu K_d}}\right]. \tag{5.13}$$

For very strong autoregulation (where K_d is very small), the result reduces to:

$$c_{\mathrm{P}}^{(ss)} = \frac{K_d}{2}\left[-1 + 2\sqrt{\frac{k}{\mu K_d}}\right] \simeq \sqrt{\frac{kK_d}{\mu}}. \tag{5.14}$$

Figure 5.4 shows $c_{\mathrm{P}}^{(ss)}$ as a function of the protein production rate, k, for several values of the dissociation constant, K_d. As the negative autoregulation gets stronger (as K_d decreases), the curves become flatter: the steady-state protein concentration becomes less dependent on the protein production rate.

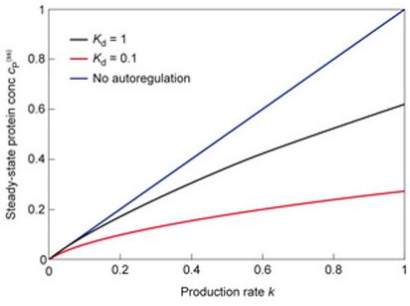


Fig. 5.4 Steady-state protein concentration for a negatively autoregulated gene.

In the cell, the protein production rate depends on the concentration of RNA polymerase, as well as the concentration of ribosomes, mRNA degradation enzymes, etc. All of these factors vary from cell to cell and over time inside any given cell. We therefore expect the protein production rate to fluctuate within and between cells. For a gene without negative autoregulation, this will cause the protein concentration to fluctuate, since $c_{\mathrm{P}}^{(ss)}$ is proportional to the production rate k. This **fluctuation problem can be avoided using negative autoregulation**. Because the curve of $c_{\mathrm{P}}^{(ss)}$ versus k is much flatter in the case of negative autoregulation, the steady-state protein concentration will remain stable even if the intracellular environment (i.e. the protein production rate) fluctuates. In other words, negative autoregulation can make the performance of a gene robust to changes in protein production rate.

You may have noticed that for negative autoregulation $c_{\mathrm{P}}^{(ss)}$ does depend on the dissociation constant, K_d. Is this a problem for robustness? Probably not: we expect K_d to fluctuate much less than k because K_d depends only on how strongly the protein binds to its DNA binding site, which is determined by the structure of the protein and the sequence of the binding site.

Negative autoregulation also has an important **effect on the rise time**, t_{rise}: the time the cell needs to turn the gene on (to the

half-maximal protein level). We saw that for the unregulated gene this time was fixed by the protein-removal rate, $t_{rise} = \ln(2)/\mu$. What happens for a negatively autoregulated gene? To calculate t_{rise}, in principle, we should solve the full version of eq. 5.12, but this is tricky analytically. If we look at early times, when c_P is small, we can approximate $c_P(t)/K_d < 1$ then

$$t_{rise} = -\frac{1}{\mu} \ln \left[1 - \frac{\mu c_P^{(ss)}}{2k} \right], \tag{5.15}$$

and if we also assume that autoregulation is strong we can substitute the previous result for $c_P^{(ss)}$, obtaining

$$t_{rise} = -\frac{1}{\mu} \ln \left[1 - \frac{\mu}{2k} \sqrt{\frac{kK_d}{\mu}} \right] = \frac{1}{\mu} \ln \left[\frac{2}{2 - \sqrt{kK_d\mu}} \right]. \tag{5.16}$$

This result is plotted in Figure 5.5: As K_d decreases (i.e. as the negative autoregulation becomes stronger), t_{rise} decreases. This important result shows that negative autoregulation can help cells to respond more rapidly to changes in their environmental conditions than they would be able to without regulation. The units chosen in Figure 5.5 are rather arbitrary. To get a feeling for some real numbers, we have already seen that a typical protein-removal rate μ in a bacterial cell would be 0.02/min, so the rise time for a typical protein without negative autoregulation would be $\ln(2)/\mu$ (~ 35 min). While protein production rates and protein-DNA dissociation constants can vary enormously, a realistic value for k might be 0.2 molecules/min per cell volume and K_d might be 0.02 molecules per cell volume (for a protein that binds very strongly to its DNA binding site). The value of t_{rise} for a negatively autoregulated gene, assuming these parameter values, would then be 12.7 min: almost a factor of three faster than the gene without negative autoregulation.

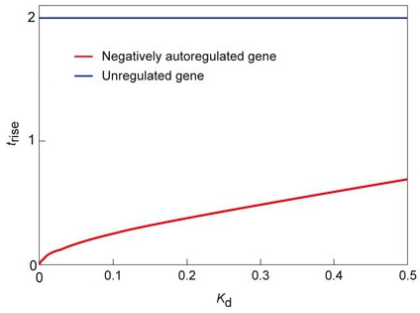


Fig. 5.5 Negatively autoregulation has strong effect on dynamics.

How are these measurements done, at population level?

Aside from noise and fluctuations, which we address below, how is the type of mRNA present in a sample (a population) of cells measured? DNA microarray chips can be used. These are large arrays (tens of thousands) of pixels (dots). Each pixel represents part of a gene, by having of the order of $10^6 - 10^9$ single-stranded DNA-mers, that are identical copies from the DNA of the gene. The chip size is of the order of 1 cm^2. The analysis consists of taking a cell sample, extracting all mRNA in this (hopefully) homogeneous sample, and translating it to cDNA (DNA that is complementary to the RNA, and thus identical to one of the strands on the original DNA). The cDNA is labeled with a fluorescent marker. The solution of many cDNAs is now flushed over the DNA chip, and the

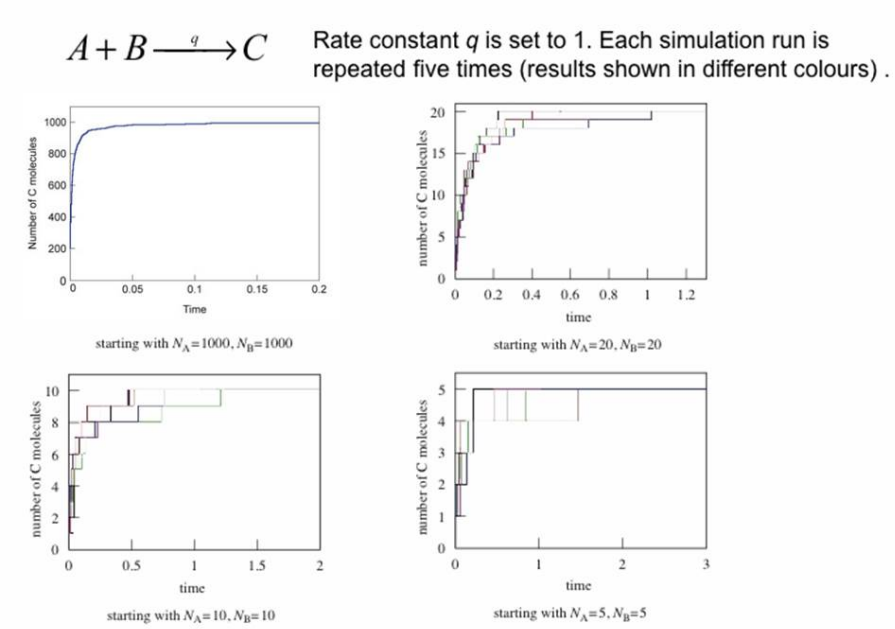


Fig. 5.6 Noise from low number of molecules can lead to different outcomes. Computer simulation results for reaction $A + B \to C$, starting with different numbers of A and B molecules.

cDNAs that are complementary to the attached single-stranded DNA-mers will bind to them. The DNA chip is washed and images (with pixel resolution) and the fluorescent light intensity thus measures the effective mRNA concentration. In its basic implementation, this technique gives one "snapshot" in time, and an average over many cells.

5.2 Biochemical noise

Cells with identical genes and environmental factors can differ chemically: we will see one way in which this can come about, using ideas about probability to model the processes mathematically.

Consider as before the reaction $A + B \to C$. Figure 5.6 shows how the number of C molecules increases in time, if we start with a 50:50 mixture of A and B. These results were obtained via computer simulations. Simulations were carried out, starting with 1000 molecules each of A and B, then with 20, 10 and 5 molecules each, with the rate constant, q, set numerically equal to 1 (to keep things simple). In each case, the simulation was repeated five times. When the total number of molecules is large, the number of C molecules rises smoothly and the repeat runs all give the same results. In this case, we can model the system with deterministic ordinary differential equations, as discussed in the previous

section.

However, if the total number of molecules is small, the system becomes very "noisy": the number of C molecules does not rise smoothly and repeat simulation runs give different results. Using standard methods from statistics, we can quantify what we mean by the number of molecules, N, being "small". It is convenient to define s as the ratio of the standard deviation in the mean to the standard deviation itself, $s = 1/\sqrt{N}$; this tends to unity for small N, and equivalently $N \simeq \sqrt{N}$. It turns out that "small molecule number" effects become important when the number of molecules becomes small enough that it is similar to its own square root.

Putting in the starting numbers of molecules for the simulations in Figure 5.6, when $N = 2000$, $s = 0.022$, but when $N = 5$, $s = 0.44$. Although Figure 5.6 shows computer simulation results, the same effect would happen in an experiment, if we could build an experimental system so small that it contained only a few molecules each of types A and B.

What is going on here? Why is our chemical reaction "noisy" when the number of molecules is small? The reason is that chemical reactions are stochastic, or random. That is, the outcome is governed by probabilities, and there are sufficiently few molecules that there is no single overwhelmingly favoured outcome. In our box of A and B molecules (the cell), we do not know the exact positions and velocities of all of the molecules and so we do not know the exact time when a pair of A and B molecules will meet and react. The exact times when reactions happen and the exact sequence of reactions that happen can be different in repeat runs of the same experiment. This may all be very interesting but why is it relevant? Even in something as small as a bacterial cell, there are many billions of molecules, so why would these stochastic effects be important? In fact, stochastic effects can be very important in cells, because even though the total number of molecules in a cell is large, the number of molecules involved in a particular biochemical reaction network can be very small. For example, in slow-growing cells, there is only one copy of the DNA (so the number of molecules of a particular gene may actually be only one). The number of messenger RNA molecules in the cell corresponding to a particular gene can also be very small for weakly expressed genes, and some proteins are only present in small numbers. Biochemical reaction networks involving genes, mRNA or proteins that are present in small numbers per cell are likely to be dramatically affected by small-molecule number fluctuations. We call these stochastic fluctuations "biochemical noise".

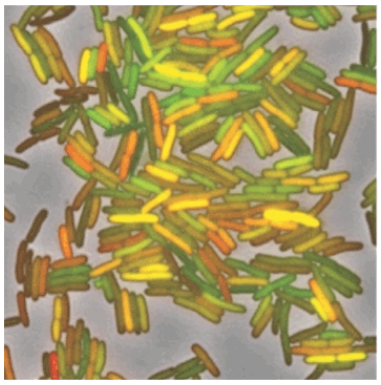


Fig. 5.7 Cells with identical genes in identical environments can behave differently. This can be explained in terms of biochemical noise. Cover image from Science Vol.297 issue 5584 (2002).

Individual cells are not identical

The fact that biochemical noise really is significant for biological cells was illustrated in an important experiment by Michael Elow-

itz *et al.* in 2002. They engineered Escherichia coli bacteria carrying two different-coloured fluorescent reporter genes. These genes encode proteins that do not interfere with any cellular functions but when excited by UV light of the right wavelength they fluoresce (i.e. they emit light of a longer wavelength). This can be detected in an epifluorescence microscope. Elowitz *et al.* were therefore able to measure the relative amounts of the two fluorescent proteins in individual bacterial cells. The question that they wanted to answer was: if two cells are genetically identical and experience the same environmental conditions, will they produce the same amount of the two fluorescent proteins?

Figure 5.7 shows the results of one of their experiments. This is an overlay of micrographs of a group of *E. coli* cells growing on a semi-solid gel under the microscope. These cells all grew from a single "ancestor" at the start of the experiment so they are genetically identical. The colours show the relative amounts of the two fluorescent proteins present in each cell: green represents protein 1 and red represents protein 2. Cells that are coloured yellow contain approximately equal amounts of proteins 1 and 2. It is clear from this image that these "identical" cells are different colours, showing that they are very far from identical in their levels of production of the fluorescent proteins. Elowitz *et al.* also showed that cells that produce the reporter proteins at low levels (small number of molecules) have much more "noisy" levels of expression than cells that produce the proteins at high levels (a large number of molecules). This is what we would expect if differences between cells are caused by small molecule number noise since $s = 1/\sqrt{N}$ is larger for small N.

Concept of intrinsic and extrinsic noise. Are the differences between cells shown in Figure 5.7 really caused by small molecule number noise in the chemical reactions involved in protein production (transcription and translation)? Or are the different colours caused by differences between the cells? For example, we can see in Figure 5.7 that some cells are short because they have just been generated, while others are much longer and are about to divide. Perhaps this affects the level of protein expression? Cells could also contain different concentrations of RNA polymerase or ribosomes, which would cause them to produce more or less fluorescent protein.

To explore the origins of the different amounts of the proteins, Elowitz *et al.* used two fluorescent proteins (in different colours) instead of just one. Within each cell, the genes encoding the two proteins should experience the same cell volume, RNA polymerase, ribosome concentration, etc. So, if the differences in protein expression are caused by differences between cells, the levels of the two colours should be correlated: cells with a lot of protein 1 should also have a lot of protein 2. However, if chemical reaction stochasticity is responsible for the differences in protein

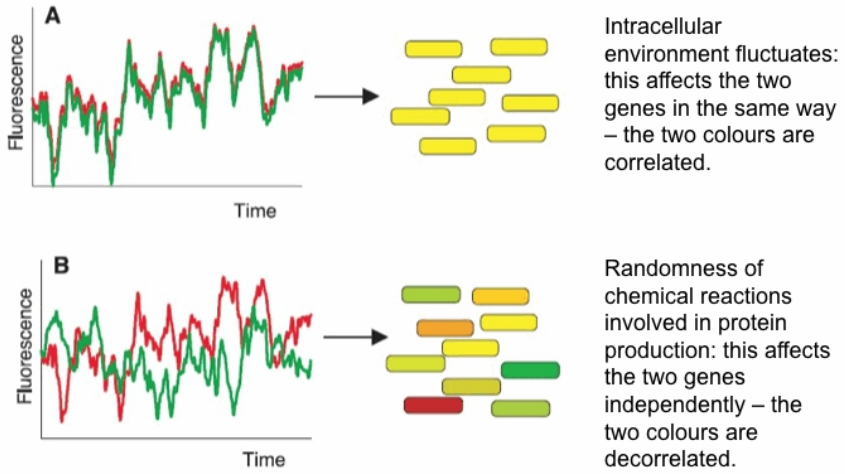


Fig. 5.8 Use of two "reporters" allows to distinguish intrinsic versus extrinsic noise. Protein levels vary because of fluctuations in the intracellular environment and of biochemical noise in transcription and translation.

expression, we would not expect the levels of protein 1 and protein 2 in individual cells to be correlated. This is illustrated in Figure 5.8.

In fact, by measuring the amount of correlation between the levels of proteins 1 and 2 in individual cells in their experiments, Elowitz *et al.* could measure how much of the cell-to-cell variation is caused by differences between cells (which they called extrinsic noise) and how much is caused by chemical reaction stochasticity (which they called intrinsic noise). In their experiments, both sources of noise played a significant role.

Why does it matter that genetically identical cells can have different levels of protein expression? One reason is that biochemical noise limits how precisely cells can control their own behaviour. If a cell needs to control precisely the concentration of a particular protein, either it must produce a large number of molecules (which is expensive) or it must use a biochemical control circuit (such as a negative feedback loop) to reduce the noise.

On a more positive note, biochemical noise may actually be useful for cells in some cases. For example, bacterial populations are often exposed to environmental stress (attack by antibiotics, changes in food availability, etc). If all of the cells in the population are identical in their protein composition, the stress may wipe them all out; but if there is large variability in protein composition among cells, it is possible that a few cells will happen to

have the right protein levels to survive the stress. The population can then regrow from these cells once the stress is over.

Theory of noise

For stochastic chemical reactions, we cannot predict exactly which reaction will happen when, or which cell in a population will contain which exact numbers of molecules of proteins, mRNA, etc. However, we can make predictions about probability distributions. For example, we might predict the probability that a randomly selected cell in a population will have 100 molecules of a particular protein, even though we cannot predict which cell this will be. The quantity we are interested in is therefore $p(N,t)$: the probability that our system contains N molecules of protein P at time t.

"Birth-death" model for gene expression. We can write down an equation for $p(N,t)$ for the simple "one-step model" of gene expression that we discussed above, in which we include chemical reactions for protein production and degradation:
source $\xrightarrow{\text{k}}$ P
P $\xrightarrow{\mu}$ sink
We assume that these reactions are "Poisson processes". This means that if we observe the system for a very short time interval from time t to time $t + dt$, the probability that the first reaction (production) happens will be

$$Prob(produce) \;=\; kdt,$$

while the probability that the second reaction (degradation) happens in this same time interval will be

$$Prob(degrade) \;=\; \mu N dt,$$

where N is the number of molecules of protein P, since the more P molecules there are, the more likely it is that this reaction will happen somewhere in the system during the time interval $t \;\rightarrow\; t + dt$.

How does the probability, $p(N)$, of having N molecules change during the time interval $t \;\rightarrow\; t + dt$? To determine this, we need to think about how the system can enter and leave the state of 'having N molecules'. To get N molecules, the system could have (a) previously had $(N-1)$ molecules and gained one more in a production reaction, or (b) previously had $(N+1)$ and lost one in a degradation reaction. These are the only ways in which the system can enter the 'state of having N molecules'. However, it can also leave this state if it already has N molecules and either (a) another one is produced (then it will have $N+1$), or (b) one is degraded (then it will have $N-1$), see Figure 5.9.

