Biological Physics - NOTES

Pietro Cicuta and Eileen Nugent

Experimental and Theoretical Physics
Part III
Michaelmas 2015

Notes version: v0.01
Release name: animated arachae

Preface

How to use these notes:

Where derivations are written out extensively here, they will probably not be reproduced in class, and vice versa. You will be expected to have understood all of these, and to be able to reproduce these results and variations that use the same methods. Derivations obtained in the question sheet exercises are also part of the course, and worked out solutions will be made available towards the end of the course.

Dos:

- Use the notes to follow progress through the course material. The structure of these notes is almost the same as the lectures.
- Integrate the lecture overheads and the notes material yourself. There is examinable material that might appear in one place only.
- Follow suggestions and think about the questions in the notes. These are distributed through the text to help you spot if you are understanding the material.

Don'ts:

- Expect to study only from these notes. You will need the other main references and attendance to lectures. Most of all you will need to understand how to use the material and methods presented, rather than memorising information.
- Expect these notes to be error free. They will contain a higher density of errors than a typical book! e-mail us if you think something is wrong or unclear, and the notes will improve.
- Expect these notes to be even in the level of presentation. Some paragraphs are minimal, and some section labels are only place holders for material that will be covered in class. Instead, use these notes to guide you through the books.

Course aims and structure

Course aims

Possible questions

Is Biological Physics well defined? Is there physics in biology?
A pragmatic answer is along the lines of what the late Sir Sam Edwards (former Cavendish professor, and founder 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. Science is one, and you need to find areas where good progress can be made and where it is worth putting effort. In this sense, if a physicist sees an opportunity to contribute in a unique way to biology, this can be pursued. Is is pretty obvious that "biology" is itself very broad (consider how many aspects there are to living systems, reflected in a mosaic of departments and institutes that is not unique to Cambridge), and important questions can be posed at many length-scales from the molecular through cellular, organ tissue, up to populations and ecology. Not to mention medicine. One also can pose questions on dynamics and evolution, and again relevant time-scales span many orders of magnitude from molecular binding processes through organism development and maintenance of tissues, up to mechanisms of evolution. So clearly there will be many ways to apply physics, and many different types of models that can be deployed or invented based on physical intuition. 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 into reductionist approaches when dealing with, say, materials, we don't really have a unified model that we expect to give quantitative predictions at the same time for all the material's behaviour, say x-ray diffraction, melting temperatures, properties of density, conductivity or elasticity... We are used and ready to accept, in this context, the idea that we can abstract the important elements that underlie a certain phenomenon. 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 have indeed captured a relevant mechanism, and hence are able to show how modifying the '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 'understanding', as opposed to other quantitative approaches more typical of engineers or mathematicians.

Is there physics in biology?
This is a corollary to the statements above. But yes, particularly at the present time recent 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 to applying and developing physical models, in exactly the same spirit as in studying condensed matter systems, or other complex systems (nonlinear optics, 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, on the experimental (and in some cases computational) frontier, developing experiments and techniques.

What can be achieved in a 24 hour course? The main aims of this course are:
(a) through good examples, and with a storyline as coherent as possible, mostly at the cell and molecular level, show how physics (particularly stat mech, soft matter, networks and nonlinear dynamics approaches) has been developed and applied in recent years to address both existing challenges, and even to define new categories in biological systems.
(b) through (a), provide 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, the term is usually meant to emphasise the molecular aspect (e.g. protein folding, biochemical interactions); we touch only some aspects. Another community (medical) defines biophysics as biomechanics and issues to do with circulation, pressure, etc - this course has none of this 'physiological' 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 does not change fast, but we do not cover

that here). Closer to the spirit of this course there would also be a lot of topics that we cannot cover due to lack of time and personal expertise, but which could lend themselves to the same type of thinking and modeling presented here: worth mentioning are embryonic development (and tissue homeostasis), evolution (which is possible in the lab, exploring influence of stresses, species competition, etc), ecology (also amenable to lab experiments, with suitable choice 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 was to provide a coherent 'story', covering various concepts and examples that we think are useful. Whilst not wanting to overburden with fact collections, a minimum of context is necessary and will be useful in any future interaction with the world of Life Sciences.

The course is structured into six modules (A-F). Modules are 3 or 4 lectures, and have a single lecturer (Dr P.Cicuta or Dr E.Nugent). Two 'guest lecturers', quite prominent biologists, will give 3 lectures (details non examinable) on their pioneering discoveries of quantitative aspects in cell biology, and how they pursued physical modeling. Those guest lecturers will also explain the experimental approach in more detail than what is possible 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: An overview of quantitative cell biology, and a primer of concepts. 3 lectures

Physical biology of the cell - information processing 'central dogma'; Life from a Physics perspective; The stuff of life; Model building in biology; How a cell adjusts to different growth rates; 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: Statistical Physics of Living systems. 3 lectures

Energy and the life of cells; Thermal and Statistical Physics of living systems; Chemical Forces; State variable descriptions of macromolecules; Two-state systems: phosphorylation, ion channels, cooperative binding; Diffusion in living systems.

Module C: Protein production and regulation of gene expression. 3 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 D: Circuits and dynamical systems. 4 lectures

Properties of dynamical systems, and intro to methods; Feedback circuits; Genetic circuits with switch and oscillating properties.

Module E: Molecular Motors. 4 lectures

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

Module F: Sensing and Neural Biophysics. 3 lectures

The electrical status of cells and their membranes; The Hodkin-Huxley Model for the generation of action potentials; Sensing: vision, hearing; Information processing in neurons.

