Biological Physics - NOTES

Pietro Cicuta and Diana Fusco

Experimental and Theoretical Physics
Part III
Michaelmas 2021

Notes version: v0.05
Release name: Evolving Emergence

Course aims and structure

Course aims

Possible questions

Is Biological Physics well defined?
A pragmatic answer is along the lines of what the late Sir Sam Edwards (former Cavendish professor, and pioneer of polymer physics) was fond of saying: "Physics is what physicists do". Sir Sam was not the first to hold this view, and in the Cavendish there has always been a strong tradition of applying physics to new areas, regardless of traditional disciplines. Often this process requires developing new physics. The rationale behind this is that Science is one enterprise, so are the challenges that are posed by the "real world". You need to find areas where good progress can be made and where it is worth putting your effort. In this sense, if a physicist sees an opportunity to contribute in a unique way to biology, this can/should be pursued. Is is pretty obvious that "biology" is itself very broad (consider how many aspects there are to living systems, reflected for example 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 how these processes come into play in medicine. One also can pose questions on dynamics and evolution, and again the relevant *time-scales also span many orders of magnitude*, from molecular binding processes through to organism development and maintenance of tissues, up to the mechanisms of evolution or ecological equilibria.

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

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

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

What can be achieved in a course of 24 hours? The main aims of this course are:
(a) through good examples, and with a storyline as coherent as possible, mostly from the molecular to the cell level, show how physics (particularly statistical mechanics, soft matter, networks and nonlinear dynamics approaches) has been developed and applied in recent years to address both existing challenges, and even to define new categories in biological systems.
(b) provide (through (a)) an exposure and an education such that interested students will be able to make informed decisions on

fields of further study.

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

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

Structure of the course

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

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

tions and question sheet problems come from (**?**), which is a very 'reader friendly' source. The book does not cover everything (and we don't use the whole book), and in some places we wanted to go deeper, so other sources are also used, and referenced in appropriate places.

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

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

Module B: Population growth, genetic and evolution 3 lectures

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

Module C: Dynamics in cells. 3 lectures

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

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

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

Module E: Spatial Patterns. 2+1 lectures

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

Module F: Protein production and regulation of gene expression. 4+1 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 G: Circuits and dynamical systems. 2+2 lectures

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

Conclusion: Outlook beyond cell biological physics - 1 lecture

Terms and quantities in cell biology

The introductory material of the first couple of lectures can be found on the overheads. Here is firstly a glossary of terms, most of which should become familiar after a few lectures and on reading the first chapters of (**?**), and secondly two tables of useful data.

Glossary

Extended and modified from p.265 of U.Alon, and p.297 of Sneppen-Zocchi books.

Activator - A transcription factor that increases the rate of transcription of a gene when it binds a specific site in the gene's promoter.

Activation threshold - Concentration of activator in its active state needed for half-maximal activation of a gene.

Adaptation - Decreasing response to a stimulus that is applied continuously.

Adaptation time - Time for output to recover to 50% of prestimulus level following a step stimulus.

Allele - One of a set of alternative forms of a gene. In a diploid organism, such as most animal cells, each gene has two alleles, one on each of the two sister chromosomes.

Amino acid - A molecule that contains both an amino group (NH_2) and a carboxyl group (COOH). Amino acids are linked together by peptide bonds and serve as the constituents of proteins.

AND gate - A logic function of two inputs that outputs a one only if both inputs are equal to one.

Anti-motif - A pattern that occurs in a network less often than expected at random.

Antibody - A protein produced by a cell of the immune system

that recognizes a protein present in or on invading microorganisms.

Antigen - A part of a protein or other molecule that is recognized by an antibody.

Arabinose - A sugar utilized by *E. coli* as an energy and carbon source, using the *ara* genes. Arabinose is not pumped into the cells if glucose, a better energy source, is present.

ATP (adenosine triphosphate) - A molecule that is the main currency in the cellular energy economy. The conversion of ATP to ADP (adenosine diphosphate) liberates energy.

***B. subtilis* (*Bacillus subtilis*)** - A bacterium commonly found in the soil. It forms durable spores upon starvation. A model organism for study, and commonly used in synthetic biology.

Binomial distribution - A statistical distribution that describes, for example, the probability for k heads out of n throws of a coin that has probability p to give heads and $l-p$ to give tails.

Chemoreceptor - A receptor that responds to the presence of a particular chemical.

Chemotaxis - Movement up spatial gradients of specific chemicals (attractants), or down gradients of specific chemicals (repellents).

Chromosome - A strand of DNA with its associated proteins, found in the nucleus; carries genetic information.