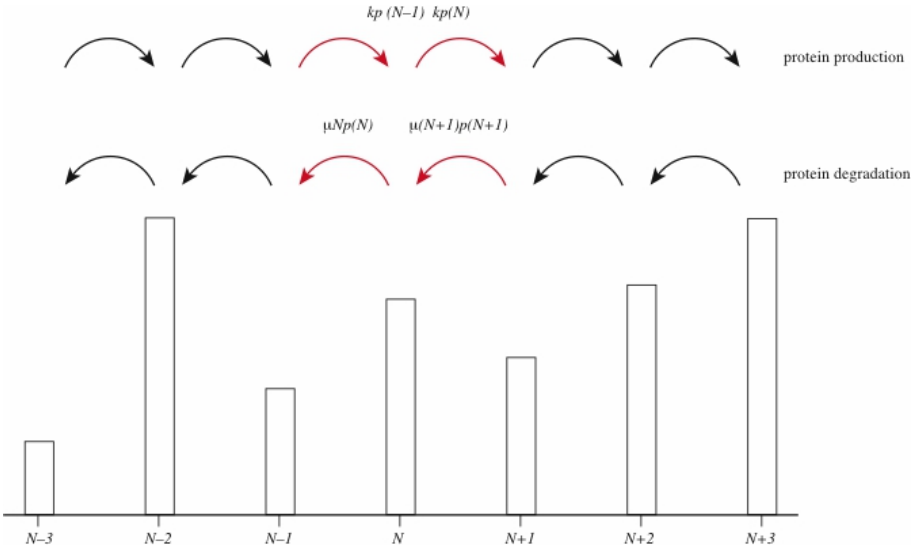


Fig. 5.9 Considering individual steps in a chemical reaction. Here, the vertical bars represent the probability of having a particular number of molecules and the arrows represent how the number of molecules is changed by the protein production and degradation reactions. In a very small time interval, $t \rightarrow t + dt$, the probability $p(N, t)$ increases due to the possibility of reactions happening from states $(N-1)$ or $(N+1)$ to N, and it decreases due to the possibility of reactions from state N to $(N-1)$ or $(N+1)$.

By summing all of the probabilities we can generate an equation called the chemical master equation:

$$\frac{dp(N,t)}{dt} = kp(N-1) + \mu(N+1)p(N+1) - kp(N) - \mu N p(N) \tag{5.17}$$

Let us suppose that we are only interested in the probability distribution $p(N)$ after a long time, once the system has reached its steady state. In that case, we have

$$\frac{dp(N,t)}{dt} = 0. \tag{5.18}$$

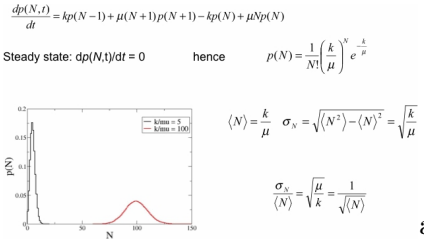


Fig. 5.10 Solution of chemical master equation for the simple one-step model of protein expression.

Solution to this is:

$$p(N) = \frac{1}{N!}\left(\frac{k}{\mu}\right)^N e^{-\frac{k}{\mu}}. \tag{5.19}$$

as you can check by substitution, noting that $p(N-1) = N(\mu/k)p(N)$ and that $p(N+1) = (k/\mu)(1/(N+1))p(N)$.
Equation 5.19 is the well known Poisson distribution.
Figure 5.10 shows the probability distribution $p(N)$ plotted for different values of (k/μ). We can see that as (k/μ) increases, the average number of molecules increases. The mean and standard

deviation σ_N of the distribution $p(N)$ are given by:

$$\begin{aligned} \langle N \rangle &= \frac{k}{\mu} \\ \sigma_N &= \sqrt{\langle N^2 \rangle - \langle N \rangle^2} = \sqrt{\frac{k}{\mu}}, \end{aligned}$$

We can estimate the importance of stochastic effects looking at the ratio of the standard deviation to the mean:

$$\frac{\sigma_N}{\langle N \rangle} = \sqrt{\frac{\mu}{k}} = \frac{1}{\sqrt{\langle N \rangle}}, \tag{5.20}$$

this explains why earlier we stated that small molecule number noise becomes important when the inverse square root of the number of molecules is close to one.

A two-step model for protein production

The model that we have just been considering may be too simple. In reality, the production of protein from a gene does not happen in a single step. We can make our model slightly more realistic by making a two-step model that includes both transcription and translation. The reaction scheme for this model would be

source $\longrightarrow$ M

M $\longrightarrow$ sink

M $\longrightarrow$ M + P

P $\xrightarrow{\mu}$ sink

Here, M represents mRNA and P represents protein. It is possible to write down a chemical master equation also for this model, and to solve it for the steady state probability distribution. In this case, there is a probability distribution for the number of messenger RNA molecules as well as for the number of protein molecules. For mRNA we only need to consider the top two reactions (since the bottom two reactions do not change the number of mRNA molecules), which are identical to our previous simpler model. So we expect the probability distribution for the number of mRNA molecules to be a Poisson distribution. However the bottom two reactions, which control the production and degradation of protein, are now different from our simple model. This means that the probability distribution of protein may be different from a Poisson distribution in this model.

Figure 5.11 shows the protein number probability distribution for this model. We set the parameters (translation rate/mRNA decay rate) so that five proteins are made on average per mRNA molecule (although some mRNA molecules will produce more and some less). We can compare this with the previous one-step model by fixing the transcription rate so that the average protein number is the same in both models. The results are shown in Figure 5.11: we can see immediately that the distribution is broader

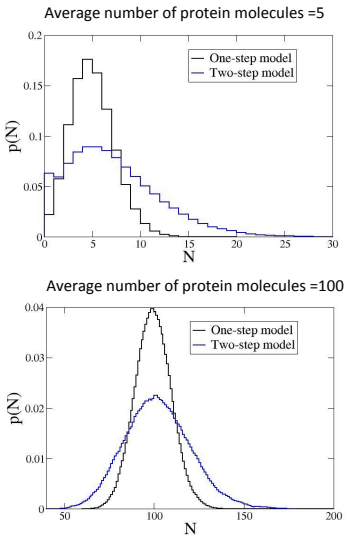


Fig. 5.11 Chemical master equation solutions for the one- and two-step models of protein expression. For the two-step process we assume that on average an mRNA produces five proteins, and we fix the transcription rate to get the same average number of proteins as in the one-step model.

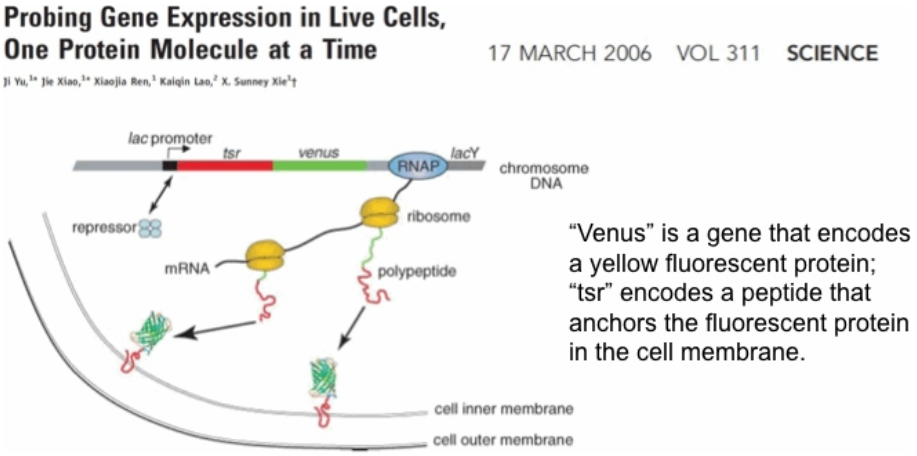


Fig. 5.12 Direct imaging of noise in gene expression. This experimental system was constructed by Yu *et al.* in 2006 to visualise in real time the production of a single protein molecule in a cell. From Yu *et al.* 2006, 'Probing gene expression in live cells, one protein molecule at a time', Science 311 1600.

in the two-step model. This model predicts more noisy protein expression than the one-step model. The reason for this is that the extra chemical reaction step amplifies the noise: the number of mRNA molecules is itself noisy, and then on top of this each mRNA molecule can produce a variable number of proteins.

Visualising noise in gene expression

How can we test whether these are good models for noisy gene expression in real cells? One way to do this is actually to carry out single molecule experiments, in other words to watch, under the microscope, the production of single protein molecules in individual cells. Since protein molecules are very small, this is a very challenging task. However, in 2006, Yu *et al.* managed to design an appropriate experiment (Figure 5.12). They made a strain of *E. coli* that produced a yellow fluorescent protein attached to a polypeptide (a chain of amino acid molecules), which could anchor this complex in the cell's lipid membrane. When the fluorescent protein is anchored in the membrane, it diffuses around much less, making it easier to see single molecules under the microscope. In this system, using advanced fluorescent microscopy, it is possible to see individual fluorescent protein molecules as dots within the cell membrane. Yu *et al.* could then grow cells under the microscope and track the moments when individual dots appeared in the membrane. In this way, they could see the production of individual protein molecules in real time. To keep the protein numbers low, the researchers included a binding site for the Lac repressor protein (see Section 5.3) When this repressor protein is bound to the operator site in front of the gene that encodes the fluorescent protein, no protein will be produced.

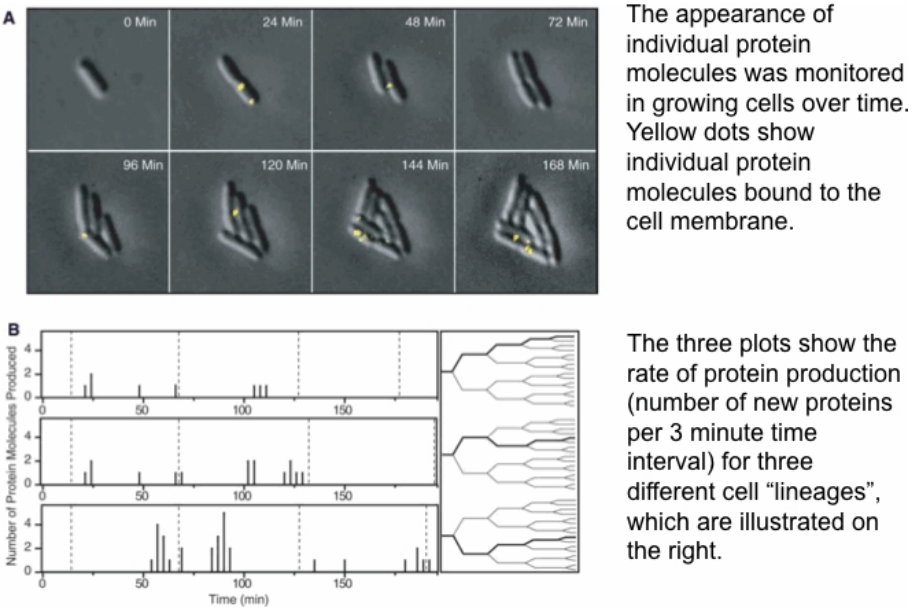


Fig. 5.13 Experiments can identify production of individual proteins. Some of Yu *et al.*'s results, showing the moment when individual protein molecules are produced in growing bacterial cells. From Yu *et al.* 2006.

Figure 5.13 shows some of Yu *et al.*'s results. The bacterial cells in the series of images grow from a single cell during the experiment. The yellow dots show individual protein molecules bound to the cell membrane. By tracking the appearance of these dots, Yu *et al.* were able to monitor the moments when protein molecules appeared in the membrane. This was done for different cell lineages, as shown in the plot, which indicates the number of protein molecules that were produced in a 3 min interval. The dotted vertical lines show the moments when the cell divided into two daughter cells.

What's really striking about Yu *et al.*'s results is that for *most of the time, no protein molecules are being produced.* Protein production occurs in short bursts, with long intervals where nothing happens. This is probably because most of the time the Lac repressor protein is bound to the DNA, thereby preventing protein expression. The bursts of expression take place during the rare moments when a stochastic fluctuation causes the repressor to fall off its DNA binding site. Yu *et al.*'s setup therefore allows us to see stochastic chemical reactions happening inside biological cells, in real time and at single-molecule resolution.

We have focused here on noise in gene expression, but the stochasticity of chemical reactions is also important in many other cell functions. Single-molecule experiments have revealed the effects of biochemical noise in the molecular machines that drive the flagellar motor that allows cells to swim and in the bacterial membrane receptors that sense environmental gradients. Other

experiments have found important effects of biochemical noise in the development of fruit-fly embryos and the mechanisms that control whether or not cells proliferate. It seems that noise is everywhere.

5.3 From a molecular to a stat mech description of regulation

We develop here a physics-based view of how gene expression is regulated, following closely the text of (Phillips et al., 2013).

RNA polymerase binding to a specific site

Following from page 242 (Phillips et al., 2013).
L ligands. Prob that 1 ligand is bound to receptor:

$$\text{weight when receptor occupied} = \mathrm{e}^{-\beta\epsilon_b} \times \sum_{solution} \mathrm{e}^{-\beta(L-1)\epsilon_{sol}},$$

where the summation is the sum over all ways of arranging the $L-1$ ligands in solution. Imagine Ω 'lattice sites' in solution. Then

$$\sum_{solution} \mathrm{e}^{-\beta(L-1)\epsilon_{sol}} = \frac{\Omega!}{(L-1)![\Omega-(L-1)]!}.$$

The partition function is

$$Z(L,\Omega) = \sum_{solution} \mathrm{e}^{-\beta L\epsilon_{sol}} + \mathrm{e}^{-\beta\epsilon_b \sum_{solution}} \mathrm{e}^{-\beta(L-1)\epsilon_{sol}}.$$

The sum in the second term has already been evaluated. The first term is

$$\sum_{solution} \mathrm{e}^{-\beta L\epsilon_{sol}} = \mathrm{e}^{-\beta L\epsilon_{sol}} \frac{\Omega!}{L!(\Omega-L)!}.$$

Bringing both together,

$$Z(L,\Omega) = \mathrm{e}^{-\beta L\epsilon_{sol}} \frac{\Omega!}{L!(\Omega-L)!} + \mathrm{e}^{-\beta\epsilon_b}\mathrm{e}^{-\beta(L-1)\epsilon_{sol}} \frac{\Omega!}{(L-1)![\Omega-(L-1)]!}.$$

If we simplify considering

$$\frac{\Omega!}{L!(\Omega-L)!} \simeq \frac{\Omega^L}{L!},$$

which is ok provided $\Omega >> L$, then we can write the probability of being bound as:

$$p_{bound} = \frac{\mathrm{e}^{-\beta\epsilon_b} \frac{\Omega^{L-1}}{(L-1)!}\mathrm{e}^{-\beta(L-1)\epsilon_{sol}}}{\frac{\Omega^L}{L!}\mathrm{e}^{-\beta L\epsilon_{sol}} + \mathrm{e}^{-\beta\epsilon_b}\frac{\Omega^{L-1}}{(L-1)!}\mathrm{e}^{-\beta(L-1)\epsilon_{sol}}}.$$

Now defining $\Delta\epsilon = \epsilon_b - \epsilon_{sol}$, we can simplify to:

$$p_{bound} = \frac{(L/\Omega)\mathrm{e}^{-\beta\Delta\epsilon}}{1 + (L/\Omega)\mathrm{e}^{-\beta\Delta\epsilon}},$$

which we can write in terms of a concentration c:

$$p_{bound} = \frac{(c/c_0)\mathrm{e}^{-\beta\Delta\epsilon}}{1 + (c/c_0)\mathrm{e}^{-\beta\Delta\epsilon}},$$

where c_0 is a reference state of full occupation. For example, if we assume our molecules to be of volume $1\,\mathrm{nm}^3$, then $c_0 = 0.6\mathrm{M}$

This, obtained here in the language of ligand/receptor binding, is a classical result known as 'Hill function', or also as a 'Langmuir adsorption isotherm' from the Part II Stat Phys course.

RNA polymerase binding: competition between specific and non specific site

This extends the calculation above. Let's assume the non-specific sites on the DNA are N_{NS} 'boxes'. Then the partition function associated with these states is:

$$Z_{NS}(P, N_{NS}) = \frac{N_{NS}!}{P!(N_{NS} - P)!} \times \mathrm{e}^{-\beta P \epsilon_{pd}^{NS}},$$

where ϵ_{pd}^{NS} is the energy of binding the polymerase to a non-specific site (and ϵ_{pd}^{S} will be later the energy of binding the polymerase to the specific site).

Now we can write the total partition function. We need to sum over the states in which the promoter is occupied (hence P-1 polymerase molecules in the non-specific sites), and those where it is not:

$$Z(P, N_{NS}) = Z_{NS}(P, N_{NS}) + Z_{NS}(P-1, N_{NS})\mathrm{e}^{-\beta\epsilon_{pd}^{S}}.$$

Hence the ratio of configuration weights where promoter is bound, to all weights, is:

$$p_{bound} = \frac{\frac{N_{NS}!}{(P-1)![N_{NS}-(P-1)]!}\mathrm{e}^{-\beta\epsilon_{pd}^{S}}\mathrm{e}^{-\beta(P-1)\epsilon_{pd}^{NS}}}{\frac{N_{NS}!}{P!(N_{NS}-P)!}\mathrm{e}^{-\beta P\epsilon_{pd}^{NS}} + \frac{N_{NS}!}{(P-1)![N_{NS}-(P-1)]!}\mathrm{e}^{-\beta\epsilon_{pd}^{S}}\mathrm{e}^{-\beta(P-1)\epsilon_{pd}^{NS}}}$$

As in the previous subsection, the factorials can be simplified, and we can write the result to show that only the energy difference matters:

$$p_{bound} = \frac{1}{1 + \frac{N_{NS}}{P}\mathrm{e}^{\beta\Delta\epsilon_{pd}}},$$

this is the familiar result for two-state models, with the unoccupied state of the promoter having weight =1, and the occupied having weight $P/N_{NS}\mathrm{e}^{-\beta\Delta\epsilon_{pd}}$.
The energy differences $\Delta\epsilon_{pd}$ are negative, and can range between minus a few to $\sim -10\,k_BT$.

Activation and repression of promoter regions

Now that the 'combinatorics' is fresh from above, we can make another construction along this line, and tackle the more complex cases of promoter regulation by transcription factors.