Conclusion: Outlook beyond cell biological physics - 1 lecture

Introduction to quantitative cell biology

1

The introductory material of the first couple of lectures can be found on the overheads. Presented 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).

1.1 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 genes 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.

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).

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).

Distributions Some common ones:
exponential

$$p(t) \sim \exp(-t/t).$$

If t is a waiting time this is the distribution for a random uncor-

related 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

$$p(x) \sim \exp(-x^{\alpha}).$$

ParetoLevi
Obtained from the sum of numbers, each drawn from a distribution $\propto x^{-\alpha}$. A ParetoLevi distribution has a typical behavior like a Gaussian, but its tail is completely dominated by the single largest event. Thus a ParetoLevi 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 (ErdosRenyi) 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 ing random network.

***E. coli* (*Escherichia Coli*)** - A rodshaped 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.

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.

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 organisms.

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).

Hill coefficient - The number of molecules that must act simultaneously in order to make a given reaction. 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 allo-lactose) or nonmetabolizable 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 $= (v/K)\,E\,S$).

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 RNApolymerase 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 produce 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 (basepair) 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 $(46) \times 10^{30}$ 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.

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. $X \rightarrow Y$ 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 two types: budding yeast (*Saccharomyces cerevisae*), 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.

1.2 Concepts in networks

full credit: 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) loc-

alization/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 $i = 1, 2, , N$ 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 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

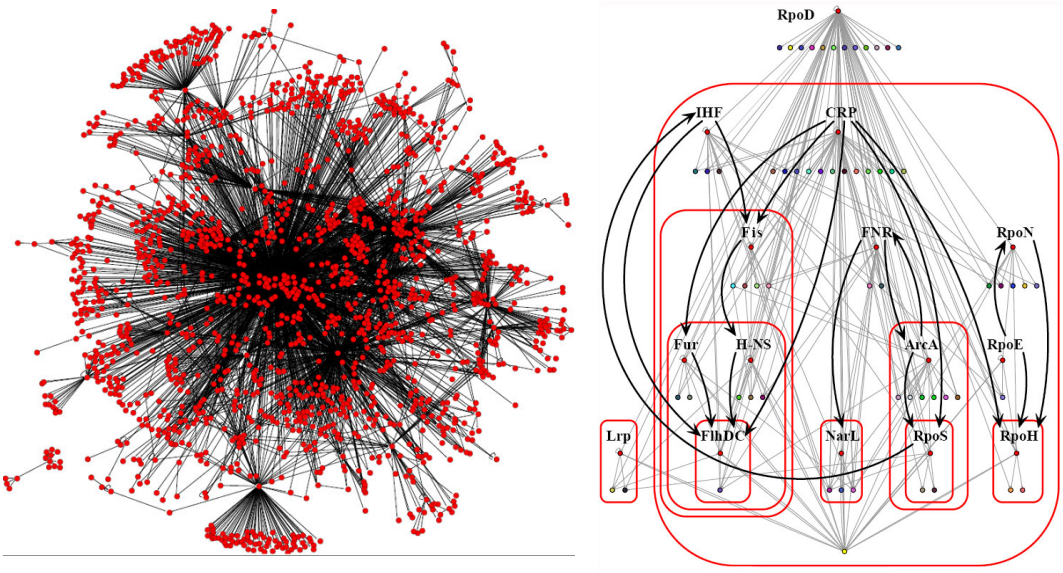


Fig. 1.1 Representations of data from RegulonDB, the database of E. coli regulation data (Salgado et al. 2006). On the left, the (known parts of) the *E. coli* transcriptional regulatory network. In this graphical representation, nodes are genes, and edges represent regulatory interactions. There is 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.

few hundred neurons in the roundworm *C. elegans*, and other examples. 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