Circadian rhythm - A daily rhythmical cycle of cellular activity. Generated by a biochemical oscillator in many different cells in animals, plants, and microorganisms. Ihe oscillations can be entrained by periodic temperature and light signals. The oscillator runs also in the absence of entraining external signals (usually with a period somewhat different than 24 hrs).

Coalescence - The merging of genetic lineages backwards in time to a recent common ancestor.

Codon - Three consecutive letters on an mRNA. There are 64 codons (each made of three letters, A, C, G, and U). These code for the 20 amino acids (with most amino acids represented by more than one codon). Three of the codons signal translational stop (end of the protein).

Coherent feed-forward loop - A feed-forward loop in which the sign of the direct path from X to Z is the same as the sign of the indirect path from X through Y to Z.

Complementary sequence Sequence of bases that can form a double-stranded structure by matching base pairs. The complementary sequence to base pairs C-T-A-G is G-A-T-C.

Cooperativity More than the sum of its parts. Acting cooperatively means that one part helps another to build a better functioning system. Cooperative bindings include dimerization, tetramerization, and binding between transcription factors on adjacent DNA sites.

Cost-benefit analysis - A theory that seeks the optimal design such that the difference between the fitness advantage gained by a system (benefit) and fitness reduction due to the cost of its parts is maximal.

Cytoplasm - The viscous, semiliquid substance contained in the interior of a cell. The cytoplasm is densely picked with proteins ('crowding').

Degree-preserving random networks - An ensemble of randomized networks that have the same degree sequence (the number of incoming and outgoing edges for each node in the network) as the real network. Despite the fact that the degree sequence is the same, the identity of which node connects to which other node is randomized. Such random networks can be generated on the computer by randomly switching pairs of edges, repeating the switching operation many times until the network is randomized. For a given real network, many thousands of different randomized degree-preserving networks can usually be readily generated.

Developmental transcription networks - Networks of transcription interactions that guide changes in cell type. Important examples are networks that guide the selection of cell fate as cells in the embryo differentiate into tissues. Developmental transcription networks work on the timescale of cell generations and often make irreversible decisions. They stand in contrast to sensory transcription networks that govern responses to environmental signals.

Differentiation - The process in which a cell changes to a different type of cell (same genome).

Diploid Individual - Individual carrying two copies of its genome. The two copies do not have to be identical.

Distributions Some common ones:
exponential

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

If t is a waiting time this is the distribution for a random uncorrelated signal. In that case the expected waiting time for the next signal does not change as time passes since the last signal.

power law

$$p(t) \sim 1/t^{\alpha}.$$

For example, if t is a waiting time, then expected waiting time for the next signal increases as time passes since the last signal.

normal or Gaussian distribution
Obtained by sum of exponentially bounded random numbers that are uncorrelated. Distribution:

$$p(x) \sim \exp\left(-x^2/\sigma^2\right).$$

log normal
Obtained by product of exponentially bounded random numbers that are uncorrelated. If x is normal distributed then $y = \exp(x)$ is log normal:

$$q(y)dy \sim \exp(-\log(y)^2/\sigma^2)dy/y \,\text{and}\, \sim\, dy/y$$

for y within a limited interval.

stretched exponentials
These are of the form

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

Pareto-Levi
Obtained from the sum of numbers, each drawn from a distribution $\propto x^{-\alpha}$. A Pareto-Levi distribution has a typical behavior like a Gaussian, but its tail is completely dominated by the single largest event. Thus a Pareto-Levi distribution has a power-law tail.

binomial
with parameters n and p is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p. Probability of k successes is:

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Poisson
expresses the probability of a given number of events (i.e. k, discrete) occurring in a fixed interval of time and/or space, if these events occur with a known average rate λ and independently of the time since the last event. The probability of a random variable being $X = k$ is:

$$p(X = k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}.$$

It has the special property that $\lambda = \langle p(X) \rangle = \mathrm{variance}(p(X))$.

DNA (deoxyribonucleic acid) - A long molecule composed of two interconnected helical strands. Contains the genetic information. Each strand in the DNA is made of four bases: A, C, T and G. The two strands pair with each other so that A pairs with T, and C with G. Thus DNA is made of a chain of base-pairs and can be represented by a string of four types of letters.

Dorsal - Side of an animal closer to its back.

Drosophila - Fruit fly. A model organism commonly used for biological research.

Edge - A link between two nodes in a network. Edges describe interactions between the component described by the nodes. Edges in most networks have a specific direction. Mutual edges are edges that link nodes in both directions. See transcription network for an example.

Endocytosis - Uptake of material into a cell.