Activators. Activators are proteins that bind to a specific site, and promote the recruitment of RNA polymerase to a nearby promoter site. We now have 4 classes of outcome to sum over to make the total partition function: the activator and promoter site can each be occupied or unoccupied. So:

$$\begin{aligned} Z_{tot}(P, A, N_{NS}) &= Z(P, A, N_{NS}) \text{ (empty)} \\ &+ Z(P-1, A, N_{NS})\mathrm{e}^{-\beta\epsilon^S_{pd}} \text{ (only RNAP on promoter)} \\ &+ Z(P, A-1, N_{NS})\mathrm{e}^{-\beta\epsilon^S_{ad}} \text{ (only activator bound)} \\ &+ Z(P-1, A-1, N_{NS})\mathrm{e}^{-\beta(\epsilon^S_{pd}+\epsilon^S_{ad}+\epsilon_{pa})}. \text{ (both RNAP and activator bound)} \end{aligned}$$

(Here A, a are the activator, P, p the polymerase, d the DNA). ϵ_{ap} is the energy that favors the activator and the RNA polymerase being close.

The algebra is more lengthy but follows the exact steps as previously. To get promoter occupancy, we can take the ratios of the weights of the two 'favorable' states, against the sum of all weights, and we get:

$$p_{bound}(P, A, N_{NS}) = \frac{1}{1 + \frac{N_{NS}}{P\, F_{reg}(A)}\mathrm{e}^{\beta\Delta\epsilon_{pd}}},$$

where the function $F_{reg}(A)$ is:

$$F_{reg}(A) = \frac{1 + (A/N_{NS})\mathrm{e}^{-\beta\Delta\epsilon_{ad}}\mathrm{e}^{-\beta\epsilon_{ap}}}{1 + (A/N_{NS})\mathrm{e}^{-\beta\Delta\epsilon_{ad}}},$$

and the $\Delta\epsilon$ are the energy differences between specifically and non-specifically bound conditions.

This is a neat result, because it shows that activating molecules make an $F > 1$, i.e. have an effect that is mathematically equivalent to increasing the number of polymerases. Given realistic values of the other energies, a few $-k_BT$ for ϵ_{ap} is enough to significantly change the bound probability, see (and reproduce your own?) Figs.19.10 and 19.11 in (Phillips et al., 2013).

If the approx $(N_{NS}/PF_{reg})\mathrm{e}^{\beta\Delta\epsilon_{pd}} >> 1$ holds, i.e. the promoter is not too strong, then you can obtain (exercise) that the fold increase is approximately $F_{reg}(A)$ itself.

Repressors. Repressor proteins occupy the promoter region, and prevent the PRNA binding there. The statistical mechanics

approach is a variant of the above. The partition function associated with binding of repressors to the non-specific sites is:

$$Z(P,R,N_{NS}) \;=\; \frac{N_{NS}!}{P!R!(N_{NS}-P-R)!}\mathrm{e}^{\beta P\epsilon_{pd}^{NS}}\,\mathrm{e}^{\beta R\epsilon_{rd}^{NS}}.$$

Now the total partition function is:

$$\begin{aligned} Z_{tot}(P,R,N_{NS}) \;&=\; Z(P,R,N_{NS}) \;\text{(empty promoter)}\\ &+ Z(P-1,R,N_{NS})\mathrm{e}^{-\beta\epsilon_{pd}^{S}} \;\text{(RNAP on promoter)}\\ &+ Z(P,R-1,N_{NS})\mathrm{e}^{-\beta\epsilon_{rd}^{S}}. \;\text{(repressor on promoter)}\end{aligned}$$

With the same algebra steps and approximations as previously, we obtain

$$p_{bound}(P,R,N_{NS}) \;=\; \frac{1}{1\;+\;\frac{N_{NS}}{P}\mathrm{e}^{\beta(\epsilon_{pd}^{S}-\epsilon_{pd}^{NS})}[1+\frac{R}{N_{NS}}\mathrm{e}^{-\beta(\epsilon_{rd}^{S}-\epsilon_{rd}^{NS})}]}.$$

To obtain a compact expression of the same form as for activators, a regulating function $F_{reg}(A)$ can be defined as:

$$F_{reg}(R) \;=\; \left(1\;+\;\frac{R}{N_{NS}}\mathrm{e}^{-\beta\Delta\epsilon_{rd}}\right)^{-1},$$

with $\Delta\epsilon_{rd} = \epsilon_{rd}^{S} - \epsilon_{rd}^{NS}$. Here, $F_{reg} < 1$, which means that the systems behaves as if fewer polymerases were present.

Towards the real case: activation and repression! In a real regulatory system, both mechanisms can interplay. Again we can build on the same lines as before, and there are now six distinct possible outcomes:

$$\begin{aligned} Z_{tot}(P,A,R,N_{NS}) \;&=\; Z(P,A,R,N_{NS}) \;\text{(empty promoter)}\\ +\quad & Z(P-1,A,R,N_{NS})\mathrm{e}^{-\beta\epsilon_{pd}^{S}} \;\text{(RNAP on promoter)}\\ +\quad & Z(P,A-1,R,N_{NS})\mathrm{e}^{-\beta\epsilon_{ad}^{S}} \;\text{(activator on promoter)}\\ +\quad & Z(P-1,A-1,R,N_{NS})\mathrm{e}^{-\beta(\epsilon_{ad}^{S}+\epsilon_{pd}^{S}+\epsilon_{pa})} \;\text{(RNAP and activator on)}\\ +\quad & Z(P,A,R-1,N_{NS})\mathrm{e}^{-\beta\epsilon_{rd}^{S}} \;\text{(repressor on promoter)}\\ +\quad & Z(P,A-1,R-1,N_{NS})\mathrm{e}^{-\beta(\epsilon_{ad}^{S}+\epsilon_{rd}^{S})}. \;\text{(activator and repressor on)}\end{aligned}$$

As before the RNA polymerase binding probability can be calculated and has the form:

$$p_{bound}(P,A,R,N_{NS}) \;=\; \frac{1}{1\;+\;\frac{N_{NS}}{P\,F_{reg}(A,R)}\mathrm{e}^{\beta(\epsilon_{pd}^{S}-\epsilon_{pd}^{NS})}},$$

where the regulating function $F_{reg}(A,R)$ is richer:

$$\begin{aligned} F_{reg}(A,R) \quad = \quad & \Big[1+(A/N_{NS})\mathrm{e}^{-\beta(\Delta\epsilon_{ad}+\epsilon_{ap})}\Big] \;/\\ & \Big[1\;+\;(A/N_{NS})\mathrm{e}^{-\beta\Delta\epsilon_{ad}}+(R/N_{NS})\mathrm{e}^{-\beta\Delta\epsilon_{rd}}+\\ & (A/N_{NS})(R/N_{NS})\mathrm{e}^{-\beta(\Delta\epsilon_{ad}+\Delta\epsilon_{pd})}\Big].\end{aligned}$$

The *lac* Operon

The *lac* Operon has played a key role historically in understanding physical and biological aspects of gene regulation. In the *lac* Operon there is an activator, the protein CAP: in order to recruit RNAP, CAP has to be bound to a molecule called cyclic AMP (cAMP), whose concentration goes up when amount of glucose decreases. There is also a repressor, the Lac repressor, which decreases the amount of transcription unless it is abound to allolactose, a byproduct of lactose metabolism.
Keep in mind that this regulation is just to ensure that the enzymes to digest lactose are produced only when glucose is not present, and lactose is present. It seems an apparent simple objective, but selecting reliably for one of four situations requires a mechanism of both activation and repression as outlined here.

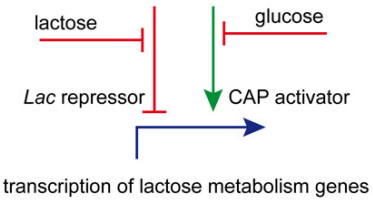


Fig. 5.14 Idealised (logic) lac network. A convenient way to illustrate the molecular interactions that make up the lac regulatory network. Here, positive molecular interactions (activation) are shown by arrows and negative molecular interactions (repression) are shown by "blocker" bars. The input to the network are the concentrations of lactose and glucose, the output is the activation of gene transcription for the machinery required to metabolise lactose. This type of diagram is often used to represent regulatory networks and is convenient when the networks are complicated, involving a lot of interactions.

Our thermodynamical model is in fact still too simple to describe quantitatively the *lac* Operon. There is another important detail which is worth mentioning, because it brings in the nature of the DNA double-helix as a polymer, with all the 'polymer physics' concepts that have been studied in other contexts. What we have not considered in the models above is the fact that (a) each lac repressor molecule has two binding sites, combined with (b) that there are three operator regions on the DNA for lac to bind (with slightly different binding energies, and situated 92 and 401bp on each side of the main operator). This fact corresponds to the possibility for the lac repressor to form a loop of DNA. Bending of double-stranded DNA carries of course a free energy cost, and this cost can in principle be regulated by the cell through associations of proteins, or physical chemistry changes, that lead to changes in DNA persistence length.

On one hand this *lac* Operon, is a classic system of study, and considered understood well enough to be used in an 'engineering' building-block spirit in synthetic biology constructions. On the other hand, it is still the study of refined experiments and models, aiming to understand it fully quantitatively. That systems open up new refined questions as we understand more of them is a familiar theme in various areas of physics.

Regulating gene expression by DNA conformation, with loops or compact regions stabilised by protein adhesion, is a very general mechanism heavily exploited in eukaryotic cells.

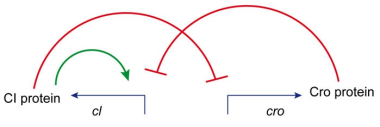


Fig. 5.15 Idealised (logic) and simplified diagram of the phage lambda genetic switch. The *cro* gene results in cell lysis; the *cI* gene promotes lysogeny.

Case study: lambda phage

This is another very well studied "hydrogen atom" situation in biology. A virus called bacteriophage lambda infects *E. coli*. Once a bacterial cell is infected, the virus has two options: it can either hijack the cell machinery to replicate itself and then kill the cell (known as lysis), resulting in its release, or it can add its DNA

to the DNA sequence of the bacterium and lie dormant inside the host cell (known as lysogeny) until conditions are more favourable for lysis. Which of these developmental pathways is adopted is determined by (a more complex version of) the regulatory network shown in Figure 5.15. This network contains two genes, *cI* and *cro*. When the *cro* gene is activated, cell lysis results; when the *cI* gene is activated, lysogeny follows. What prevents both pathways from being activated simultaneously?

As shown in Figure 5.15, the *cI* gene encodes a protein, CI, which acts as a repressor of the *cro* gene and an activator of its own gene. Thus, when *cI* is active, *cro* is repressed and remains inactive, while *cI* remains active. Likewise, the *cro* gene encodes a protein, Cro, which acts as a repressor of the *cI* gene. Thus, when the *cro* pathway to lysis has been adopted, the *cI* pathway to lysogeny is automatically shut down. In this way, the virus ensures that a binary all-or-nothing "decision" is made between lysis and lysogeny. This is an example of a bistable switch: a regulatory network with two distinct outcomes. Bistable switches are important not just for bacteriophage lambda but also in developmental and cell-fate decisions in many other cells, including human ones. (Bistable switches are also used in electronic control networks, where they maintain a circuit in one of two stable states until some external trigger is applied - very similar to their biological analogue.)

5.4 Simulating chemical reaction dynamics

Gene expression in a living organism is not a steady state process: at the embryo development level, regulation evolves within a cell cycle, and very significantly at cell division; the cell cycle also defines genes that are only expressed at certain times; 'zooming in' at even shorter times, the gene expression can often be seen to be happening in bursts. A dynamical description of concentrations can be important. Careful experiments, and models, can highlight the various sources of 'noise' (stochasticity) in expression, which can be quite different in origin: for example from the molecular binding event, to fluctuations in concentrations, to noise that comes from the coupling of the dynamical process.

Other processes in the cell for example the translation of proteins, or reaction networks of proteins, also can exhibit transients in time, and noise. How can we model this? Except in the simplest cases, there is not much that can be one analytically. Given a set of coupled differential equations, one can solve numerically. In a brute-force approach, a constant timestep for integration could be chosen: this would have to be much smaller than any reaction or decay timescales, and can be very wasteful of simulation time. A very elegant way to address these problems computationally was

proposed by Gillespie in 1977, and his algorithm is still in current use.

the Gillespie algorithm

The Gillespie algorithm instead of working with a constant Δt provides a strategy for adapting the timestep to the problem, by choosing it at random from a particular probability distribution. A second (biased) random number then determines which of the reactions take place at the simulation step. Running this algorithm is equivalent to following one particular realisation of the stochastic dynamics of a system. It is powerful because it has 'real time', and because by running it with several iterations one can build up distributions.

Let's see how the algorithm works (what is the correct probability distribution for Δt, and how to choose the reaction) with the example of the unregulated promoter. There are two reactions:
(1) an mRNA can be can be produced, with probability k per unit time;
(2) an mRNA can decay, with probability γ per unit time and per unit molecule.
Let's call $m(t)$ the number of mRNA molecules at time t.
Once we have a timestep Δt , we want to determine $P(i, \Delta t)dt$, the probability that reaction i takes place in the interval $\Delta t, \Delta t + dt$. First, we note that we also want to impose no reaction to have occurred before Δt. We call this probability $P_0(\Delta t)$. Thus the probability that reaction i takes place in the interval $\Delta t, \Delta t + dt$ is

$$P(i, \Delta t)dt \; = \; P_0(\Delta t)k_i dt.$$

How do we calculate $P_0(\Delta t)$? We can write

$$P_0(\Delta t + dt) \; = \; P_0(\Delta t) \left(1 - \sum_i k_i dt \right),$$

i.e. the product of the probability of no reaction having occurred up to Δt, time the probability of no reaction taking place in dt. The first term can be Taylor expanded around Δt, and we obtain

$$\frac{dP_0(\Delta t)}{d\Delta t} \; = \; -P_0(\Delta t) \sum_i k_i,$$

which has solution

$$P_0(\Delta t) \; = \; \mathrm{e}^{-\sum_i k_i \Delta t} \; = \; \mathrm{e}^{-k_0 \Delta t},$$

where we have used $P_0(\Delta t = 0) = 1$ and defined $k_0 = \sum_i k_i$.
Substituting back, we get

$$P(i, \Delta t)dt \; = \; \mathrm{e}^{-k_0 \Delta t} k_i dt.$$

If we sum this over all i, we get the probability that any of the possible reactions happens in the interval $\Delta t, \Delta t + dt$:

$$P(\Delta t)dt \;=\; \mathrm{e}^{-k_0 \Delta t} k_0 dt.$$

This is the distribution from which one needs to pick Δt.

Now we need to work out how to make a distribution from which to pick the random choice of which reaction takes place. The probability that reaction i happens at *some* time is:

$$P(i) \;=\; \int_0^\infty P(i, \Delta t) d(\Delta t) \;=\; \frac{k_i}{k_0}.$$

This tells us that the probability of a reaction to take place is just the ratio of its rate, and the sum of all the possible rates. This gives us the criterion to choose (randomly, but with the right bias) which reaction will take place at the simulation timestep.

In algorithm form, the steps in this example are:
1. given $m(t)$, calculate the rates. In this case only k_2 depends on $m(t)$.
2. draw a uniform random number x between $[0, 1]$. Compute k_0. $\Delta t = (1/k_0) \ln(1/x)$. This last formula is a way (you can check) to turn the uniform random number in a random number from the exponential distribution we want, calculated above. Advance simulation clock by Δt.
3. draw a uniform random number between $[0, 1]$. If the number is between $[0, k_1/k_0[$, increase the mRNA molecule number by one. If it is between $[k_1/k_0, 1]$ then decrease the mRNA molecule number by one.
4. loop back to step (1).

Check that the distribution of m at steady state is well described by a Poisson distribution. This is a result that could have been obtained analytically, in this simple example.

Dynamical Systems: Systems and Circuits

6

Recall the introduction to one and two-dimensional dynamical systems, from previously in the course. Here we investigate, following section 19.3.5 of (Phillips et al., 2013), some interesting examples of processes at the cell biology level that can be modelled as dynamical systems.

6.1 Gene regulation switches

Let's consider a synthetic switch that was created in *E.coli* as a simple construction to understand possibly more complex biological switches. The construction consists of two repressor proteins, whose transcription is mutually regulated. This arrangement gives rise to feedback, and we will see that it allows for a very non-trivial switch between steady states, depending on the initial conditions of the system.

The concentrations of the two proteins are c_1 and c_2, and we want to write equations for the time derivatives of concentration. Each protein is subject to two processes:
(1) degradation at a rate γ, and
(2) its expression, but regulated via the concentration of the other protein. Let's assume that there is a basal (un-repressed) rate r, and that the actual rate of expression is $r(1 - p_{bound})$. If we take the rate of binding to be a Hill function of some order n,

$$p_{bound}(c_1) = \frac{K_b c_1^n}{1 + K_b c_1^n},$$

with K_b the binding constant for the repressor. The expression of protein 2 will then be given by:

$$r(1 - p_{bound}) = \frac{r}{1 + K_b c_1^n}$$

This gives us the coupled equations:

$$\begin{aligned} \frac{dc_1}{dt} &= -\gamma c_1 + \frac{r}{1 + K_b c_2^n} \\ \frac{dc_2}{dt} &= -\gamma c_2 + \frac{r}{1 + K_b c_1^n}. \end{aligned}$$

These can be made dimension-less by expressing concentrations in units of $K_b^{-1/n}$, and time in units of γ^{-1}. Then the equations are:

$$\begin{aligned} \frac{du}{dt} &= -u + \frac{\alpha}{1+v^n} \\ \frac{dv}{dt} &= -v + \frac{\alpha}{1+u^n}, \end{aligned}$$

where $\alpha = rK_b^{1/n}/\gamma$.

We can see that there is always one steady state solution:

$$u^* = v^* = \frac{\alpha}{1+v^{*n}}.$$

Let's see if there are other steady state solutions. Let's consider $n = 2$ to proceed with calculus. The steady state values have to satisfy

$$u^* = \frac{\alpha}{1+\left(\frac{\alpha}{1+u^{*2}}\right)^2},$$

and the corresponding equation for v^*. This can be expanded as:

$$(u^{*2} - \alpha u^* + 1)(u^{*3} + u^* - \alpha) \; = \; 0.$$

The cubic polynomial here can be shown to have only one zero, and by some inspection you can see that it is the solution with $u^* = v^*$. The quadratic however can have 0 (if $\alpha < 2$), 1 (if $\alpha = 2$), or 2 (if $\alpha > 2$) solutions, depending on the value of α. In the 2-solution regime, the concentrations are not the same! The solution with $u^* = v^*$ exists for all α, but it is unstable for $\alpha > 2$.