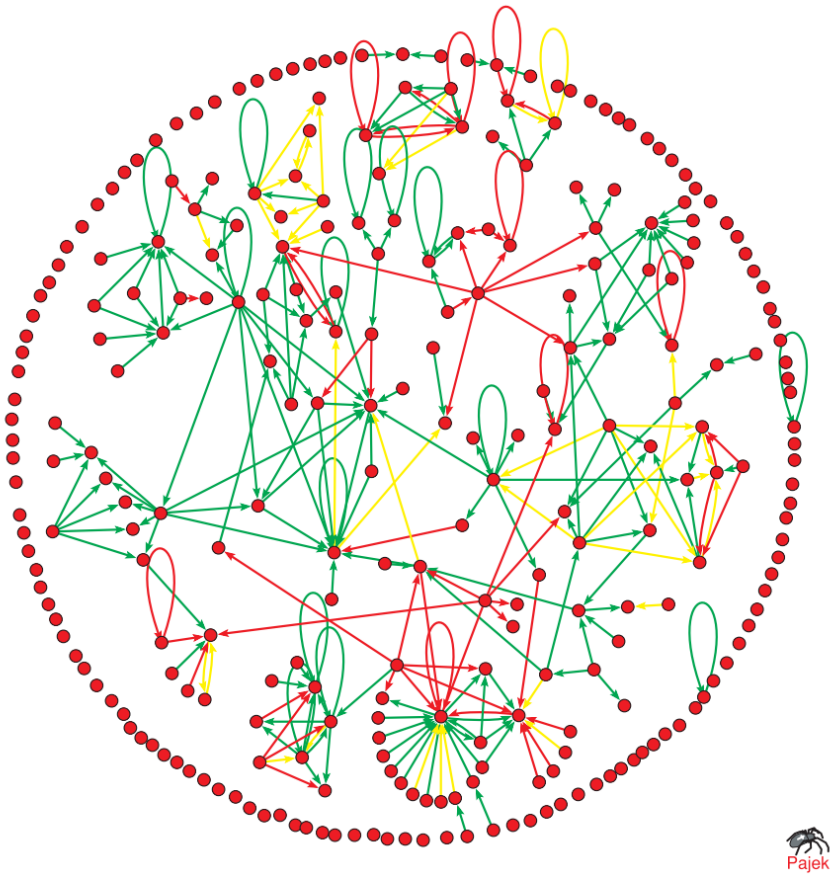


Fig. 1.2 Networks of transcriptional regulatory proteins in yeast. All proteins that are known to regulate at least one other protein are shown. Arrows indicate the direction of control, which may be either positive or negative. Functionally the network is roughly divided into an upper half that regulate 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.

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

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

1.2.1 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 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.

1.2.2 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

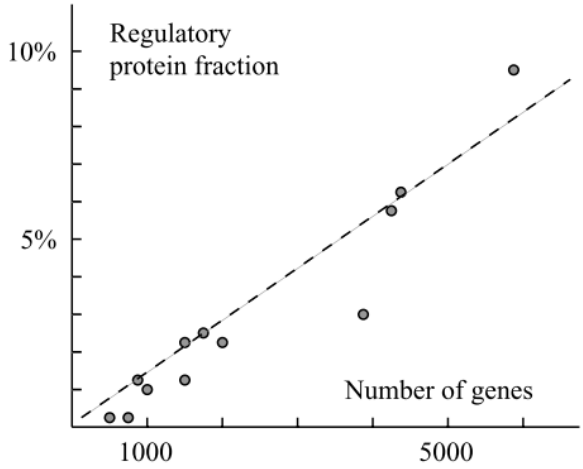


Fig. 1.3 Fraction of proteins that regulate other proteins, as a function of size of the organisms' gene pool. These data are for prokaryotes: smallest genome is *M. Genitalium* (480 genes); the largest genome is *P. Aeruginosa* (5570 genes). The linear relation demonstrates that each added gene should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

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 proteinprotein binding networks, and exponent $\gamma \simeq 1.5 \pm 0.5$ for "out-degree" distribution of transcription 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.

Mechanical and Chemical equilibrium inside the living cell. Entropy.

2

Energy and the life of cells
Thermodynamics of living systems
Biological swimmers as minimisers
Quantitative rules of metabolism (Hwa)
Diffusion and transport
Law of mass action
Ligand - receptor Chemical dynamics
Two state systems: from ion channels to cooperative binding
Macromolecules with multiple states
State variable description of binding

Physics of regulation: Reactions and Stat Mech of promoters

3

[1] full credit for this section: Dr Rosalind Allen, Univ. Edinburgh, through IoP Biological Physics online teaching material

3.1 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!)

3.1.1 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

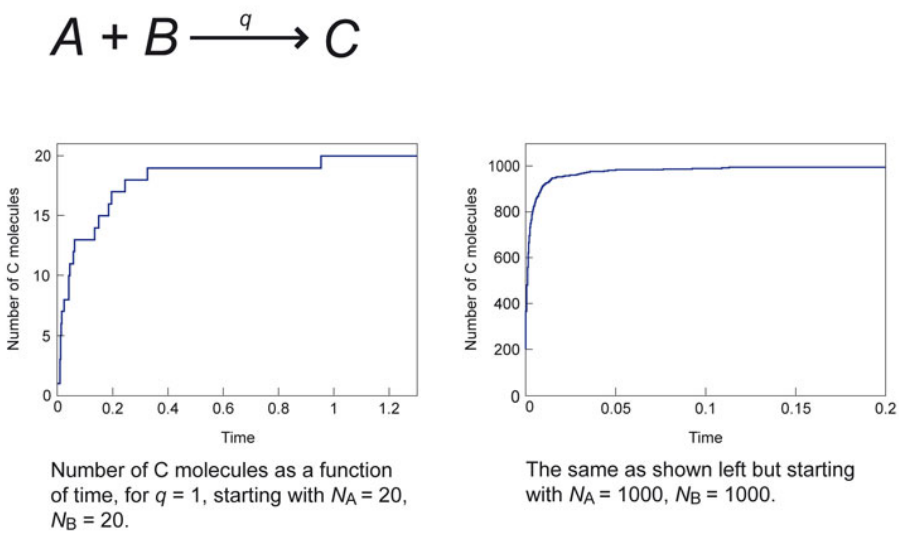


Fig. 3.1 Simulations of simple chemical reaction.

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 3.1 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 $c_C(0) = 0$.

3.1.2 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

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 $c_{\rm P}$ of P molecules:

$$\frac{dc_P(t)}{dt} = k - \mu c_P(t). \tag{3.1}$$

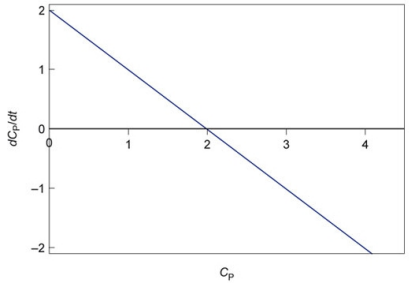


Fig. 3.2 Rate of change of protein concentration.