Enzyme - A protein that facilitates a biochemical reaction. The enzyme catalyzes the reaction and does not itself become part of the end product.

ER (Erdos-Renyi) random networks - An ensemble of random networks with a given number of nodes, N, and edges, E. The edges are placed randomly between the nodes. This model can be used for comparison to real networks. A more stringent random model is the degree-preserving random network.

***E. coli* (*Escherichia Coli*)** - A rod-shaped bacterium normally found in the colon of humans and other mammals. It is widely studied as a model organism.

Eukaryotic cells and organisms - Organisms made of cells with a nucleus. Includes all forms of life except for viruses and bacteria (prokaryotes). Yeast is a single-celled eukaryotic organism.

Exponential phase - A phase of bacterial (possible also for other cell types) growth in which cells double with a constant cell generation time, resulting in exponentially increasing cell numbers. This occurs in a test tube when there are so few cells that nutrients are not depleted from the medium, and waste products do not accumulate to high levels. See also stationary phase.

Extinction - The process by which an allele (or a species) disappears from a population.

Feedback - A process whereby some proportion or function of the output signal of a system is passed (fed back) to the input.

Feedback inhibition - A common control mechanism in metabolic networks, in which a product inhibits the first enzyme in the pathway that produces that product.

Feed-forward loop (FFL) - A pattern with three nodes, X, Y and Z, in which X has a directed edge to Y and Z, and Y has a directed edge to Z. The FFL is a network motif in many biological networks and can perform a variety of tasks (such as sign-sensitive delay, sign-sensitive acceleration, and pulse generation).

Fine-tuned property - A property of a biological circuit that depends sensitively on the biochemical parameters of the circuit (opposite to robust property).

First-order kinetics - Mathematical description of the rate of an enzymatic reaction in the limit where the substrate concentration is very low and is far from saturating the enzyme, such that the rate is equal to $(v/K)\,E\,S$, where v is the rate per enzyme, E is the enzyme concentration, K is the Michaelis constant, and S is the substrate concentration. See also Michaelis-Menten kinetics, zero-order kinetics.

Fixation - The process by which one allele reaches frequency 100% in the population.

Fitness - The growth advantage associated to specific genetic traits.

Flagellum (plural flagella) - A long filament whose rotation drives bacteria through a fluid medium. Rotated by the flagellar

motor.

Functionalism - The strategy of understanding an organism's structural or behavioral features by attempting to establish their usefulness with respect to survival or reproductive success.

Gene - The functional unit of a chromosome, which directs the synthesis of one protein (or several alternate forms of a protein). The gene is transcribed into mRNA. which is then translated into the protein. The gene is preceded by a regulatory DNA region called the promoter that includes binding sites for transcription factors that regulate the rate of transcription.

Gene circuit - A term used here to mean a set of biomolecules that interact to perform a dynamical function. An example is a feed-forward loop.

Gene product - The protein encoded by a gene. Sometimes, the RNA transcribed from the gene, when the RNA has specific functions.

Generation time - Mean time for an organism to produce offspring.

Genetic code - The mapping between the 64 codons and the 20 amino acids. The genetic code is identical in nearly all 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).

Haploid Individual - An individual carrying only one copy of its genome.

Heterozygous - Diploid individual whose two genome copies differ at a specific site.

Homozygous- Diploid individual whose two genome copies are identical at a specific site.

Hill coefficient - See also 'Michaelis-Menten kinetics'. The Hill coefficient in kinetics reactions is the power exponent on the concentration of reagent. Mechanistically, it corresponds to the number of molecules that must act simultaneously in order to make a given reaction happen. The higher the Hill coefficient, the sharper the transition.

Histones - only in in eukaryotes, these are DNA binding proteins that regulate the condensation of DNA, i.e. determine the physical structure. The DNA makes two turns around each histone. Histones play a major role in gene silencing in eukaryotes, and a large fraction of transcription regulators in yeast, for example, is associated with histone modifications.

Homeostasis - The process by which the organism's substances and characteristics are maintained at their steady (optimal) level. Typically the result of a negative (stabilizing) regulative feedback.

Homologous - Similar by virtue of a common evolutionary origin. Homologous genes generally show similarity in their sequence.

Hormone - A chemical substance liberated by an endocrine gland that has effects on target cells in other organs.

Immune system - The system by which an organism protects

itself from foreign proteins. In mammals there are an innate and an adaptive system. The innate system triggers inflammation and recruitment of further immune cells. In response to an infection, the white blood cells (adaptive system) can produce antibodies that recognize and attack invading microorganisms, and typically some memory of this remains in the organism.

Integral feedback - Feedback on a device in which the integral over time of the error (output minus the desired output) is negatively fed back into the input of the device. Integral feedback can lead to robust exact adaptation.