Calculate phase portraits of this system.

6.2 Oscillations in gene expression

Another ubiquitous dynamical element are coupled equations capable of sustaining oscillations. It has even been proposed that, much like FM vs. AM radio, oscillatory dynamics is used by some cell processes to code and transmit information robustly. One simple set of equations that gives rise to oscillations is a gene regulated by both an activator and a repressor:
- the repressor binds as a dimer, and represses production of the activator
- the activator also binds as a dimer, and increases the production of itself, and also of the repressor.

Then the rate equations can be written as:

$$\begin{aligned}\frac{dc_A}{dt} &= -\gamma_A c_A + r_{0A}\frac{1}{1+(C_A/K_d)^2+(C_R/K_D)^2} + \\ &\quad + r_A\frac{(c_A/K_d)^2}{1+(C_A/K_d)^2+(C_R/K_D)^2} \\ \frac{dc_R}{dt} &= -\gamma_R c_R + r_{0R}\frac{1}{1+(C_A/K_d)^2} + r_R\frac{(c_A/K_d)^2}{1+(C_A/K_d)^2},\end{aligned}$$

where r_{0A}, r_{0R} are the basal expression rates, and r_A, r_R are the regulated rates in the presence of the activator bound.
As before, it is possible to write the equations in dimension-less form:

$$\begin{aligned}\frac{d\tilde{c}_A}{dt} &= -\tilde{\gamma}_A\tilde{c}_A + \frac{\tilde{r}_{0A}+\tilde{r}_A\tilde{c}_A^2}{1+\tilde{c}_A^2+\tilde{c}_R^2} \\ \frac{d\tilde{c}_R}{dt} &= -\tilde{c}_R + \frac{\tilde{r}_{0R}+\tilde{r}_R\tilde{c}_A^2}{1+\tilde{c}_A^2}.\end{aligned}$$

Oscillations can arise if there is a separation of timescales between the activator and repressor dynamics. 'Nullclines' are the locus of points achieved by the repressor or activator at steady state, given fixed values of activator or repressor, respectively. They are obtained by setting the time derivatives equal to zero, and we have:

$$\begin{aligned}\tilde{c}_R &= \sqrt{-1-\tilde{c}_A^2+\frac{\tilde{r}_{0A}+\tilde{r}_A\tilde{c}_A^2}{\tilde{\gamma}_A\tilde{c}_A}} \\ \tilde{c}_R &= \frac{\tilde{r}_{0R}+\tilde{r}_{AR}\tilde{c}_A^2}{1+\tilde{c}_A^2}.\end{aligned}$$

See fig.19.51 (Phillips et al., 2013).

Evolution

7

7.1 Concepts in evolution

full credit: Notes on Population Genetics, Graham Coop, UC Davies

Evolution is change over time. Biological evolution is change over time in the genetic composition of populations. We define the genetic composition of a population to be the set of genomes that the individuals in our population carry. While at first this definition of evolution seems at odds with the common textbook view of the evolution of phenotypes (such as the changing shape of the finch beaks over generations) it is genetic changes that underpin these phenotypic changes. The genetic composition of the population can alter due to the death of individuals or the migration of individuals in or out of the population. If our individuals have different numbers of children, this also alters the genetic composition of the population in the next generation. Every new individual born into the population subtly changes the genetic composition of the population. Their genome is a unique combination of their parents' genomes, having been shuffled by segregation and recombination during meioses, and possibly changed by mutation. Population genetics is the study of the genetic composition of natural populations. It seeks to understand how this composition has been changed over time by the forces of mutation, recombination, selection, migration, and genetic drift. To understand how these forces interact, it is helpful to develop simple theoretical models to help our intuition. In these notes we will work through these models and summarize the major areas of population genetic theory. We will also highlight areas in which physics as contributed to create new and more realistic models to describe the evolution of populations. Throughout the course we will see that these simple models yield accurate predictions, such that much of our understanding of the process of evolution is built on these models.

7.1.1 Allele and genotype frequencies

Population genetics emerged from early efforts to reconcile Mendelian genetics with Evolution. Thanks to Mendel and Mendelian genetics (and a lot of prior and subsequent work), we understand that the genome of an individual is formed from two gametes that

fused to form a zygote. The genomes of each gamete originate from a parental genome through meiosis, in particular the segregation and recombination of the parental genome's two gametes. Loci and alleles are the basic currency of population genetics–and indeed of genetics. Each individual's genetic makeup is defined in their genome. A locus (plural: loci) is a specific spot in the genome. A locus may be an entire gene, or a single nucleotide base pair such as A-T. At each locus, there may be multiple genetic variants segregating in the population—these different genetic variants are known as alleles. For example, at a particular nucleotide site in the genome, a population may segregate for A-T and G-C base pairs (note that due to the complementary nature of DNA, it will suffice to say the site segregates for A and G variants). If all individuals in the population carry the same allele, we say that the locus is monomorphic; at this locus there is no genetic variability in the population. If there are multiple alleles in the population at a locus, we say that this locus is polymorphic.

Note that not all populations are made of diploid individuals. Bacteria carry only one copy of their genome and they reproduce asexually, meaning that the daughter will be genetically identical to mother, modulus errors that occur in the DNA replication step.

7.1.2 Mutation, selection and genetic drift

New alleles can be introduced in a population through genetic mutations. Mutations occur when errors are made in the process of DNA replication. These errors are normally very rare (see Fig. 7.1), since often mutations can result in a protein not folding correctly or not functioning properly. In these cases, mutations are called *deleterious*, as they will negatively affect the growth rate of the individual carrying them, or even impede reproduction alltogether (*lethal mutations*). Why then does nature allow the possibility of mutations? Occasionally mutations can confer big advantages, such as resistance to an antibiotic, or the ability to metabolize a new carbon source. These are called *beneficial* mutations, as they positively affect the growth rate of the carrier. Finally, there are mutations that do not lead to any change in growth rate. These are called *neutral mutations* and are often linked to synonymous substitutions, which do not alter the aminoacid composition of a protein.

If mutations occur randomly and most of them are deleterious, why don't we observe them routinely in populations? As we mentioned, deleterious mutations confer a *fitness* disadvantage to the individual carrying them, meaning that such individual will be less likely to generate offspring in the next generation. Since these changes are hereditary, the limited offspring will itself be affected by the same disadvantage and therefore replicate even less. Eventually, *selection* will purge these deleterious variants from

organism	mutations/ base pair/ replication	mutations/ base pair/ generation	mutations/ genome/ replication	BNID
multicellular				
human *H. sapiens*	10^{-10}	$1–4 \times 10^{-8}$ (mitochondria: 3×10^{-5})	0.2–1	105813, 10(109959, 10 111228
mouse *M. musculus*	2×10^{-10}	10^{-8}	0.5	100315, 106
D. melanogaster	3×10^{-10}	10^{-8}	0.06	100365, 106
C. elegans	$10^{-10}–10^{-9}$	10^{-8}	0.02–0.2	100290, 100 103520, 107
unicellular				
bread mold *N. crassa*		10^{-10}	0.003	100355, 100
budding yeast		$10^{-10}–10^{-9}$	0.003	100458, 1004
E. coli		$10^{-10}–10^{-9}$	0.0005–0.005	106748, 1002
DNA viruses				
bacteriophage T2 & T4		2×10^{-8}	0.004	103918, 1039
bacteriophage lambda		10^{-7}	0.004	100222, 1057
bacteriophage M13		10^{-6}	0.005	106788
RNA viruses				
bacteriophage Qβ		10^{-3}	7	106762
poliovirus		10^{-4}	1	106760
vesicular stomatitis virus		3×10^{-4}	4	106760
influenza A		10^{-5}	1	106760
RNA retroviruses				
spleen necrosis virus		2×10^{-5}	0.2	106762
moloney murine leukemia virus		4×10^{-6}	0.03	106760
rous sarcoma virus		5×10^{-5}	0.4	106762

Fig. 7.1 Table summarizing typical mutation rates in different organisms (from Physical Biology of the Cell).

the population and we won't be able to observe them. The opposite process occurs for beneficial mutations, which instead tend to spread in the population.

What about neutral mutations? If growth was a purely deterministic process with no source of noise, one would expect neutral mutations to maintain their frequency over time. However, this is not the case, because individuals do not always replicate exactly once at each generation. The stochastic fluctuations associated with the random process of replication come under the name of *genetic drift*. Analogously to selection, genetic drift tends to reduce the population diversity generated by mutations. Unlike selection, though, genetic drift's action is purely random and in many circumstances can act against the purging/enriching process of deleterious/beneficial mutations, respectively. In the following, we will provide a mathematical framework to model genetic drift and then its relationship with selection.

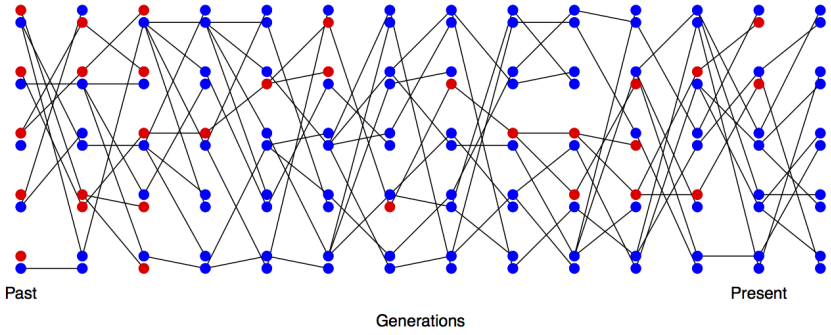


Fig. 7.2 Loss of heterozygosity over time, in the absence of new mutations. A diploid population of 5 individuals over the generations, with lines showing transmission. In the first generation every individual is a heterozygote.

7.2 Genetic drift and neutral diversity

Various sources of randomness are inherent in evolution. One major source of stochasticity in population genetics is genetic drift. Genetic drift occurs because more or less copies of an allele by chance can be transmitted to the next generation. This can occur because by chance the individuals carrying a particular allele can leave more or less offspring in the next generation. In a sexual population genetic drift also occurs because Mendelian transmission means that only one of the two alleles in an individual, chosen at random at a locus, is transmitted to the offspring.

7.2.1 Loss of heterozygosity due to genetic drift

Genetic drift will, in the absence of new mutations, slowly purge our population of neutral genetic diversity as alleles slowly drift to high or low frequencies and are lost or fixed over time.

Imagine a population of a constant size N diploid individuals, and that we are examining a locus segregating for two alleles that are neutral with respect to each other. This population is randomly mating with respect to the alleles at this locus. See Figure 7.2 to see how genetic drift proceeds, by tracking alleles within a small population.

In generation t our current level of heterozygosity is H_t, i.e. the probability that two randomly sampled alleles in generation t are non-identical is H_t. Assuming that the mutation rate is zero (or vanishing small), what is our level of heterozygosity in generation $t+1$?

In the next generation $t+1$ we are looking at the alleles in the offspring of generation t. If we randomly sample two alleles in generation $t+1$ which had different parental alleles in generation t then it is just like drawing two random alleles from generation t. So the probability that these two alleles in generation $t+1$, that

have different parental alleles in generation t, are non-identical is H_t.

Conversely, if our pair of alleles have the same parental allele in the proceeding generation (i.e. the alleles are identical by descent one generation back) then these two alleles must be identical (as we are not allowing for any mutation).

In a diploid population of size N individuals there are $2N$ alleles. The probability that our two alleles have the same parental allele in the proceeding generation is $\frac{1}{2N}$, the probability that they have different parental alleles is $1 - \frac{1}{2N}$ So by the above argument the expected heterozygosity in generation $t + 1$ is

$$H_{t+1} = \frac{1}{2N} \times 0 + (1 - \frac{1}{2N})H_t.$$

By this argument if the heterozygosity in generation 0 is H_0 our expected heterozygosity in generation t is

$$H_t = (1 - \frac{1}{2N})^t H_0$$

i.e. the expected heterozygosity with our population is decaying geometrically with each passing generation. If we assume that $\frac{1}{2N} \ll 1$ then we can approximate this geometric decay by an exponential decay such that

$$H_t = H_0 \exp(-\frac{t}{2N})$$

7.2.2 Levels of diversity maintained by a balance between mutation and drift

Looking backwards in time from one generation to the next, we are going to say that two alleles which have the same parental allele (i.e. find their common ancestor) in the preceding generation have *coalesced*, and refer to this event as a *coalescent event*.

The probability that our pair of randomly sampled alleles have coalesced in the preceding generation is $\frac{1}{2N}$, the probability that our pair of alleles fail to coalesce is $1 - \frac{1}{2N}$.

The probability that a mutation changes the identity of the transmitted allele is μ per generation. So the probability of no mutation occurring is $1 - \mu$. We'll assume that when a mutation occurs it creates some new allelic type which is not present in the population. This assumption (commonly called the infinitely-many-alleles model) makes the math slightly cleaner, and also is not too bad an assumption biologically. See Figure 7.3 for a depiction of mutation-drift balance in this model over the generations.

This model let's us calculate when our two alleles last shared a common ancestor and whether these alleles are identical as a result of failing to mutate since this shared ancestor. For example we can work out the probability that our two randomly sampled alleles coalesced 2 generations in the past (i.e. they fail to coalesce

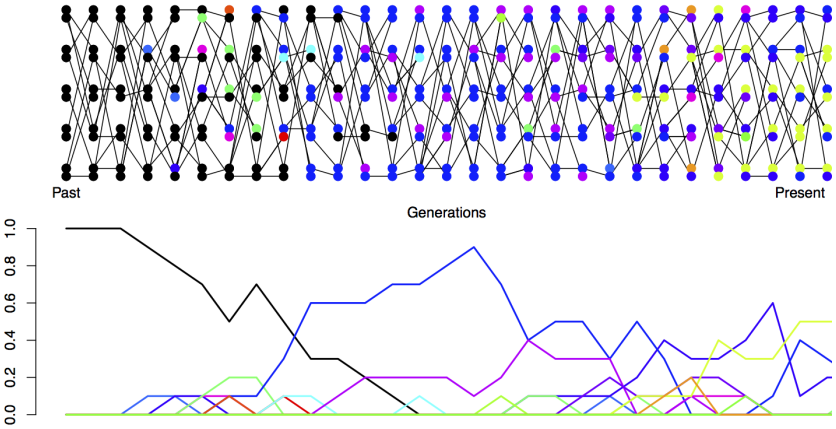


Fig. 7.3 Mutation-drift balance. A diploid population of 5 individuals. In the first generation everyone has the same allele (black). Each generation the transmitted allele can mutate and we generate a new colour. In the bottom plot I trace the frequency of alleles in our population over time.

in generation 1 and then coalescent in generation 2), and that they are identical as

$$(1-\frac{1}{2N})\frac{1}{2N}(1-\mu)^4$$

note the power of 4 is because our two alleles have to have failed to mutate through 2 meioses each.

More generally the probability that our alleles coalesce in generation $t+1$ and are identical due to no mutation to either allele in the subsequent generations is

$$P_{\text{coal}}(t+1)=\frac{1}{2N}(1-\frac{1}{2N})^t(1-\mu)^{2(t+1)}.$$

Over long times, $t\approx t+1$, so we can write

$$P_{\text{coal}}(t+1)=\frac{1}{2N}(1-\frac{1}{2N})^t(1-\mu)^{2t}.$$

This gives us the approximate probability that two alleles will coalesce in the $(t+1)^{th}$ generation. In general, we may not know when two alleles may coalesce: they could coalesce in generation t=1,t=2,..., and so on. Thus, to calculate the probability that two alleles coalesce in any generation before mutating, we can write:

$$P_{\text{coal}}=\sum_{t=1}^{\infty}P_{\text{coal}}(t)$$

If we assume a large population, $\frac{1}{2N}\ll 1$ and small mutation rate $\mu\ll 1$, then we can approximate the geometric decay with an exponential decay,

$$P_{\text{coal}}(t+1)\approx\frac{1}{2N}\exp[-t(2\mu+1/(2N))],$$

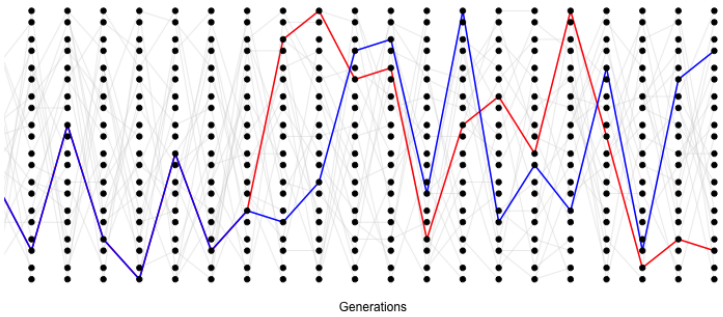


Fig. 7.4 A simple simulation of the coalescent process. The simulation consists of a diploid population of 10 individuals (20 alleles). In each generation, each individual is equally likely to be the parent of an offspring (and the allele transmitted is indicated by a light grey line). We track a pair of alleles, chosen in the present day, back 14 generations untill they find a common ancestor.

and the summation with an integral:

$$P_{\text{coal}} \approx \frac{1}{2N} \int_0^{\infty} \exp[-t(2\mu + 1/(2N))]dt = \frac{1}{1 + 4N\mu}$$

Then, the complementary probability that our pair of alleles are non-identical (or heterozygous) is simply one minus this. This gives the equilibrium heterozygosity in a population at equilibrium between mutation and drift:

$$H = \frac{4N\mu}{1 + 4N\mu}.$$

The component $4N\mu$ is often referred to as θ and represents the population-scaled mutation rate. So all else being equal, species with larger population sizes should have proportionally higher levels of neutral polymorphism.

7.2.3 The effective population size

In practice populations rarely conform to our assumptions of being constant in size with low variance in reproduction success. Real populations experience dramatic fluctuations in size, and there is often high variance in reproductive success. Thus rates of drift in natural populations are often a lot higher than the census population size would imply. To cope with this population geneticists often invoke the concept of an effective population size N_e. In many situations (but not all), departures from model assumptions can be captured by substituting N_e for N.