We can tell a lot about the system without actually solving this equation. Figure 3.2 shows the rate of change, $dc_{\rm P}/dt$, plotted for different concentrations of protein, $c_{\rm P}$, for parameter values $k = 2$ and $\mu = 1$. When the concentration of protein is small ($c_{\rm P} < k/\mu$), the rate is positive. This means that there will be net protein production ($c_{\rm P}$ will increase). However, when the concentration of protein is large ($c_{\rm P} > k/\mu$), $dc_{\rm P}/dt$ is negative. This means there will be a net loss of protein. We can also see that for $c_{\rm P} = k/\mu$, $dc_{\rm P}/dt$ 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, $c_{\rm P}^{(ss)}$.

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 $c_{\rm P}^{(ss)}$ 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

$$\mu = \frac{\ln 2}{\tau}. \tag{3.2}$$

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, $c_{\rm P}(t)$. 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,

$$c_{\rm P}(t) = \frac{k}{\mu}(1 - {\rm e}^{-\mu t}) + c_{\rm P}(0){\rm e}^{-\mu t}. \tag{3.3}$$

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: $c_{\rm P}(0) = 0$. 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

$$c_{\mathrm{P}}(t) \;=\; \frac{k}{\mu}(1 - \mathrm{e}^{-\mu t}). \tag{3.4}$$

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

$$t_{rise} \;=\; -\frac{1}{\mu} \ln \left[1 - \frac{\mu c_{\mathrm{P}}^{(ss)}}{2k} \right]. \tag{3.5}$$

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

$$t_{rise} \;=\; \ln(2)/\mu. \tag{3.6}$$

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.

3.1.3 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:

$$p_{bound} \;=\; \frac{\left(\frac{c}{c_0}\right) \exp -\beta\Delta\epsilon}{1 \;+\; \left(\frac{c}{c_0}\right) \exp -\beta\Delta\epsilon}, \tag{3.7}$$

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

$$K_d \;=\; c_0 \mathrm{e}^{\beta\Delta\epsilon}. \tag{3.8}$$

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

$$p_{bound} = \frac{\left(\frac{c_P}{K_d}\right)}{1 + \left(\frac{c_P}{K_d}\right)}, \tag{3.9}$$

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

$$p_{unbound} = 1 - p_{bound} = \frac{1}{1 + \left(\frac{c_P}{K_d}\right)}. \tag{3.10}$$

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:

$$\frac{dc_P(t)}{dt} = kp_{unbound} - \mu c_P = \frac{k}{1 + \left(\frac{c_P}{K_d}\right)} - \mu c_P. \tag{3.11}$$

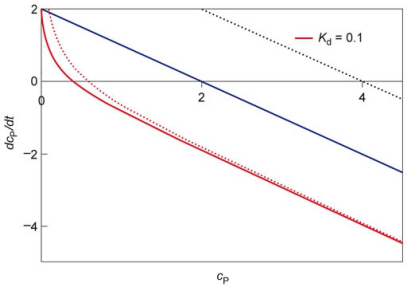


Fig. 3.3 Rate of change of protein concentration with negative autoregulation. The solid lines are for $k = 2$ and $m = 1$, and the dotted lines are for $k = 4$ and $m = 1$. The blue lines show the result without negative feedback (for the same k and m).

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 3.3 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 3.3 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:

$$\frac{dc_{\mathrm{P}}(t)}{dt} = \frac{k}{1 + \left(\frac{c_{\mathrm{P}}}{K_d}\right)} - \mu c_{\mathrm{P}} = 0, \tag{3.12}$$

obtaining

$$c_{\mathrm{P}}^{(ss)} = \frac{K_d}{2}\left[-1 + \sqrt{1 + \frac{k}{\mu K_d}}\right]. \tag{3.13}$$

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

$$c_{\mathrm{P}}^{(ss)} = \frac{K_d}{2}\left[-1 + 2\sqrt{\frac{k}{\mu K_d}}\right] \simeq \sqrt{\frac{kK_d}{\mu}}. \tag{3.14}$$

Figure 3.4 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.

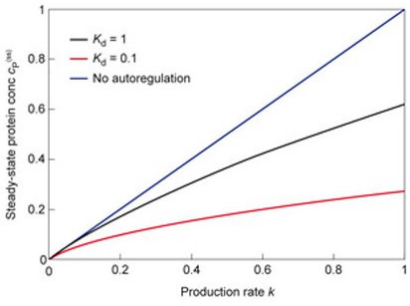


Fig. 3.4 Steady-state protein concentration for a negatively autoregulated gene.

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. 3.12, 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

$$t_{rise} = -\frac{1}{\mu} \ln \left[1 - \frac{\mu c_P^{(ss)}}{2k} \right], \tag{3.15}$$

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

$$t_{rise} = -\frac{1}{\mu} \ln \left[1 - \frac{\mu}{2k} \sqrt{\frac{kK_d}{\mu}} \right] = \frac{1}{\mu} \ln \left[\frac{2}{2 - \sqrt{\frac{kK_d}{\mu}}} \right]. \tag{3.16}$$

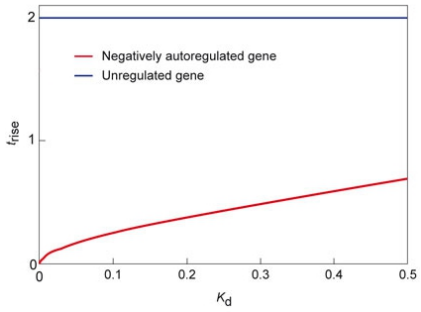


Fig. 3.5 Negatively autoregulation has strong effect on dynamics.