Kinase - An enzymatic protein that transfers a phosphate group (PO_4) from a phosphate donor to an acceptor amino acid in a substrate protein (an important example of 'post-transcriptional modification', i.e. the regulation mechanisms that a cell deploys on proteins, the final products of gene expression). Kinases have been classified after acceptor amino acids.

Lac operon - A group of three genes in *E. coli* that are adjacent on the chromosome and transcribed on the same mRNA. These genes are *lacZYA*, encoding for the metabolic enzyme LacZ which cleaves lactose into glucose and galactose; the permease (pump) LacY, which pumps lactose into the cells; and LacA, whose function is unknown. Lactose is not pumped into the cells if glucose (a better energy source) is present, a phenomenon called "inducer exclusion". The *lac* operon is repressed by LacI and activated by CRP. LacI unbinds from the DNA and the system is induced in the presence of lactose (LacI binds a derivative of lactose called allolactose) or non-metabolizable analogs of lactose such as IPTG. As well as having a key importance in bacteria, this switch has been and continues to be a test-bed for quantitative work on understanding regulation of gene expression.

Lactose - A sugar utilized by *E. coli* as an energy and carbon source, using the *lac* genes expressed from the *lac* operon.

Ligand - A molecule that specifically binds the binding site of a receptor.

Mathematically controlled comparison - A comparison that is carried out with equivalence of as many internal and external parameters as possible between the alternative model mechanisms. Internal parameters include biochemical parameters, such as the lifetime of the proteins that make up the circuit and external parameters include desired output properties, such as steady-state levels.

Membrane - A structure consisting principally of lipid molecules that define the outer boundaries of a cell or organelle.

Membrane potential - The difference in electrical potential inside and outside of the cell expressed as voltage relative to the outside voltage. Membrane potential is maintained by protein pumps that transport ions across the membrane at the expense of energy supplied by ATP.

Michaelis-Menten kinetics - A mathematical description of the rate of an enzymatic reaction as a function of the concentration of the substrate. The rate is equal to $v\,E\,S/(K\,+\,S)$, where v is the rate per enzyme, E is the enzyme concentration, S is the substrate concentration, and K is the Michaelis constant. When $S >> K$ one obtains zero-order kinetics (rate $= v\,E$), and when $S << K$ one obtains first-order kinetics (rate $= (v/K)\,E\,S$).

Modularity - A property of a system which can be separated into nearly independent sub-systems.

Morphogen - A molecule (protein) that determines spatial patterns. Morphogens bind specific receptors to trigger signal transduction pathways within the cells to be patterned. The signaling leads the cells to assume different cell fates according to the morphogen level.

Morphology - Physical shape and structure.

mRNA - A macromolecule made of a sequence of four types of bases: A, C, G and U. Transcription is the process by which an RNA-polymerase enzyme produces an mRNA molecule that corresponds to the base sequence on the DNA (where DNA T is mapped to RNA U). The mRNA is read by ribosomes, which 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 (base-pair) in the DNA.

Poisson distribution - A distribution that characterizes a random process such as the number of heads in a coin-toss experiment, with many tosses, N, and a small probability for heads, $p << 1$. The mean number of heads is $m = pN$. The variance in a Poisson process is equal to the mean. $\sigma^2 = m$ and hence the standard deviation is the square root of the mean, $\sigma = \sqrt{m}$.

Prokaryotes - Single-celled organisms without a membrane around the nucleus. It is estimated that there are $(4 - 6) \times 10^{30}$ prokaryotes on Earth. The number of prokaryote divisions per year is $\approx 1.7 \times 10^{30}$. Prokaryotes are estimated to contain about the same amount of carbon as all plants on Earth (5×10^{14} kg). Some 5000 species have been described, but there are estimated

to be more than 10^6 species.

Promoter - A regulatory region of DNA that controls the transcription rate of a gene. The promoter contains a binding site for RNA polymerase (RNAp), the enzyme that transcribes the gene to produce mRNA. Each promoter also usually contains binding sites for transcription factor proteins. The transcription factors, when bound, affect the probability that RNAp will initiate transcription of an mRNA.

Protease - An enzyme that degrades proteins. Proteins are often targeted for degradation in biologically regulated ways. For example, many eukaryotic proteins are targeted for degradation in the proteosome by enzymes that attach a chain of ubiquitin molecules to the target protein. Different proteins can have different degradation rates.

Protein - A long chain of amino acids (a polymer, on the order of tens to hundreds of amino acids) that can serve in a structural capacity or as an enzyme. Each protein is encoded by a gene. Proteins are produced in ribosomes, based on information encoded on an mRNA that is transcribed from the gene.