Specifically the effective population size N_e is the population size that would result in the same rate of drift in an idealized constant population size as that observed in our true population

(following our modeling assumptions). If population sizes vary rapidly in size, we can (if certain conditions are met) replace our population size by the harmonic mean population size. Consider a diploid population of variable size, whose size is N_t t generations into the past. The probability our pairs of alleles have not coalesced by the generation t^{th} is given by

$$\prod_{i=1}^{t}(1-\frac{1}{2N_i}).$$

Note that this simply collapses to our original expression $(1-\frac{1}{2N})^t$ if N_i is constant in time. If $\frac{1}{N_i}$ is small, then $1-\frac{1}{2N_i} \approx \exp(-1/2N_i)$, then

$$\prod_{i=1}^{t}(1-\frac{1}{2N_i}) \approx \Pi_{i=1}^{t}\exp(-1/2N_i) = \exp(-\sum_{i=1}^{t}1/2N_i).$$

So the variable population coalescent probabilities are still of the same form but the exponent has changed. Comparing the exponent in the two cases we see

$$t/2N = \sum_{i=1}^{t}1/2N_i$$

so that we can define the effective population size as

$$N_e = \frac{1}{1/t\sum_{i=1}^{t}1/N_i}$$

Many processes can affect the value of the effective population size, such as reproductive success, population structure, etc... We will see more examples later in the chapter.

7.2.4 The coalescent and patterns of neutral diversity

Thinking back to our calculations we made about the loss of neutral heterozygosity and equilibrium levels of diversity, you'll note that we could first specify what generation a pair of sequences coalesce in, and then calculate some properties of heterozygosity based on that. That's because neutral mutations do not affect the probability that an individual transmits that allele, so don't affect the way in which we can trace ancestral lineages back. As such it will often be helpful to consider the time to the common ancestor of a pair of sequences, and then think of the impact of that on patterns of diversity. The probability that a pair of alleles have failed to coalesce in t generations and then coalesce in the $t+1$ generation back is

$$\frac{1}{2N}(1-\frac{1}{2N})^t \approx \frac{1}{2N}e^{-t/2N}$$

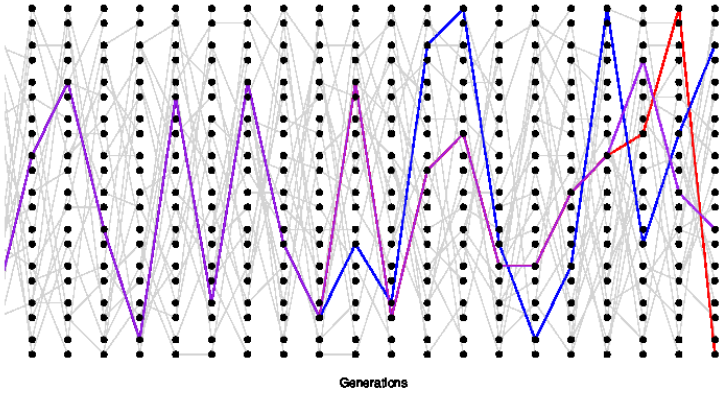


Fig. 7.5 A simple simulation of the coalescent process for three lineages. We track the ancestry of three modern-day alleles, the first pair (blue and purple) coalesce four generations back their are then two independent lineages we are tracking, this pair then coalesces twelve generations in the past. Note that different random realizations of this process will differ from each other a lot.

thus the coalescent time of a pair of sequences T_2 is approximately exponentially distributed with a rate $1/(2N)$. The mean coalescent time of a pair of a pair of alleles is $2N$ generations.

Conditional on a pair of alleles coalescing t generations ago there are $2t$ generations in which a mutation could occur. Thus the probability of our pair of alleles are separated by j mutations since they last shared a common ancestor is

$$P(j|T_2 = t) = \binom{2t}{j} \mu^j (1-\mu)^{2t-j}$$

i.e. mutations happen in j generations, and do not happen in $2t - j$ generations. Assuming that $\mu \ll 1$ and $j \ll 2t$ then we can approximate the equation above with a Poisson distribution

$$P(j|T_2 = t) = \frac{(2\mu t)^j e^{-2\mu t}}{j!}.$$

As our expected coalescent time is $2N$ generations (which follows from the expected value of exponential distributions), the expected number of mutations separating two alleles drawn at random from the population is

$$E(j) = 2\mu E(t) = 4N\mu = \theta$$

If experimentally we observe a given number of differences between two alleles in the sample, this formula allows to estimate the population size.

Usually we are not just interested pairs of alleles, or the average pairwise diversity, we are interested in the properties of diversity

in samples of a number of alleles drawn from the population. To allow for this instead of just following a pair of lineages back until they coalesce, we can follow the history of a sample of alleles back through the population.

When we sample i alleles there are $i(i-1)/2$ pairs, thus the probability that no pair of alleles coalesces in the preceding generation is

$$(1-\frac{1}{2N})^{\binom{i}{2}} \approx (1-\frac{\binom{i}{2}}{2N})$$

while the probability of any coalescing is $\approx \frac{\binom{i}{2}}{2N}$.

When there are i alleles the probability that we wait until the $t+1$ generation before any pair of alleles coalesce is

$$\frac{\binom{i}{2}}{2N}(1-\frac{\binom{i}{2}}{2N})^t \approx \frac{\binom{i}{2}}{2N}\exp(-\frac{\binom{i}{2}}{2N}t)$$

thus the waiting time T_i to the first coalescent event in a sample of i allels is exponentially distributed with a rate $\frac{\binom{i}{2}}{2N}$. When a pair of alleles first find a common ancestral allele some number of generations back further into the past we only have to keep track of that common ancestral allele for the pair. Thus when a pair of alleles in our sample of i alleles coalesce, we then switch to having to follow $i-1$ alleles back. Then when a pair of these $i-1$ alleles coalesce, we then have to follow $i-2$ alleles back. This process continues until we coalesce back to a sample of two, and from there to a single most recent common ancestor (MRCA).

$$T_{MRCA} = \sum_{i=2}^{n} T_i.$$

As our coalescent times for different i are independent, the expected time to the most recent common ancestor is

$$E(T_{MRCA}) = \sum_{i=2}^{n} E(T_i) = \sum_{i=2}^{n} \frac{2N}{\binom{i}{2}}.$$

Using the fact that $\frac{1}{i(i-1)} = \frac{1}{i-1} - \frac{1}{i}$ with some rearrangement we can write

$$E(T_{MRCA}) = 4N(1-\frac{1}{n}).$$

Mutations fall on lineages of the coalescent genealogy. These mutations affect all descendants of this lineage, and under the infinitely-many-sites assumption, create a new segregating site for each new mutation. The mutation process is a Poisson process, and the longer a particular lineage branch, the more mutations that can accumulate on it. The total number of segregating sites in the genealogy is thus a function of the total amount of time in the genealogy of the sample, or the sum of all the genealogy branch lengths, T_{tot}. Since our coalescent genealogies are bifurcating (only

two lineages coalesce at once), our total amount of time in the genealogy is:

$$T_{tot} = \sum_{i=2}^{n} iT_i$$

as when there are i lineages each contributes a time T_i to the total time. Taking the expectation of the total time in the genealogy

$$E(T_{tot}) = \sum_{i=2}^{n} i\frac{2N}{\binom{i}{2}} = \sum_{i=2}^{n} \frac{4N}{i-1} = \sum_{i=1}^{n-1} \frac{4N}{i},$$

so our expected total amount of time in the genealogy scales linearly with our population size. Our expected total amount of time is also increasing with sample size but is doing so very slowly. To see this more carefully we can see that for large n

$$E(T_{tot}) = \sum_{i=1}^{n-1} \frac{4N}{i} \approx 4N \int_{1}^{n-1} \frac{di}{i} = 4N \log(n-1)$$

We saw above that the number of mutational differences between a pair of alleles that coalescence T_2 generations ago was Poisson with a mean of $2\mu T_2$. A mutation that occurs on any branch of our genealogy will cause a segregating polymorphism in the sample (making our infinitely-many-sites assumption). Thus if the total time in the genealogy is T_{tot} there is T_{tot} generations for mutations. So the total number of mutations segregating in our sample S is Poisson with mean μT_{tot}. Thus the expected number of segregating in history a sample of size n is

$$E(S) = \mu E(T_{tot}) = \sum_{i=1}^{n-1} \frac{4N\mu}{i} = \theta \sum_{i=1}^{n-1} \frac{1}{i}.$$

Thus we can use this formula to derive another estimate of the population scaled mutation rate, by setting our observed number of segregating sites in a sample S equal to this expectation.

7.2.5 The fixation of neutral alleles

It is very unlikely that a rare neutral allele accidentally drifts up to fixation; more likely, such an allele will be eventually lost from the population. However, populations experience a large and constant influx of rare alleles due to mutation, so even if it is very unlikely that an individual allele fixes within the population, some neutral alleles will fix.

An allele which reaches fixation within a population is an ancestor to the entire population. In a particular generation there can be only single allele that all other alleles at the locus in later generation can claim as an ancestor. A neutral locus, the actual allele does not affect the number of descendents that the allele has (this follows from the definition of neutrality: neutral alleles don't

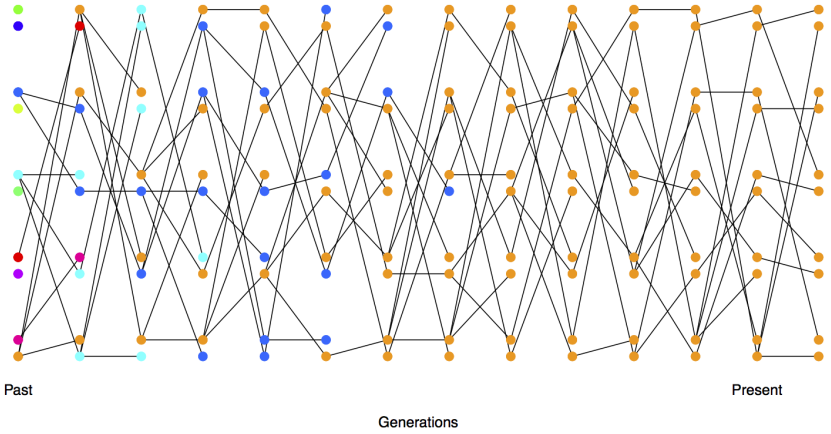


Fig. 7.6 Each allele initially present in a small diploid population is given a different colour so we can track their descendants over time. By the 9th generation all of the alleles present in the population can trace their ancestry back to the orange allele.

leave more or less descendents on average). An equivalent way to state this is that the allele labels don't affect anything; thus the alleles are exchangeable. As a consequence of this, any allele is equally likely to be the ancestor of the entire population. In a diploid population size of size N, there are $2N$ alleles all of which are equally likely to be the ancestor of the entire population at some later time point. So if our allele is present in a single copy, the chance that it is the ancestor to the entire population in some future generation is $\frac{1}{2N}$.

An allele newly arisen mutation only becomes a fixed difference if it is lucky enough to be the ancestor of the entire population. How long does is take on average for such an allele to fix within our population? We've seen that it takes $4N$ generations for a large sample of alleles to all trace their ancestry back to a single most recent common ancestor. Thus it must take roughly $4N$ generations for a neutral allele present in a single copy within the population to the ancestor of all alleles within our population. This argument can be made more precise, but in general we would still find that it takes $\approx 4N$ generations for a neutral allele to go from its introduction to fixation with the population.

7.3 Introducing selection effects

Natural selection occurs when there are differences between individuals in fitness. We may define fitness in various ways. Most commonly, it is defined with respect to the contribution of a phenotype or genotype to the next generation. Differences in fitness can arise at any point during the life cycle. For instance, different genotypes or phenotypes may have different survival probabilities from one stage in their life to the stage of reproduction (viability),

or they may differ in the number of offspring produced (fertility), or both. Here, we define the absolute fitness of a genotype as the expected number of offspring of an individual of that genotype.

7.3.1 Haploid selection model

We start out by modelling selection in a haploid model, as this is mathematically relatively simple. Let the number of individuals carrying alleles A_1 and A_2 in generation t be P_t and Q_t. Then, the relative frequencies at time t of alleles A_1 and A_2 are $p_t = P_t/(P_t + Q_t)$ and $q_t = Q_t/(P_t + Q_t) = 1 - p_t$. Further, assume that individuals of type A_1 and A_2 on average produce W_1 and W_2 offspring individuals, respectively. We call W_i the absolute fitness. Therefore, in the next generation, the absolute number of carriers of A_1 and A_2 are $P_{t+1} = W_1 P_t$ and $Q_{t+1} = W_2 Q_t$, respectively. The mean absolute fitness of the population at time t is

$$\overline{W_t} = W_1 p_t + W_2 q_t,$$

i.e. the sum of the fitness of the two types weighted by their relative frequencies. Note that the mean fitness depends on time, as it is a function of the allele frequencies, which are themselves time dependent.

The frequency of allele A_1 in the next generation is then given by

$$p_{t+1} = \frac{P_{t+1}}{P_{t+1} + Q_{t+1}} = \frac{W_1}{\overline{W_t}} p_t.$$

Importantly, this equation tells us that the change in p only depends on a ratio of fitnesses. Therefore, we need to specify fitness only up to an arbitrary constant. As long as we multiply all fitnesses by the same value, that constant will cancel out and the equation will hold. Based on this argument, it is very common to scale absolute fitnesses by the absolute fitness of one of the genotypes, e.g. the most or the least fit genotype, to obtain relative fitnesses. Here, we will use w_i for the relative fitness of genotype i. If we choose to scale by the absolute fitness of genotype A_1, we obtain the relative fitnesses $w_1 = W_1/W_1 = 1$ and $w_2 = W_2/W_1$.

Without loss of generality, we can therefore write

$$p_{t+1} = \frac{w_1}{\overline{w}} p_t,$$

dropping the dependence of the mean fitness on time in our notation, but remembering it. The change in frequency from one generation to the next is then given by

$$\Delta p_t = p_{t+1} - p_t = \frac{w_1 p_t}{\overline{w}} - p_t = \frac{w_1 - w_2}{\overline{w}} p_t q_t.$$

Assuming that the fitnesses of the two alleles are constant over time, the number of the two allelic types τ generations after time

t are $P_{t+\tau} = W_1^\tau P_t$ and $Q_{t+\tau} = W_2^\tau Q_t$, respectively. Therefore, the relative frequency of allele A_1 after τ generations past t is

$$p_{t+\tau} = \frac{p_t}{p_t + (w_2/w_1)^\tau q_t}.$$

Rearranging the equation and setting $t = 0$, we can work out the time τ for the frequency of A_1 to change from p_0 to p_τ. First, we write

$$p_\tau = \frac{p_0}{p_0 + (w_2/w_1)^\tau q_0}$$

and rearrange to obtain

$$\frac{p_\tau}{q_\tau} = \frac{p_0}{q_0}(\frac{w_1}{w_2})^\tau.$$

Solving for τ gives

$$\tau = \log(\frac{p_\tau q_0}{q\tau p_0}) / \log(\frac{w_1}{w_2}).$$

In practice, it is often helpful to parametrize the relative fitnesses w_i in a specific way. For example, we may set $w_1 = 1$ and $w_2 = 1 - s$, where s is called the selection coefficient. Using this parametrization, s is simply the difference in relative fitnesses between the two alleles. The equation becomes

$$p_{t+\tau} = \frac{p_t}{p_t + q_t(1-s)^\tau}.$$

If $s \ll 1$ (weak selection), then we can approximate

$$p_{t+\tau} = \frac{p_t}{p_t + q_t e^{-s\tau}}.$$

This equation takes the form of a logistic function. That is because we are looking at the relative frequencies of two 'populations' (of alleles A_1 and A_2) that are growing (or declining) exponentially, under the constraint that $p+q = 1$. Moreover, the equation for the time τ it takes for a certain change in frequency to occur becomes

$$\tau = -\log(\frac{p_\tau q_0}{q\tau p_0}) / \log(1-s).$$

If we assume again $s \ll 1$, we can write

$$\tau \approx \frac{1}{s}\log(\frac{p_\tau q_0}{q\tau p_0}).$$

One particular case of interest is the time it takes to go from an absolute frequency of $1/N$ to near fixation in a population of size N. In this case, we have $p_0 = 1/N$, and we may set $p_\tau = 1 - 1/N$, which is very close to fixation. Then, plugging these values, we obtain

$$\tau \approx \frac{2}{s}\log(N).$$

7.4 Interplay between selection and genetic drift

7.4.1 Stochastic loss of strongly selected alleles

Even strongly selected alleles can be lost from the population when they are sufficiently rare. This is because the number of offspring left by individuals to the next generation is fundamentally stochastic. A selection coefficient of $s = 1\%$ is a strong selection coefficient, which can drive an allele through the population in a few hundred generations once the allele is established. However, if individuals have on average a small number of offspring per generation the first individual to carry our allele who has on average 1% more children could easily have zero offspring, leading to the loss of our allele before it ever get a chance to spread. To take a first stab at this problem let's think of a very large haploid population, and in order for this population to stay constant in size we'll assume that individuals without the selected mutation have on average one offspring per generation. While individuals with our selected allele have on average $1 + s$ offspring per generation. We'll assume that the distribution of offspring number of an individual is Poisson distributed with this mean, i.e. the probability that an individual with the selected allele has i children is

$$P_i = \frac{(1+s)^i e^{-(1+s)}}{i!}.$$

Consider starting from a single individual with the selected allele, and ask about the probability of eventual loss P_L of our selected allele starting from this single copy. To derive this we'll make use of a simple argument (derived from branching processes). Our selected allele will be eventually lost from the population if every individual with the allele fails to leave descendants.