This result is plotted in Figure 3.5: 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 3.5 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.

3.1.4 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

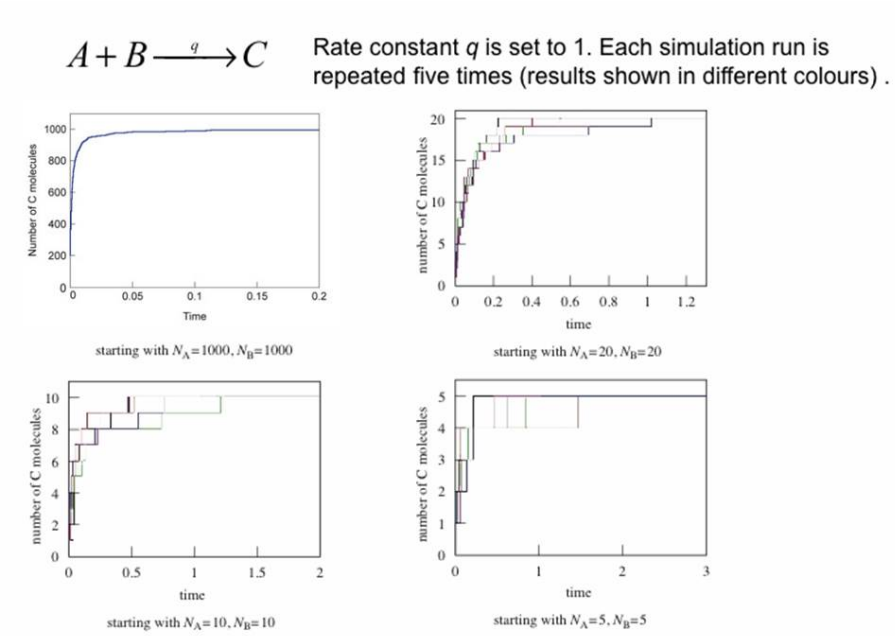


Fig. 3.6 Noise from low number of molecules can lead to different outcomes. Computer simulation results for reaction $A + B \to C$, starting with different numbers of A and B molecules.

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.

3.2 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 3.6 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 3.6, when $N = 2000$, $s = 0.022$, but when $N = 5$, $s = 0.44$. Although Figure 3.6 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".

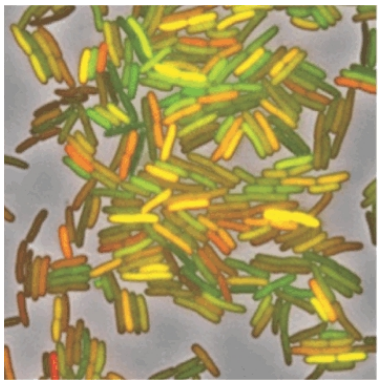


Fig. 3.7 Cells with identical genes in identical environments can behave differently. This can be explained in terms of biochemical noise. Cover image from Science Vol.297 issue 5584 (2002).

3.2.1 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 3.7 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 3.7 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 3.7 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

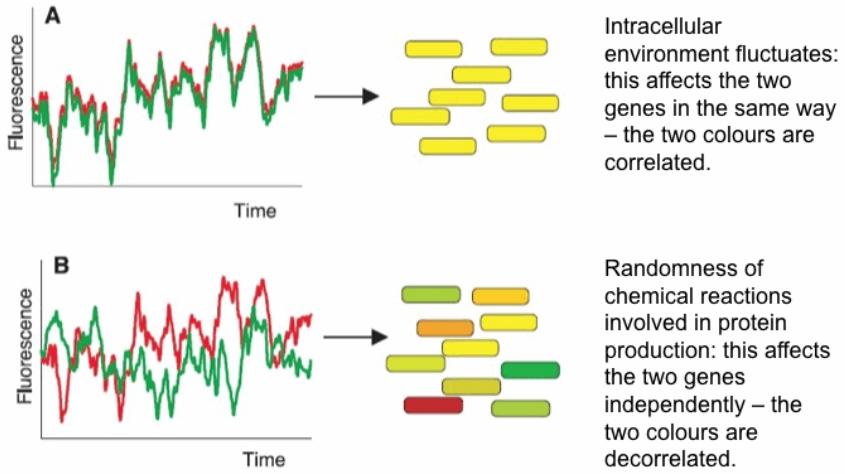


Fig. 3.8 Use of two "reporters" allows to distinguish intrinsic versus extrinsic noise. Protein levels vary because of fluctuations in the intracellular environment and of biochemical noise in transcription and translation.

should also have a lot of protein 2. However, if chemical reaction stochasticity is responsible for the differences in protein expression, we would not expect the levels of protein 1 and protein 2 in individual cells to be correlated. This is illustrated in Figure 3.8.

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 com-

position 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.

3.2.2 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

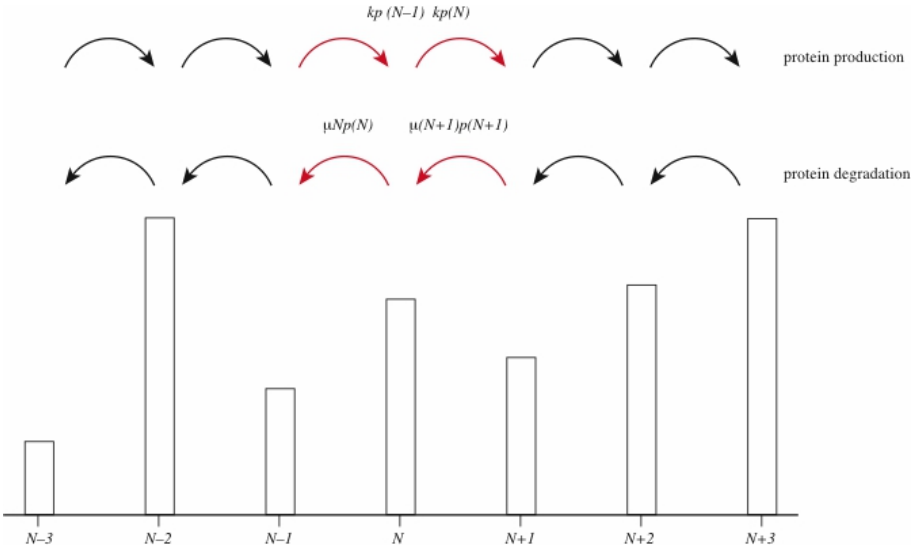


Fig. 3.9 Considering individual steps in a chemical reaction. Here, the vertical bars represent the probability of having a particular number of molecules and the arrows represent how the number of molecules is changed by the protein production and degradation reactions. In a very small time interval, $t \rightarrow t + dt$, 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 $(N+1)$.

(a) another one is produced (then it will have $N + 1$), or (b) one is degraded (then it will have $N - 1$), see Figure 3.9.

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

$$\frac{dp(N,t)}{dt} = kp(N-1) + \mu(N+1)p(N+1) - kp(N) - \mu N p(N) \tag{3.17}$$

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

$$\frac{dp(N,t)}{dt} = 0. \tag{3.18}$$

Solution to this is:

$$p(N) = \frac{1}{N!}\left(\frac{k}{\mu}\right)^N e^{-\frac{k}{\mu}}. \tag{3.19}$$

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 3.19 is the well known Poisson distribution.
Figure 3.10 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:

$$\frac{\sigma_N}{\langle N \rangle} = \sqrt{\frac{\mu}{k}} = \frac{1}{\sqrt{\langle N \rangle}}, \tag{3.20}$$

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.

3.2.3 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 3.11 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 3.11: we can see immediately that the distribution is broader

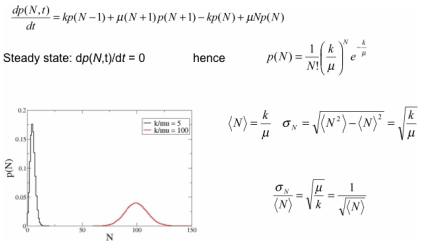


Fig. 3.10 Solution of chemical master equation for the simple one-step model of protein expression.

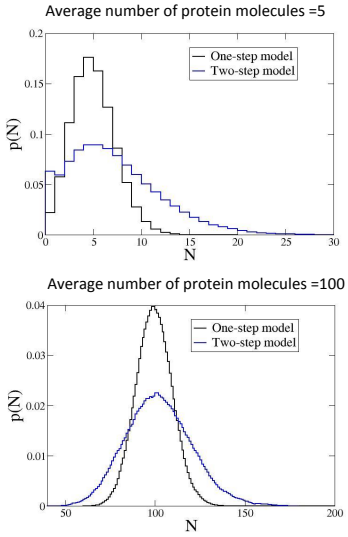


Fig. 3.11 Chemical master equation solutions for the one- and two-step models of protein expression. For the two-step process we assume that on average an mRNA produces five proteins, and we fix the transcription rate to get the same average number of proteins as in the one-step model.

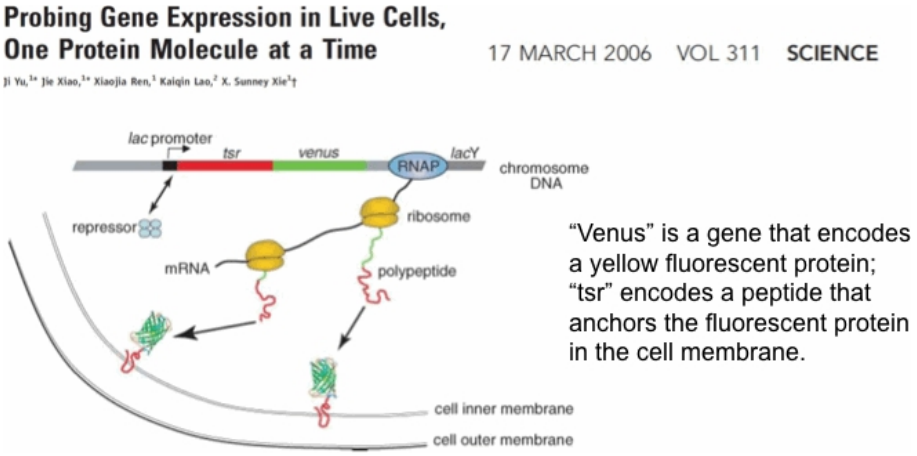


Fig. 3.12 Direct imaging of noise in gene expression. This experimental system was constructed by Yu *et al.* in 2006 to visualise in real time the production of a single protein molecule in a cell. From Yu *et al.* 2006, 'Probing gene expression in live cells, one protein molecule at a time', Science 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.

3.2.4 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 3.12). 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 3.3.4) 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.