(1) In the first generation with probability P_0 our individual leaves no copies of itself to the next generation, in which case our allele is lost

(2) Alternatively it could leave one copy of itself to the next generation (with probability P_1), in which case with probability P_L this copy eventually goes extinct

(3) It could leave two copies of itself to the next generation (with probability P_2), in which case with probability P_L^2 both of these copies eventually goes extinct

(4) More generally it could leave could leave k copies ($k > 0$) of itself to the next generation (with probability P_k), in which case with probability P_L^k all of these copies eventually go extinct

summing over these probabilities we get

$$P_L = \sum_{k=0}^{\infty} P_k P_L^k = e^{-(1+s)} \left(\sum_{k=0}^{\infty} \frac{(P_L(1+s))^k}{k!} \right) = e^{(1+s)(P_L-1)}$$

The probability of escaping loss $P_F = 1 - P_L$ then satisfies

$$1 - P_F = e^{-P_F(1+s)}.$$

If we consider a small selection coefficient $s \ll 1$ such that $P_F \ll 1$ and expanded the exponential we obtain

$$1 - P_F = 1 - P_F(1+s) + P_F^2(1+s)^2/2$$

which results into $P_F = 2s$. Thus even an allele with a 1% selection coefficient has a 98% probability of being lost when it is first introduced into the population by mutation.

We can also adapt this result to a diploid setting. Assuming that heterozygotes for the 1 allele have $1 + (1-h)s$ children, the probability of allele 1 is not lost, starting from a single copy in the population, is

$$P_F = 2(1-h)s$$

for $h > 0$.

7.4.2 The interaction between genetic drift and weak selection

For strongly selected alleles, once the allele has escaped initial loss at low frequencies, their path will be determined deterministically by their selection coefficients. However, if selection is weak the stochasticity of reproduction can play a role in the trajectory an allele takes even when it is common in the population. To see this lets think of our simple Wright-Fisher model. Each generation we allow a deterministic change in our allele frequency, and then binomially sample two alleles for each of our offspring to construct our next generation. So the expected change in our allele frequency within a generation is given just by our deterministic formula. To make things easy on our self lets assume an additive model, i.e., $h = 1/2$, and that $s \ll 1$ so that $\overline{w} \approx 1$. This gives us

$$E(\Delta p) = \frac{s}{2} p(1-p)$$

while our variance in allele frequency is given by

$$Var(\Delta p) = Var(p' - p) = Var(p') = \frac{p'(1-p')}{2N} \approx \frac{p(1-p)}{2N}$$

this variance in our allele frequency follows from the fact that we are binomially sampling $2N$ new alleles in the next generation from a frequency p' and assuming that $p' \approx p$, since $s \ll 1$.

To get our first look at the relative effects of selection vs drift we can simply look at when our change in allele frequency caused selection within a generate is reasonably faithfully passed across the generations. In particular if our expected change in frequency is much great than the variance around this change, genetic drift will play little role in the fate of our selected allele (once the allele is not too rare within the population). When does selection dominates over genetic drift? This will happen if $E(\Delta p) \gg Var(\Delta p)$ when $Ns \gg 1$. Conversely any hope of our selected allele following its deterministic path will be quickly undone if our change in allele frequencies due to selection is much less than the variance induced by drift. So if $Ns \ll 1$ then drift will dominate the fate of our allele.

To make further progress on understanding the fate of alleles with selection coefficients of the order $1/N$ requires more careful modeling. However, we can obtain the probability that under our diploid model, with an additive selection coefficient s, the probability of allele 1 fixing within the population starting from a frequency p is given by

$$\pi(p) = \frac{1 - e^{-2Nsp}}{1 - e^{-2Ns}}.$$

(try to prove this as an exercise). A new allele will arrive in the population at frequency $p = 1/(2N)$, then its probability of reaching fixation is

$$\pi(1/2N) = \frac{1 - e^{-s}}{1 - e^{-2Ns}}.$$

if $s \ll 1$ but $Ns \gg 1$, the $\pi(1/2N) \approx s$, which nicely gives us back our result that we obtained above. To recover our neutral result we can take the limit $s \to 0$ to obtain our neutral fixation probability $1/2N$.

7.4.3 The fixation of slightly deleterious alleles

We can see that weakly deleterious alleles can fix, especially in small populations. To understand how likely it is that deleterious alleles accidently reach fixation by genetic drift, lets assume a diploid model with additive selection (with a selection coefficient of $-s$ against our allele 2). If $Ns \gg 1$ then our deleterious allele (allele 2) can not possibly reach fixation. However, if Ns is not large then

$$\pi(1/2N) \approx \frac{s}{e^{2Ns} - 1}$$

7.5 Beyond well-mixed populations: the role of space

In many realistic scenarios, a population is not well-mixed, where each individual feels the same environmental conditions, but it's

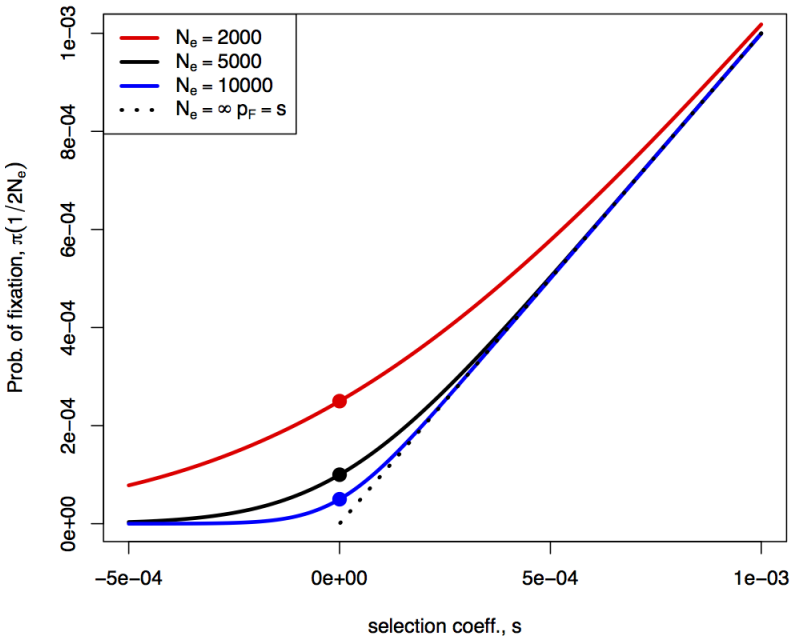


Fig. 7.7 The probability of the fixation of a new mutation with selection coefficient s ($h = 1/2$) in a diploid population of effective size N_e. The dashed line gives the infinite population solution. The dots give the solution for $s \to 0$, i.e. $1/2N_e$.

distributed in space. In the following, we will present different models that have tried to deal with the issue of space and how to include it when describing evolutionary dynamics.

7.5.1 Island model

The most classical example of spatially distributed population is the so-called "island model". The model is based on dividing the population into subpopulations that live in different "islands" or spatially distinct locations. Often some degree of migration is allowed between islands so that subpopulations can exchange individuals. In these models, often the islands have different conditions, therefore mutations that are neutral or beneficial in some islands can be deleterious in others. Because of migration there can be a constant influx of deleterious mutations creating a migration-selection balance which is very similar to the mutation-selection balance. In these models, often migration between islands can be treated very similarly to mutations within one population.

7.5.2 Some theory of spatial distribution of alleles frequencies under deterministic models of selection

Imagine a continuous haploid population spread out along a line. An individual disperses a random distance Δx from its birthplace to the location where it reproduces, where Δx is drawn from the probability density $g()$. To make life simple we will assume that $g(\Delta x)$ is normally distributed with mean zero and standard deviation σ i.e. migration is unbiased an individuals migrate an average distance of σ.

The frequency of allele 2 at time t in the population at spatial location x is $q(x,t)$. Assuming that only dispersal occurs, how does our allele frequency change in the next generation? Our allele frequency in the next generation at location x reflects the migration from different locations in the proceeding generation. Our population at location x receives a contribution $g(\Delta x)q(x+\Delta x,t)$ of allele 2 from the population at location $x+\Delta x$, such that the frequency of our allele at x in the next generation is

$$q(x,t+1) = \int_{-\infty}^{\infty} g(\Delta x)q(x+\Delta x,t)d\Delta x.$$

To obtain $q(x+\Delta x,t)$, let's take a Taylor expansion

$$q(x+\Delta x,t) = q(x,t) + \Delta x\partial_x q|_{x,t} + 1/2\Delta x^2\partial_x^2 q|_{x,t} + ...$$

then

$$q(x,t+1) = q(x,t)+\partial_x q|_{x,t}\int_{-\infty}^{\infty}\Delta x g(\Delta x)d\Delta x+1/2\partial_x^2 q|_{x,t}\int_{-\infty}^{\infty}\Delta x^2 g(\Delta x)d\Delta x$$

Remembering that $g()$ has zero mean and variance σ^2, we find that

$$q(x,t+1) = q(x,t) + \frac{\sigma^2}{2}\partial_x^2 q.$$

In the limit of continuous time, we get

$$\partial_t q = \frac{\sigma^2}{2}\partial_x^2 q$$

This is a diffusion equation, so that migration is acting to smooth out allele frequency differences with a diffusion constant of $D = \sigma^2/2$. This is exactly analogous to the equation describing how a gas diffuses out to equal density, as both particles in a gas and our individuals of type 2 are performing Brownian motion (blurring our eyes and seeing time as continuous)

We will now introduce fitness differences into our model and set the relative fitnesses of allele 1 and 2 at location x to be 1 and $1+s\gamma(x)$, respectively. To make analytical progress in this model we'll have to assume that selection isn't too strong i.e. $s\gamma(x) \ll 1$

for all x. The change in frequency of allele 2 obtained within a generation due to selection is

$$q'(x,t) - q(x,t) \approx s\gamma(x)q(x,t)(1-q(x,t)),$$

which described logistic growth of the favoured allele. Putting our selection and migration together we find

$$\partial_t q(x,t) = s\gamma(x)q(x,t)(1-q(x,t)) + D\partial_x^2 q(x,t).$$

This is a reaction-diffusion equation, which describes the spreading of a beneficial allele in a population. If $\gamma(x) = 1$, this is known as the Fisher-Kolmogorov-Petrovski-Piskinov (FKPP) deterministic equation. Importantly, this equation is satisfied by a travelling wave solution of characteristic speed $v_F = 2\sqrt{D/s}$, which is called the Fisher velocity.

7.5.3 Emergence of resistance in an antibiotic gradient

We will now use the determinstic FKPP equation to derive the probability of resistant mutations arising in an antibiotic gradient (for exersice compare this probability to the well-mixed case). In class, we saw an experiment on *E. coli* growing over a large agar plate made of slabs with step-wise increasing antibiotic concentration. Let's focus on one particular step of the plate and determine the probability that the wild-type population will be able to develop resistance to the next antibiotic concentration.

We assume that the wild-type population density $c(x,t)$ (rescaled by the carrying capacity K) is described by the reaction-diffusion equation

$$\partial_t c(x,t) = D\partial_x^2 c + s_{WT}(x)c - a_{WT}(x)c^2, \tag{7.1}$$

where $a_{WT}(x)$ is the local wild type birth rate, b_{Wt} is the local antibiotic-induced death rate, $s_{WT}(x) = a_{WT}(x) - b_{WT}(x)$ is the local net growth rate of the wild type.

The model ensures that the steady-state local population density c_{SS} depends explicitly on the local death rate $b_{WT}(x)$ when the death and birth rate profiles change sufficiently slowly in space,

$$c_{SS}(x) = 1 - \frac{b_{WT}(x)}{a_{WT}(x)}. \tag{7.2}$$

Given a single resistance mutation arises in the population at position x, its probability $u(x,t)$ to survive for a time t can be derived by modeling the mutant lineage as a branching random walk.

Let $u_x(t)$ denote the probability that a mutation born at lattice site x survives for time t. Denote by $a(x)$ and $b(x)$ the local birth and death rate, respectively, and let D be the migration rate over

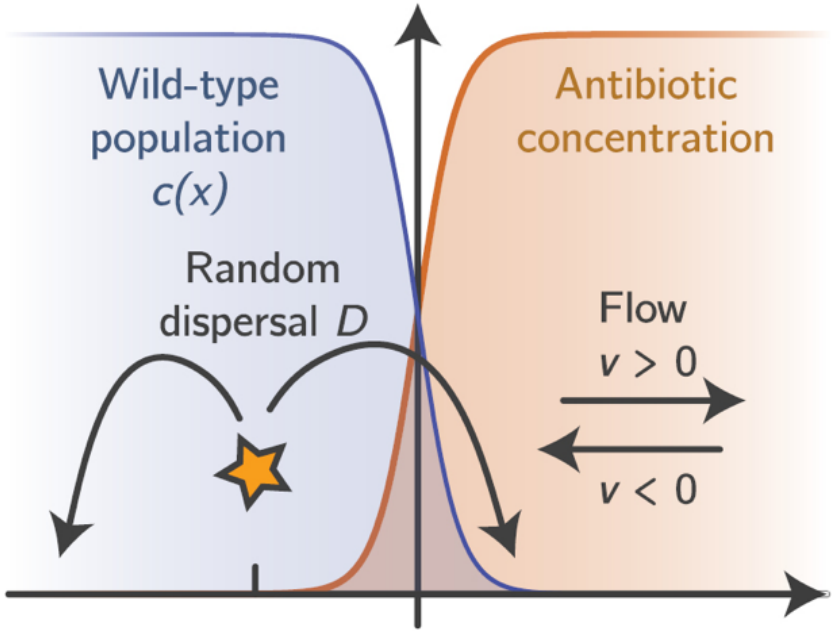


Fig. 7.8 Representation of an antibiotic gradient and the corresponding steady-state wild-type population ($c(x)$). Mutations that confer resistance can occur and generate a mutant population able to live in the antibiotic region. The populations can in principle also be subject to flow.

a distance δx (i.e., one lattice site). Then, the $u(x,t)$ after a short time interval ϵ satisfies the equation

$$u_x(t+\epsilon) = \epsilon a(x)\left\{1-[1-u_x(t)]^2\right\} + \{1-\epsilon\,[a(x)+b(x)]\}\,u_x(t) + \epsilon D\left\{u_{x+\delta x}(t)+u_{x-\delta x}(t)-2u_x(t)\right\}. \tag{7.3}$$

The first term on the right-hand side accounts for the fact that when the initial mutant divides then there are two mutants, and the probability of survival of at least one lineage is 1 minus the square of both lineages disappearing. The second term describes the case of nothing happening in the time interval ϵ.

Letting $\epsilon \to 0$ and performing the Taylor expansion in δx similarly to what we did to derive the FKPP, we obtain

$$\partial_t u = D\partial_x^2 u + s(x)u - a(x)u^2, \tag{7.4}$$

where $s(x) = a(x) - b(x)$.

We assume that the birth rate of wild-type and mutant is identical and constant, $a_{WT}(x) = a_{MT}(x) = a_0$, while the wild-type drug-induced death rate $b_{WT}(x)$ ranges from $-a_0$ to a_0. This implies that effect of the antibiotic is to increase the death rate of the wild type, while the drug-induced death rate of the resistant mutant is zero. The effective growth rate of the mutants is thus determined purely through competition with the wild type, i.e., $s_{\mathrm{MT}}(x) = a_0[1-c(x)]$.

A step-like increase in concentration at $x = 0$ gives rise a net growth rate of a_0 for $x < 0$ and $-a_0$ for $x > 0$, i.e.,

$$s_{WT}(x) = a_0 \left[1 - 2\Theta(\xi)\right]. \tag{7.5}$$

Such a sharp gradient could emerge, for instance, at the boundary of different tissues or organs with different affinities to store antibiotics. Upon rescaling the spatial coordinate by the characteristic length scale $\ell = \sqrt{D/a_0}$, which can be intuitively understood as the typical distance that a mutant individual travels through random dispersal before replicating, the wild-type population density in this case becomes

$$0 = \partial_\xi^2 c + [1 - 2\Theta(\xi)]c - c^2 \tag{7.6}$$

where $\xi = x/\ell$.

For both $\xi > 0$ and $\xi < 0$, this equation can be solved by a mechanical analogy with a particle in the "potential" $U(c) = \pm\frac{c^2}{2} - \frac{c^3}{3}$. For $\xi < 0$, we have the potential $U(c) = c^2/2 - c^3/3$ and the boundary condition $c(-\infty) = 1$ and $c(0) = c_0$, which gives a total energy $E = K + U = 1/6$ because the kinetic energy $K = 0$ at $-\infty$. The population density $c(\xi)$ is then determined through the integral

$$\xi = \int_{c_0}^{c(\xi)} \frac{dc'}{\sqrt{-2U(c') + 2E}} = \int_{c_0}^{c(\xi)} \frac{dc'}{\sqrt{-c^2 + c^3/3 + 1/3}}. \tag{7.7}$$

Similarly, for $\xi > 0$, we have $U(c) = -c^2/2 - c^3/3$ and $E = 0$, and $c(\xi)$ is determined through the integral

$$\xi = \int_{c_0}^{c(\xi)} \frac{dc'}{\sqrt{c^2 + c^3/3}}. \tag{7.8}$$

Both integrals can be solved exactly, and the derivatives matched at $\xi = 0$. The result is

$$c(\xi) = \begin{cases} \frac{3}{2}\tanh\left[\frac{\xi - \xi_-}{2}\right]^2 - \frac{1}{2}, & \xi < 0, \\ \frac{3}{2}\tanh\left[\frac{\xi + \xi_+}{2}\right]^2 - \frac{3}{2}, & \xi \geq 0, \end{cases} \tag{7.9}$$

where $\xi_\pm = 2\text{arctanh}(\frac{1}{3}\sqrt{6 \pm 3 + \sqrt{6}})$.

The population density transitions from 1 to 0 exponentially fast, and we can approximate $c(\xi) \approx \Theta(-\xi)$ when computing the establishment probability, which then approximately obeys the equation

$$0 = \partial_\xi^2 u + \Theta(\xi)u - u^2. \tag{7.10}$$

Using the same mechanical analogy as above, and again matching derivatives at $\xi = 0$, we find

$$u(\xi) = \begin{cases} \frac{1/\sqrt{3}}{\left(1 - \xi/\sqrt{6\sqrt{3}}\right)^2}, & \xi < 0, \\ \frac{3}{2}\tanh\left[\frac{\xi + \xi_u}{2}\right]^2 - \frac{1}{2}, & \xi \geq 0, \end{cases} \tag{7.11}$$

where $\xi_u = 2\text{arctanh}(\frac{1}{3}\sqrt{3 + 2\sqrt{6}})$.