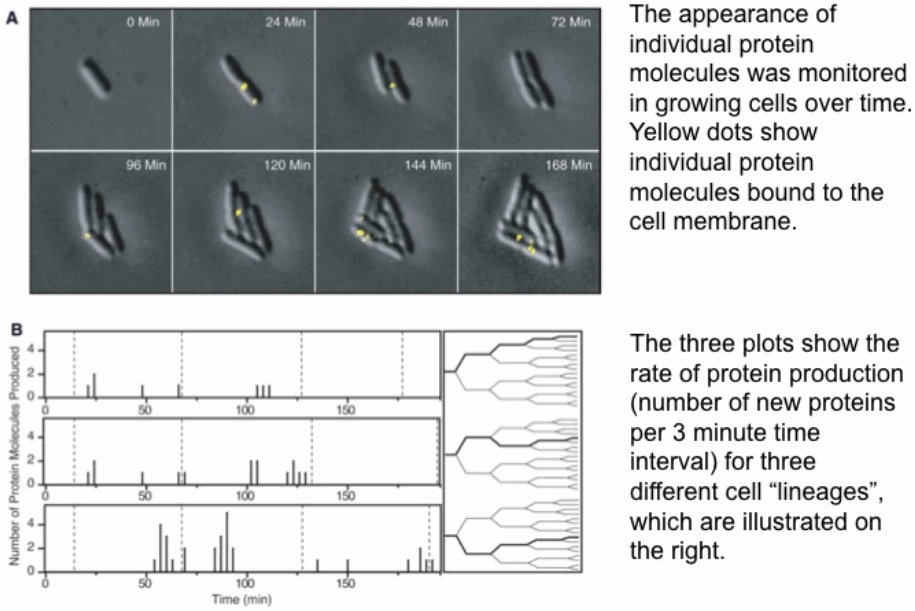


Fig. 3.13 Experiments can identify production of individual proteins. Some of Yu *et al.*'s results, showing the moment when individual protein molecules are produced in growing bacterial cells. From Yu *et al.* 2006.

Figure 3.13 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.

3.3 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).

3.3.1 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

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

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.

3.3.2 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 $\sim -10\,k_BT$.

3.3.3 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:

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

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:

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

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

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

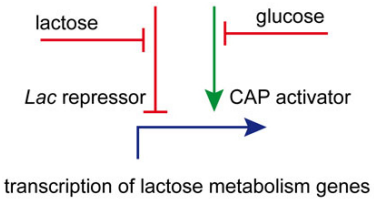


Fig. 3.14 Idealised (logic) lac network. A convenient way to illustrate the molecular interactions that make up the lac regulatory network. Here, positive molecular interactions (activation) are shown by arrows and negative molecular interactions (repression) are shown by "blocker" bars. The input to 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.

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

$$F_{reg}(A,R) = \left[1+(A/N_{NS})\mathrm{e}^{-\beta(\Delta\epsilon_{ad}+\epsilon_{ap})}\right] / \left[1+(A/N_{NS})\mathrm{e}^{-\beta\Delta\epsilon_{ad}}+(R/N_{NS})\mathrm{e}^{-\beta\Delta\epsilon_{rd}}+(A/N_{NS})(R/N_{NS})\mathrm{e}^{-\beta(\Delta\epsilon_{ad}+\Delta\epsilon_{pd})}\right].$$

3.3.4 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.

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.

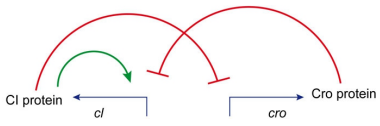


Fig. 3.15 Idealised (logic) and simplified diagram of the phage lambda genetic switch. The *cro* gene results in cell lysis; the *cI* gene promotes lysogeny.

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.

3.3.5 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 3.15. 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 3.15, 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.)

3.4 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.

3.4.1 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:

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

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

4

4.1 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, in turn, the one-variable system ("flow on the line"), the two-variable system ("flow on the plane") and the three-variable system ("3-D flow"). In general, an n-variable system requires n equations to represent it.

4.1.1 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,

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

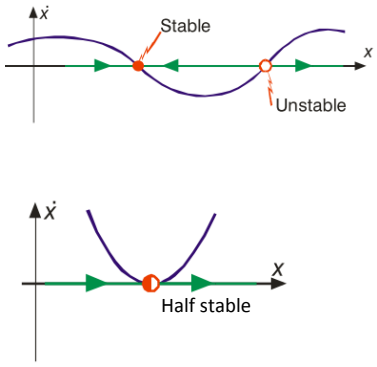


Fig. 4.1 Illustrations of the types of fixed points in 1-D systems. Note the notation: stable fixed points are denoted by filled circles; unstable fixed points by open circles, and half-stable points by half- filled circles, as shown in the examples. Note the notation: stable fixed points are denoted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

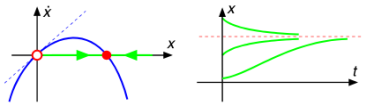


Fig. 4.2 Fixed points of the autocatalytic system.

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

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

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.

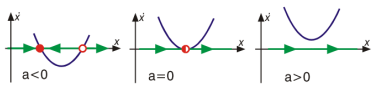


Fig. 4.3 In this example, a is control variable. Its value determines the stability of the system.

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.

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.