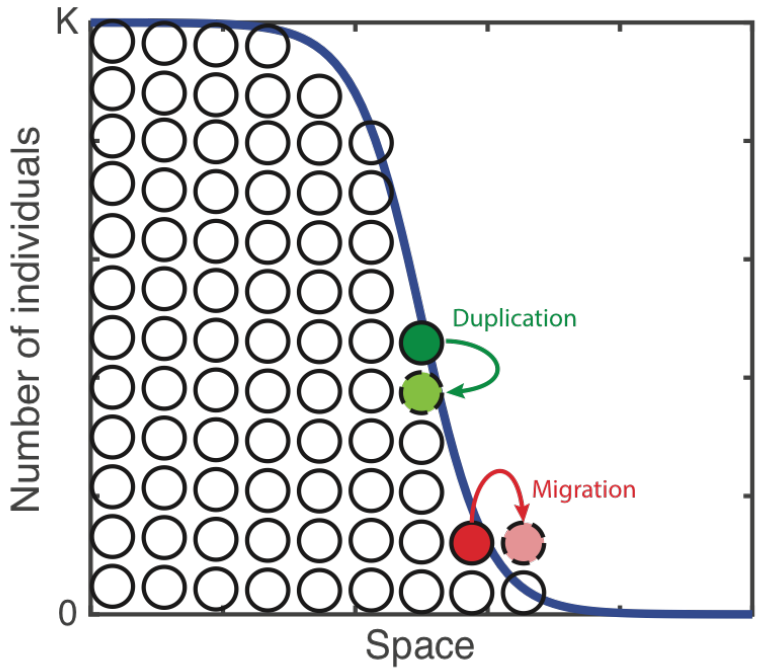


Fig. 7.9 Representation of a stochastic simulation that mimics the growth of a population in 1 dimensional space.

7.5.4 Stochasticity in the FKPP equation

The sections above removed any stochasticity from the FKPP equation. However, genetic drift equally acts in spatial models as it did in well-mixed conditions. In particular, since in the FKPP equation the dynamics is dominated by the individuals at the front of the wave, the corresponding effective population size N_e can be very small and genetic fluctuations very high. Indeed, in these conditions, stochastic effects can be so strong that selection becomes very inefficient and deleterious mutations may, by chance, accumulate in the population, with a phenomenon called expansion load. To better understand this, let's look at the stochastic nature of the FKPP equation by imagining to simulate it.

The FKPP can be used to model the expansion of a population in virgin territory (see Fig. 7.9). At each point in time and space, the population density is defined by $q(x,t)$. At each step in our simulation, one individual is selected and can replicate. A generation will occur once every individual at position x is given a chance to replicate, therefore 1 generation consists of N steps. Or, in other words, 1 simulation step happens at a rate $1/N$. Because of logistic growth, replication occurs if we pick a random individual *and* a random free position (see Fig. 7.9). This means

that at time $t + dt$

$$q(t+dt) = \begin{cases} q+1 & \text{with prob. } \frac{dt}{N} q(N-q) \\ q & \text{with prob. } 1 - \frac{dt}{N} q(N-q) \end{cases}$$

Therefore, at position x,

$$\langle q(x, t+dt) \rangle = q(x,t) + \frac{dt}{N} q(x,t)(N - q(x,t))$$

and variance

$$Var(q(x, t+dt) = \frac{dt}{N} q(x,t)(N - q(x,t))$$

Defining the white noise R_t such that $\langle R_t \rangle = 0$ and $\langle R_t^2 \rangle = 1$, we can write for position x

$$q(x, t+dt) = q(x,t) + \frac{dt}{N} q(x,t)(N - q(x,t)) + R_t \sqrt{\frac{dt}{N} q(x,t)(N - q(x,t))}$$

In the limit of $dt \to 0$, at position x

$$\partial_t q = \frac{q(N-q)}{N} + \eta_t \sqrt{\frac{q(N-q)}{N}}$$

where $\langle \eta_t \eta_{t'} \rangle = \delta(t - t')$. Adding the diffusion term and defining $c = q/N$ as the population density, we get the stochastic FKPP equation:

$$\partial_t c = \partial_x^2 c + c(1-c) + \eta_t \sqrt{\frac{c(1-c)}{N}}$$

.

From this equation, you can see that at the very front of the wave, where we have only one individual, $c \approx 1/N$, and therefore the noise term and the growth term are both $\mathcal{O}(1/N)$ and fluctuations cannot be neglected.

One of the consequences of these large fluctuations is that, similarly to the well-mixed case, weak selection effects can behave almost neutrally so that weakly beneficial mutations may easily get lost and weakly deleterious mutations may instead accumulate at the front of the wave.

7.6 Ecological scales: multispecies evolution

Credit: Models of Life by Kim Sneppen

If we look at evolution over much larger timescales, we often observe signs of cooperative behavior, where many species have often been replaced almost "simultaneously". To model macro-evolutionary patterns we start with units on the size of the main

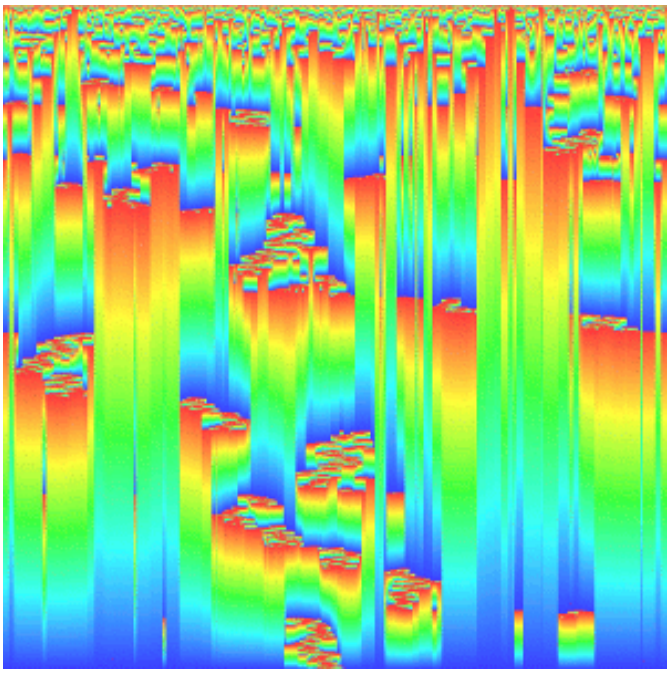


Fig. 7.10 Sample of Bak–Sneppen model evolution: the x-axis represent the position of the different species in the ecosystem and on the y-axis (from top to the bottom) we see the history of the population. Each discontinuity represents an evolution. The color codes the age of the species (red=origin, blue=extinction). (From Wikipedia) Note the spatial clustering that naturally emerges along the simulation.

players at this scale: species. A species consists of many individuals, and the dynamics of a species represent the coarse-grained view of the dynamics of these entire populations. Whereas population dynamics may be influenced by fitness, species dynamics are governed by stabilities. We characterize a species by a number B_i that specifies its stability on very long timescales. An ecosystem of species then consistst of N numbers B_i, each representing a different species, with links that could be predation, collaboration or niche maintanance.

A simple model that describes the dynamics of these species is the Bak-Sneppen model. Let's assume that the numbers B_i with $i = 1, 2, ..., N$ are placed on a line, mimicking a one dimensional model ecosystem. At each time step, one changes the least stable of these species. The fitness of a given species if a function of the species it interacts with, and accordingly, the neighboring species will also change their stabilities. We will use the following update rule: at each step, the smallest $\{B_i\}_{i=1,N}$ are located. Its value and those of its nearest neighbors are replaces by new random numbers in $[0, 1]$.

This is the simplest model that exhibits a phenomenon called self-organized criticality. As the system evolves, the smallest of B_i is eliminated and after a transient period, a statistically stationary

distribution of B is obtained. For a system $N \to \infty$, this distribution is a step-function, where the selected minimum B_{min} is always below a critical B_c, and the distribution is constant above B_c. The value of B_c depends on dimensionality and the details of the update move: for $d = 1$ and two nearest neighbors update, $B_c = 0.667$.

Importantly, over the course of evolution, active sites tend to be localized in the same region of the model ecosystem (Fig. 7.10. Thus evolution is reinforced locally, bridging punctuated equilibrium in single-species evolution to larger quantum evolution and origination of new taxonomic groups.

Terms and quantities in cell biology

8

The introductory material of the first couple of lectures can be found on the overheads. Here is firstly a glossary of terms, most of which should become familiar after a few lectures and on reading the first chapters of (Phillips et al., 2013), and secondly a table of useful data.

Glossary

Extended and modified from p.265 of U.Alon, and p.297 of Sneppen-Zocchi books.

Activator - A transcription factor that increases the rate of transcription of a gene when it binds a specific site in the gene's promoter.

Activation threshold - Concentration of activator in its active state needed for half-maximal activation of a gene.

Adaptation - Decreasing response to a stimulus that is applied continuously.

Adaptation time - Time for output to recover to 50% of prestimulus level following a step stimulus.

Allele - One of a set of alternative forms of a gene. In a diploid organism, such as most animal cells, each gene has two alleles, one on each of the two sister chromosomes.

Amino acid - A molecule that contains both an amino group (NH_2) and a carboxyl group (COOH). Amino acids are linked together by peptide bonds and serve as the constituents of proteins.

AND gate - A logic function of two inputs that outputs a one only if both inputs are equal to one.

Anti-motif - A pattern that occurs in a network less often than expected at random.

Antibody - A protein produced by a cell of the immune system that recognizes a protein present in or on invading microorganisms.

Antigen - A part of a protein or other molecule that is recognized by an antibody.

Arabinose - A sugar utilized by *E. coli* as an energy and carbon source, using the *ara* genes. Arabinose is not pumped into the cells if glucose, a better energy source, is present.

ATP (adenosine triphosphate) - A molecule that is the main currency in the cellular energy economy. The conversion of ATP to ADP (adenosine diphosphate) liberates energy.

***B. subtilis* (*Bacillus subtilis*)** - A bacterium commonly found in the soil. It forms durable spores upon starvation. A model organism for study, and commonly used in synthetic biology.

Binomial distribution - A statistical distribution that describes, for example, the probability for k heads out of n throws of a coin that has probability p to give heads and $l-p$ to give tails.

Cell - see the earlier chapters of the handout.

Cell line - Unicellular organisms like bacteria and some algae species quite naturally grow and divide in the right conditions. On the other spectrum of what we imagine as 'complexity' of lifeforms, for example in vertebrates, after the embryo develops into the adult there is very little cell growth and division. A cell taken out of such an organism is called a 'primary cell', and has 'differentiated' into a particular cell type. In the right biochemical environment it might be possible to make that cell grow and divide a few times ("passages") but not indefinitely. Disruption of some specific genes, which are often the same ones at the heart of a cancer mutation, can turn a cell into a 'cell line' where the cells grow and divide indefinitely. This can be extremely useful in research, although their biology is certainly a bit different. See also: 'model organism' and 'stem cell'.

Chemoreceptor - A receptor that responds to the presence of a particular chemical.

Chemotaxis - Movement up spatial gradients of specific chemicals (attractants), or down gradients of specific chemicals (repellents).

Chromosome - A strand of DNA with its associated proteins,

found in the nucleus; carries genetic information.

Circadian rhythm - A daily rhythmical cycle of cellular activity. Generated by a biochemical oscillator in many different cells in animals, plants, and microorganisms. Ihe oscillations can be entrained by periodic temperature and light signals. The oscillator runs also in the absence of entraining external signals (usually with a period somewhat different than 24 hrs).

Coalescence - The merging of genetic lineages backwards in time to a recent common ancestor.

Codon - Three consecutive letters on an mRNA. There are 64 codons (each made of three letters, A, C, G, and U). These code for the 20 amino acids (with most amino acids represented by more than one codon). Three of the codons signal translational stop (end of the protein).

Coherent feed-forward loop - A feed-forward loop in which the sign of the direct path from X to Z is the same as the sign of the indirect path from X through Y to Z.

Complementary sequence Sequence of bases that can form a double-stranded structure by matching base pairs. The complementary sequence to base pairs C-T-A-G is G-A-T-C.

Cooperativity More than the sum of its parts. Acting cooperatively means that one part helps another to build a better functioning system. Cooperative bindings include dimerization, tetramerization, and binding between transcription factors on adjacent DNA sites.

Cost-benefit analysis - A theory that seeks the optimal design such that the difference between the fitness advantage gained by a system (benefit) and fitness reduction due to the cost of its parts is maximal.

Cytoplasm - The viscous, semiliquid substance contained in the interior of a cell. The cytoplasm is densely picked with proteins ('crowding').

Degree-preserving random networks - An ensemble of randomized networks that have the same degree sequence (the number of incoming and outgoing edges for each node in the network) as the real network. Despite the fact that the degree sequence is the same, the identity of which node connects to which other node is randomized. Such random networks can be generated on the computer by randomly switching pairs of edges, repeating the

switching operation many times until the network is randomized. For a given real network, many thousands of different randomized degree-preserving networks can usually be readily generated.

Developmental transcription networks - Networks of transcription interactions that guide changes in cell type. Important examples are networks that guide the selection of cell fate as cells in the embryo differentiate into tissues. Developmental transcription networks work on the timescale of cell generations and often make irreversible decisions. They stand in contrast to sensory transcription networks that govern responses to environmental signals.

Differentiation - The process in which a cell changes to a different type of cell (same genome).

Diploid Individual - Individual carrying two copies of its genome. The two copies do not have to be identical.

Distributions Some common ones:
exponential

$$p(t) \sim \exp(-t/t).$$

If t is a waiting time this is the distribution for a random uncorrelated signal. In that case the expected waiting time for the next signal does not change as time passes since the last signal.

power law

$$p(t) \sim 1/t^{\alpha}.$$

For example, if t is a waiting time, then expected waiting time for the next signal increases as time passes since the last signal.

normal or Gaussian distribution
Obtained by sum of exponentially bounded random numbers that are uncorrelated. Distribution:

$$p(x) \sim \exp\left(-x^2/\sigma^2\right).$$

log normal
Obtained by product of exponentially bounded random numbers that are uncorrelated. If x is normal distributed then $y = \exp(x)$ is log normal:

$$q(y)dy \sim \exp(-\log(y)^2/\sigma^2)dy/y \,\text{and}\, \sim dy/y$$

for y within a limited interval.

stretched exponentials
These are of the form

$$p(x) \sim \exp(-x^{\alpha}).$$

Pareto-Levi
Obtained from the sum of numbers, each drawn from a distribution $\propto x^{-\alpha}$. A Pareto-Levi distribution has a typical behavior like a Gaussian, but its tail is completely dominated by the single largest event. Thus a Pareto-Levi distribution has a power-law tail.

binomial
with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p. Probability of k successes is:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Poisson
expresses the probability of a given number of events (i.e. k, discrete) occurring in a fixed interval of time and/or space, if these events occur with a known average rate λ and independently of the time since the last event. The probability of a random variable being $X = k$ is:

$$p(X = k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}.$$

It has the special property that $\lambda = \langle p(X) \rangle = \mathrm{variance}(p(X))$.

DNA (deoxyribonucleic acid) - A long molecule composed of two interconnected helical strands. Contains the genetic information. Each strand in the DNA is made of four bases: A, C, T and G. The two strands pair with each other so that A pairs with T, and C with G. Thus DNA is made of a chain of base-pairs and can be represented by a string of four types of letters.

Dorsal - Side of an animal closer to its back.

Drosophila - Fruit fly. A model organism commonly used for biological research.

Edge - A link between two nodes in a network. Edges describe interactions between the component described by the nodes. Edges in most networks have a specific direction. Mutual edges are edges that link nodes in both directions. See transcription network for an example.

Endocytosis - Uptake of material into a cell.

Enzyme - A protein that facilitates a biochemical reaction. The enzyme catalyzes the reaction and does not itself become part of the end product.

ER (Erdos-Renyi) random networks - An ensemble of random networks with a given number of nodes, N, and edges, E. The edges are placed randomly between the nodes. This model can be used for comparison to real networks. A more stringent random model is the degree-preserving random network.

***E. coli* (*Escherichia Coli*)** - A rod-shaped bacterium normally found in the colon of humans and other mammals. It is widely studied as a model organism.

Eukaryotic cells and organisms - Organisms made of cells with a nucleus. Includes all forms of life except for viruses and bacteria (prokaryotes). Yeast is a single-celled eukaryotic organism.

Exponential phase - A phase of bacterial (possible also for other cell types) growth in which cells double with a constant cell generation time, resulting in exponentially increasing cell numbers. This occurs in a test tube when there are so few cells that nutrients are not depleted from the medium, and waste products do not accumulate to high levels. See also stationary phase.

Extinction - The process by which an allele (or a species) disappears from a population.

Feedback - A process whereby some proportion or function of the output signal of a system is passed (fed back) to the input.

Feedback inhibition - A common control mechanism in metabolic networks, in which a product inhibits the first enzyme in the pathway that produces that product.

Feed-forward loop (FFL) - A pattern with three nodes, X, Y and Z, in which X has a directed edge to Y and Z, and Y has a directed edge to Z. The FFL is a network motif in many biological networks and can perform a variety of tasks (such as sign-sensitive

delay, sign-sensitive acceleration, and pulse generation).

Fine-tuned property - A property of a biological circuit that depends sensitively on the biochemical parameters of the circuit (opposite to robust property).

First-order kinetics - Mathematical description of the rate of an enzymatic reaction in the limit where the substrate concentration is very low and is far from saturating the enzyme, such that the rate is equal to $(v/K)\,E\,S$, where v is the rate per enzyme, E is the enzyme concentration, K is the Michaelis constant, and S is the substrate concentration. See also Michaelis-Menten kinetics, zero-order kinetics.

Fixation - The process by which one allele reaches frequency 100% in the population.

Fitness - The growth advantage associated to specific genetic traits.

Flagellum (plural flagella) - A long filament whose rotation drives bacteria through a fluid medium. Rotated by the flagellar motor.

Functionalism - The strategy of understanding an organism's structural or behavioral features by attempting to establish their usefulness with respect to survival or reproductive success.

Gene - The functional unit of a chromosome, which directs the synthesis of one protein (or several alternate forms of a protein). The gene is transcribed into mRNA. which is then translated into the protein. The gene is preceded by a regulatory DNA region called the promoter that includes binding sites for transcription factors that regulate the rate of transcription.

Gene circuit - A term used here to mean a set of biomolecules that interact to perform a dynamical function. An example is a feed-forward loop.

Gene product - The protein encoded by a gene. Sometimes, the RNA transcribed from the gene, when the RNA has specific functions.

Generation time - Mean time for an organism to produce offspring.

Genetic code - The mapping between the 64 codons and the 20 amino acids. The genetic code is identical in nearly all organ-

isms.

Genetic drift - The statistical change over time of gene frequencies in a population due to random sampling effects in the formation of successive generations.

Genome - The total genetic information in a cell or organism.

Glucose - A simple sugar, a major source of energy in metabolism.

GFP (green fluorescent protein) - GFP was originally found in jellyfish. When irradiating the protein with some short wavelength light, it emits light at some specific longer wavelength. Many colors have now been developed. The GFP proteins in a single cell can then be seen in a microscope. The fluorescent property of GFP is preserved in virtually any organism that it is expressed in, including bacteria. It has revolutionised live biological imaging in two broad classes of experiments: (i) By subjecting its expression to a promoter region that one wants to monitor, one can measure ongoing activity of the selected promoters (this construction is called 'reporter'); (ii) it can be genetically linked ('fused') to other proteins (by a covalent bond along the peptide backbone, then allowing to track movements or localisations of this protein inside the living cell.

In the best cases, this linking with GFP does not influence the properties of the particular protein, and does not perturb the cell too much. The main worries with these experimental approaches are (a) "phototoxicity", whereby the photon flux, and the byproducts of the fluorescence chemistry, affect the cell; (b) the possible metabolic cost of expressing these extra proteins; (c) in experiments where dynamics is important, to pay attention to the time required for transcription+translation+maturation (maturation may be from a few minutes in some variants, up to some hours).

Haploid Individual - An individual carrying only one copy of its genome.

Heterozygous - Diploid individual whose two genome copies differ at a specific site.

Homozygous- Diploid individual whose two genome copies are identical at a specific site.

Hill coefficient - See also 'Michaelis-Menten kinetics'. The Hill coefficient in kinetics reactions is the power exponent on the concentration of reagent. Mechanistically, it corresponds to the

number of molecules that must act simultaneously in order to make a given reaction happen. The higher the Hill coefficient, the sharper the transition.

Histones - only in in eukaryotes, these are DNA binding proteins that regulate the condensation of DNA, i.e. determine the physical structure. The DNA makes two turns around each histone. Histones play a major role in gene silencing in eukaryotes, and a large fraction of transcription regulators in yeast, for example, is associated with histone modifications.

Homeostasis - The process by which the organism's substances and characteristics are maintained at their steady (optimal) level. Typically the result of a negative (stabilizing) regulative feedback.

Homologous - Similar by virtue of a common evolutionary origin. Homologous genes generally show similarity in their sequence.

Hormone - A chemical substance liberated by an endocrine gland that has effects on target cells in other organs.

Immune system - The system by which an organism protects itself from foreign proteins. In mammals there are an innate and an adaptive system. The innate system triggers inflammation and recruitment of further immune cells. In response to an infection, the white blood cells (adaptive system) can produce antibodies that recognize and attack invading microorganisms, and typically some memory of this remains in the organism.

Integral feedback - Feedback on a device in which the integral over time of the error (output minus the desired output) is negatively fed back into the input of the device. Integral feedback can lead to robust exact adaptation.

Kinase - An enzymatic protein that transfers a phosphate group (PO_4) from a phosphate donor to an acceptor amino acid in a substrate protein (an important example of 'post-transcriptional modification', i.e. the regulation mechanisms that a cell deploys on proteins, the final products of gene expression). Kinases have been classified after acceptor amino acids.

Lac operon - A group of three genes in *E. coli* that are adjacent on the chromosome and transcribed on the same mRNA. These genes are *lacZYA*, encoding for the metabolic enzyme LacZ which cleaves lactose into glucose and galactose; the permease (pump) LacY, which pumps lactose into the cells; and LacA, whose function is unknown. Lactose is not pumped into the cells if glucose (a better energy source) is present, a phenomenon called "inducer

exclusion". The *lac* operon is repressed by LacI and activated by CRP. LacI unbinds from the DNA and the system is induced in the presence of lactose (LacI binds a derivative of lactose called allolactose) or non-metabolizable analogs of lactose such as IPTG. As well as having a key importance in bacteria, this switch has been and continues to be a test-bed for quantitative work on understanding regulation of gene expression.

Lactose - A sugar utilized by *E. coli* as an energy and carbon source, using the *lac* genes expressed from the *lac* operon.

Ligand - A molecule that specifically binds the binding site of a receptor.

Mathematically controlled comparison - A comparison that is carried out with equivalence of as many internal and external parameters as possible between the alternative model mechanisms. Internal parameters include biochemical parameters, such as the lifetime of the proteins that make up the circuit and external parameters include desired output properties, such as steady-state levels.

Membrane - A structure consisting principally of lipid molecules that define the outer boundaries of a cell or organelle.

Membrane potential - The difference in electrical potential inside and outside of the cell expressed as voltage relative to the outside voltage. Membrane potential is maintained by protein pumps that transport ions across the membrane at the expense of energy supplied by ATP.

Michaelis-Menten kinetics - A mathematical description of the rate of an enzymatic reaction as a function of the concentration of the substrate. The rate is equal to $v\,E\,S/(K\,+\,S)$, where v is the rate per enzyme, E is the enzyme concentration, S is the substrate concentration, and K is the Michaelis constant. When $S >> K$ one obtains zero-order kinetics (rate $= v\,E$), and when $S << K$ one obtains first-order kinetics (rate $= (v/K)\,E\,S$).

Model - One of the terms that causes the most confusion in science. Every field means something different by 'model'. If we want to specify to each other a physics view of a model (typically: a mathematical structure, where the ingredients correspond directly to a mechanism, usually corresponds to a solid understanding of one aspect isolated out of a more complex system, where the parameters have physical meaning and are a limited number, can be tested experimentally, has predicting value, can be somewhat extrapolated, ...) we sometimes say 'physical model'. Warning:

other words here like ‘mechanism’ and ‘solid understanding’ are also not used the same across scientific disciplines.

Model Organism - There are estimated to be 8.7 million species on Earth. They are very diverse but have evolved probably from one single ancestor organism. Fortunately they share many molecular parts, and many ways of functioning, across all life or at least across large groups. ‘Model organisms’ are those that either by chance, or by virtue of being easier to source or maintain in labs, have become the ‘standard’ organisms for particular biological questions. Over decades of research, in seeking ‘general’ or ‘fundamental’ biological principles, biologists have developed extremely detailed knowledge on a very limited number of species: certain viruses, the bacterium E.coli and a few others, two types of yeast, the fruit fly drosophila, the nematode worm C.elegans, the frog Xenopus, the plant Arabidopsis, the ‘zebrafish’ and the mouse (listed in rough order of complexity of questions, space/time scale, and also complexity of experiments). The simplest species that displays the process to be studied is usually the most appropriate species to be studied. The other 8.7 million species are studied by a very small number of labs and only for particular reasons, usually different from seeking general principles of life (e.g. because of their importance in particular ecological systems, or as infectious pathogens, or as food or energy sources for humans, or because they display some amazing unique property that has attracted human curiosity). Warning: evolution is a driver towards diversity, and has had a very long time to act; among those 8.7 million species there are a lot of unique behaviours and processes that seem to mock any quest for ’generality’ - getting the right approach and the right questions is very important in biological physics.

Modularity - A property of a system which can be separated into nearly independent sub-systems.

Morphogen - A molecule (protein) that determines spatial patterns. Morphogens bind specific receptors to trigger signal transduction pathways within the cells to be patterned. The signaling leads the cells to assume different cell fates according to the morphogen level.

Morphology - Physical shape and structure.

mRNA - A macromolecule made of a sequence of four types of bases: A, C, G and U. Transcription is the process by which an RNA-polymerase enzyme produces an mRNA molecule that corresponds to the base sequence on the DNA (where DNA T is mapped to RNA U). The mRNA is read by ribosomes, which pro-

duce a protein according to the mRNA sequence.

Mutation - A heritable change in the base-pair sequence of the chromosome.

Network motif - A pattern of interactions that recurs in a network in many contexts. Network motifs can be detected as patterns that occur much more often than in randomized networks.

Neuron (nerve cell) - Cell specialized to receive, transmit and conduct signals in the nervous system.

Nucleus - A structure enclosed by a membrane found in eukaryotic cells (not in bacteria) that contains the chromosomes.

Nucleoid - region within the cell of a prokaryote that contains all or most of the genetic material, and the proteins associated to that. Proteins that shape the chromosome in bacteria are called Nucleoid-Associated Proteins (NAP).

Nucleosome An important structural unit of the chromosome in eukaryotes, made up of 146 bp of DNA wrapped 1.75 times around an octamer of histone proteins.

Operon - Only in prokaryotes. A group of contiguous genes *transcribed* on the same mRNA, plus the regulatory elements that control their transcription. Each gene is separately *translated*. Operons are found only in prokaryotes.

Peptide - A chain of amino acids joined together by peptide bonds. Proteins are long peptides.

Phage - Also known as a bacteriophage, this is a virus that attacks a bacteria.

Plasmid - A piece of double-stranded DNA that encodes some proteins (which are expressed in the host of the plasmid) and replicates alongside the host chromosomes. It may be viewed as an extrachromosomal DNA element, and as such it can be transmitted from host to host. Plasmids are, for example, carriers of antibiotic resistance, and when transmitted between bacteria thereby help these to share survival strategies. Plasmids often occur in multiple-copies in a given organism, and can thus be used to greatly overproduce certain proteins. This is often used for industrial mass production of proteins.

Point mutation - A change of a single letter (base-pair) in the

DNA.

Poisson distribution - A distribution that characterizes a random process such as the number of heads in a coin-toss experiment, with many tosses, N, and a small probability for heads, $p << 1$. The mean number of heads is $m = pN$. The variance in a Poisson process is equal to the mean. $\sigma^2 = m$ and hence the standard deviation is the square root of the mean, $\sigma = \sqrt{m}$.

Prokaryotes - Single-celled organisms without a membrane around the nucleus. It is estimated that there are $(4 - 6) \times 10^{30}$ prokaryotes on Earth. The number of prokaryote divisions per year is $\approx 1.7 \times 10^{30}$. Prokaryotes are estimated to contain about the same amount of carbon as all plants on Earth (5×10^{14} kg). Some 5000 species have been described, but there are estimated to be more than 10^6 species.

Promoter - A regulatory region of DNA that controls the transcription rate of a gene. The promoter contains a binding site for RNA polymerase (RNAp), the enzyme that transcribes the gene to produce mRNA. Each promoter also usually contains binding sites for transcription factor proteins. The transcription factors, when bound, affect the probability that RNAp will initiate transcription of an mRNA.

Protease - An enzyme that degrades proteins. Proteins are often targeted for degradation in biologically regulated ways. For example, many eukaryotic proteins are targeted for degradation in the proteosome by enzymes that attach a chain of ubiquitin molecules to the target protein. Different proteins can have different degradation rates.

Protein - A long chain of amino acids (a polymer, on the order of tens to hundreds of amino acids) that can serve in a structural capacity or as an enzyme. Each protein is encoded by a gene. Proteins are produced in ribosomes, based on information encoded on an mRNA that is transcribed from the gene.

Receptor - A protein molecule, usually situated in the membrane of the cell (but sometimes in the cytoplasm of the cell) that is sensitive to a particular chemical. When the appropriate chemical (the ligand) binds to the binding site of the receptor, signal transduction cascades are triggered within the cell.

Repression threshold - Concentration of active repressor needed for half-maximal repression of a gene.

Repressor - A transcription factor that decreases the rate of

transcription when it binds a specific site in the promoter of a gene.

Ribosome - A structure in the cytoplasm made of about 100 proteins and special RNA molecules that serves as the site of production of proteins translated from mRNA. In the ribosome, amino acids are assembled to form the protein chain according to an order specified by the codons on the mRNA. The amino acids are brought into the ribosome by tRNA molecules, which read the mRNA codons. Each tRNA is released when its amino acid is linked to the translated protein chain.

RNA Polymerase (RNAp) - A complex of several proteins that form an enzyme that transcribes DNA into RNA. There is also DNA polymerase, the complex used to make copies of DNA before cell division.

Robust Property - Property X is robust with respect to parameter Y, if X is insensitive to changes in parameter Y.

Sensory transcription networks - Transcription networks that respond to environmental and internal signals such as nutrients and stresses, and lead to changes in gene expression. These networks need to function rapidly, usually within less than a cell generation time, and usual make reversible decisions. They stand in contrast to developmental transcription networks.

Stationary phase - A state in which cells cease to divide and grow, that occurs when growth conditions are unfavorable, such as when the bacteria run out of an essential nutrient. See also exponential phase.

Stem cell - A cell type found in the adult form of many animals and plants. It is a cell that has retained the property to differentiate into multiple other types of cell. Obviously the fertilised egg is a stem cell that makes all other cells, but breakthroughs in medicine like regenerative treatments can come from understanding adult stem cells.

stop codons - Triplets (UAG, UGA, and UAA) of nucleotides in RNA that signal a ribosome to stop translating an mRNA and release the translated polypeptide.

Terminator - Stop sign for transcription at the DNA. In *E. coli* it is typically a DNA sequence that codes for an mRNA sequence that forms a short hair-pin structure plus a sequence of subsequent Us. For example, the RNA sequence CCCGCCUAAUGAGCGGGCUUUUUUUU terminates RNAp elongation in *E. coli*.

Transcription - The process of copying the DNA template to an RNA.

Transcription factor - A protein that regulates the transcription rate of specific target genes. Transcription factors usually have two molecular states, active and inactive. They transit between these states on a rapid timescale (e.g. microseconds). When active, the transcription factor binds specific sites on the DNA to affect the rate of transcription initiation of target genes. Also called transcriptional regulator. See activator, repressor.

Transcription network - The set of transcription interactions in a cell. The network is made of nodes linked by directed edges. Each node represents a gene (or, in bacteria, an operon), Each edge is a transcriptional interaction. $X \to Y$ means that the protein encoded by gene X is a transcription factor that transcriptionally regulates gene Y.

Translation - The process of copying RNA to protein. It is done in the ribosome with the help of tRNA.

tRNA - This is transfer RNA - small RNA molecules that are recruited to match the triplet codons on the mRNA with the corresponding amino acid. This matching takes place inside the ribosome. For each amino acid there is at least one tRNA.

XOR gate (exclusive OR) - A logic function of two inputs that outputs a one if either, but not both, inputs is equal to one.

Yeast - A single-celled eukaryote, a unicellular fungus. There are many types but two are fundamental to humanity: budding yeast (*Saccharomyces cerevisae*) is most commonly used in baking and brewing, and fission yeast (*Schizosaccharomyces pombe*). Both are also common research model organisms.

Zero-order kinetics - Mathematical description of the rate of an enzymatic reaction in the limit where the substrate concentration is saturating, such that the rate is equal to $v\,E$ where v is the rate per enzyme, and E is the enzyme concentration. See also Michaelis-Menten kinetics, and first order kinetics.

Useful Estimation Data

	Quantity of interest	Symbol	Rule of thumb
E. coli			
	Cell volume	$V_{E.\,coli}$	$\approx 1\,\mu m^3$
	Cell mass	$m_{E.\,coli}$	≈ 1 pg
	Cell cycle time	$t_{E.\,coli}$	≈ 3000 s
	Cell surface area	$A_{E.\,coli}$	$\approx 6\,\mu m^2$
	Macromolecule concentration in cytoplasm	$c_{E.\,coli}^{macromol}$	≈ 300 mg/mL
	Genome length	$N_{bp}^{E.\,coli}$	$\approx 5 \times 10^6$ bp
	Swimming speed	$v_{E.\,coli}$	$\approx 20\,\mu m/s$
Yeast			
	Volume of cell	V_{yeast}	$\approx 60\,\mu m^3$
	Mass of cell	m_{yeast}	≈ 60 pg
	Diameter of cell	d_{yeast}	$\approx 5\,\mu m$
	Cell cycle time	t_{yeast}	≈ 200 min
	Genome length	N_{bp}^{yeast}	$\approx 10^7$ bp
Organelles			
	Diameter of nucleus	$d_{nucleus}$	$\approx 5\,\mu m$
	Length of mitochondrion	l_{mito}	$\approx 2\,\mu m$
	Diameter of transport vesicles	$d_{vesicle}$	≈ 50 nm
Water			
	Volume of molecule	V_{H_2O}	$\approx 10^{-2}\,nm^3$
	Density of water	ρ	$1\,g/cm^3$
	Viscosity of water	η	≈ 1 centipoise (10^{-2} g/(cm s))
	Hydrophobic embedding energy	$\approx E_{hydr}$	2500 cal/(mol nm^2)
DNA			
	Length per base pair	l_{bp}	$\approx 1/3$ nm
	Volume per base pair	V_{bp}	$\approx 1\,nm^3$
	Charge density	λ_{DNA}	$2\,e/0.34$ nm
	Persistence length	ξ_p	50 nm
Amino acids and proteins			
	Radius of "average" protein	$r_{protein}$	≈ 2 nm
	Volume of "average" protein	$V_{protein}$	$\approx 25\,nm^3$
	Mass of "average" amino acid	M_{aa}	≈ 100 Da
	Mass of "average" protein	$M_{protein}$	$\approx 30{,}000$ Da
	Protein concentration in cytoplasm	$c_{protein}$	≈ 150 mg/mL
	Characteristic force of protein motor	F_{motor}	≈ 5 pN
	Characteristic speed of protein motor	v_{motor}	≈ 200 nm/s
	Diffusion constant of "average" protein in cytoplasm	$D_{protein}$	$\approx 10\,\mu m^2/s$
Lipid bilayers			
	Thickness of lipid bilayer	d	≈ 5 nm
	Area per molecule	A_{lipid}	$\approx \frac{1}{2}\,nm^2$
	Mass of lipid molecule	m_{lipid}	≈ 800 Da

Bibliography

Phillips, R., Kondev, J., Theriot, J., Garcia, H., and Orme, N. (2013). *Physical Biology of the Cell, 2nd Ed.* Garland Science, London.

Sneppen, K. and Zocchi, G. (2005). *Physics in Molecular Biology.* Cambridge University Press, Cambridge.

Strogatz, S. S. (2014). *Nonlinear Dynamics and Chaos.* Westview Press, Boulder.