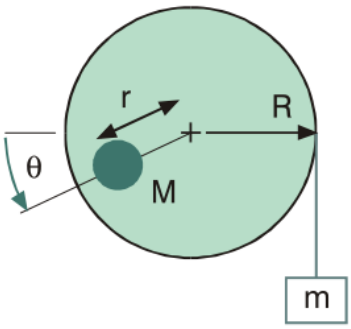
(2) **Asymmetric cubic potential: The saddle-node bifurcation.** The potential has the form

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

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 the figure.
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.
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 $\dot{x} = x^2 + a$. 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. [1]

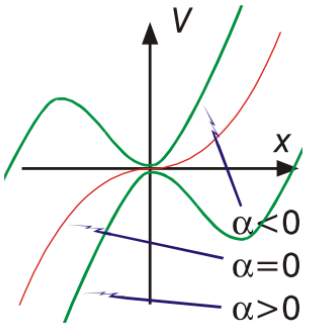
[1] **A mechanical example, the Weighted Pulley.** The gravitational potential is given by $V = mR\theta - Mr\sin\theta$, we can simplify notation $V = A\theta - B\sin\theta$. For small θ we can approximate this as $V \simeq (A-B)\theta + \frac{B}{6}\theta^3$. That has the same behavior as $V = \alpha\theta + \theta^3$, with $\alpha = 6(A-B)/B$. The system will thus be stable as long as $\alpha < 0$, i.e. $B > A$, i.e. $Mr > mR$.



(3) **Cubic potential with quadratic term: The transcritical bifurcation.** The potential this time includes a term in x 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 the figure. This is generally known as the transcritical bifurcation. One physical example of such a system is the laser.

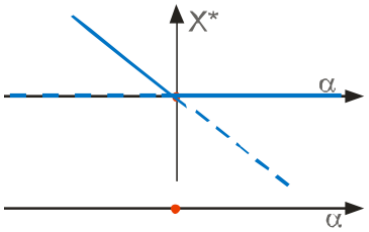


(4) **Symmetric quartic potential: The pitchfork bifurcation.** The potential is:

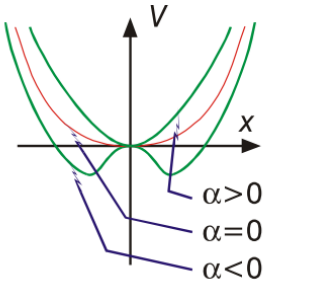
$$V(x) = x^4 + \alpha x^2.$$

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.

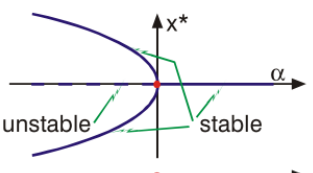


Plotted on the side here is the case of positive term on the 4^{th} power. In this case we refer to the Stable Symmetric Transition. It is also known as a Pitchfork Bifurcation (see Strogatz) from the shape of the bifurcation diagram, as shown at right. One example of this sort of potential is the Euler 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).



(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.

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.

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.

4.2 Gene regulation switches

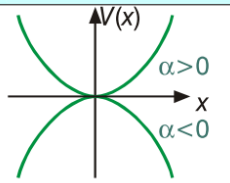
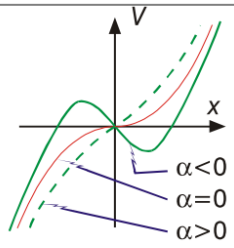
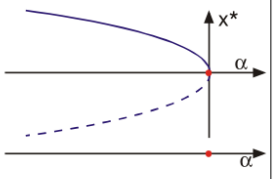
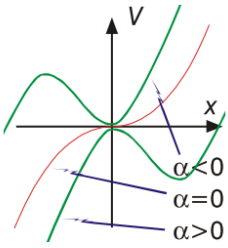
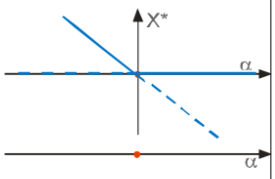
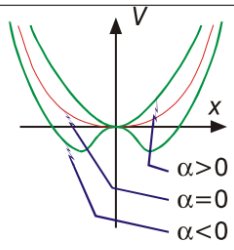
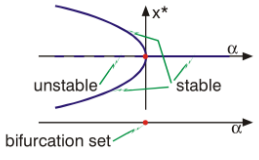
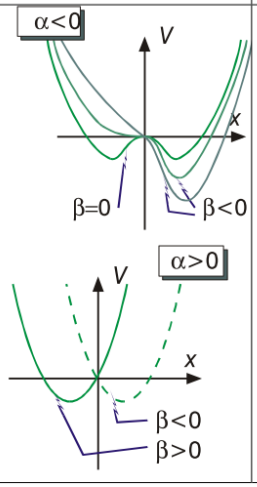
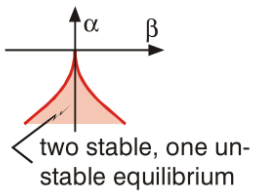
Following section 19.3.5 of (Phillips et al., 2013). Let's consider a synthetic switch that was created in *E.coli* as a simple construction to understand possibly more complex biological switches. The construction consists of two repressor proteins, whose transcription is mutually regulated. This arrangement 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:

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

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	$2\alpha x + 4x^3 + \beta$	$\alpha x^2 + x^4 + \beta x$		

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.

4.3 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:

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

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).

Life in crowded environments: Cytoskeleton and Cytoplasm

5

Depletion interaction
Hindered diffusion
Size-dependant effects
Cytoskeleton polymerisation dynamics
Dynamics of molecular motors
Cell mechanics

Membrane potential and neurons: Information transfer and processing

6

Membrane potential
Biological electricity and the Hodkin-Huxley model
Applications to sensing: vision, hearing, information processing

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.