Receptor - A protein molecule, usually situated in the membrane of the cell (but sometimes in the cytoplasm of the cell) that is sensitive to a particular chemical. When the appropriate chemical (the ligand) binds to the binding site of the receptor, signal transduction cascades are triggered within the cell.

Repression threshold - Concentration of active repressor needed for half-maximal repression of a gene.

Repressor - A transcription factor that decreases the rate of transcription when it binds a specific site in the promoter of a gene.

Ribosome - A structure in the cytoplasm made of about 100 proteins and special RNA molecules that serves as the site of production of proteins translated from mRNA. In the ribosome, amino acids are assembled to form the protein chain according to an order specified by the codons on the mRNA. The amino acids are brought into the ribosome by tRNA molecules, which read the mRNA codons. Each tRNA is released when its amino acid is linked to the translated protein chain.

RNA Polymerase (RNAp) - A complex of several proteins that form an enzyme that transcribes DNA into RNA. There is also DNA polymerase, the complex used to make copies of DNA before cell division.

Robust Property - Property X is robust with respect to parameter Y, if X is insensitive to changes in parameter Y.

Sensory transcription networks - Transcription networks that respond to environmental and internal signals such as nutrients and stresses, and lead to changes in gene expression. These networks need to function rapidly, usually within less than a cell generation time, and usual make reversible decisions. They stand in contrast to developmental transcription networks.

Stationary phase - A state in which cells cease to divide and grow, that occurs when growth conditions are unfavorable, such as when the bacteria run out of an essential nutrient. See also exponential phase.

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.

Useful Estimation Data

	Quantity of interest	Symbol	Rule of thumb
E. coli			
	Cell volume	$V_{E.\,coli}$	$\approx 1\,\mu m^3$
	Cell mass	$m_{E.\,coli}$	≈ 1 pg
	Cell cycle time	$t_{E.\,coli}$	≈ 3000 s
	Cell surface area	$A_{E.\,coli}$	$\approx 6\,\mu m^2$
	Macromolecule concentration in cytoplasm	$c_{E.\,coli}^{\text{macromol}}$	≈ 300 mg/mL
	Genome length	$N_{\text{bp}}^{E.\,coli}$	$\approx 5 \times 10^6$ bp
	Swimming speed	$v_{E.\,coli}$	$\approx 20\,\mu m/s$
Yeast			
	Volume of cell	V_{yeast}	$\approx 60\,\mu m^3$
	Mass of cell	m_{yeast}	≈ 60 pg
	Diameter of cell	d_{yeast}	$\approx 5\,\mu m$
	Cell cycle time	t_{yeast}	≈ 200 min
	Genome length	$N_{\text{bp}}^{\text{yeast}}$	$\approx 10^7$ bp
Organelles			
	Diameter of nucleus	d_{nucleus}	$\approx 5\,\mu m$
	Length of mitochondrion	l_{mito}	$\approx 2\,\mu m$
	Diameter of transport vesicles	d_{vesicle}	≈ 50 nm
Water			
	Volume of molecule	V_{H_2O}	$\approx 10^{-2}\,nm^3$
	Density of water	ρ	$1\,g/cm^3$
	Viscosity of water	η	≈ 1 centipoise (10^{-2} g/(cm s))
	Hydrophobic embedding energy	$\approx E_{\text{hydr}}$	2500 cal/(mol nm^2)
DNA			
	Length per base pair	l_{bp}	$\approx 1/3$ nm
	Volume per base pair	V_{bp}	$\approx 1\,nm^3$
	Charge density	λ_{DNA}	$2\,e/0.34$ nm
	Persistence length	ξ_p	50 nm
Amino acids and proteins			
	Radius of "average" protein	r_{protein}	≈ 2 nm
	Volume of "average" protein	V_{protein}	$\approx 25\,nm^3$
	Mass of "average" amino acid	M_{aa}	≈ 100 Da
	Mass of "average" protein	M_{protein}	$\approx 30{,}000$ Da
	Protein concentration in cytoplasm	c_{protein}	≈ 150 mg/mL
	Characteristic force of protein motor	F_{motor}	≈ 5 pN
	Characteristic speed of protein motor	v_{motor}	≈ 200 nm/s
	Diffusion constant of "average" protein in cytoplasm	D_{protein}	$\approx 10\,\mu m^2/s$
Lipid bilayers			
	Thickness of lipid bilayer	d	≈ 5 nm
	Area per molecule	A_{lipid}	$\approx \frac{1}{2}\,nm^2$
	Mass of lipid molecule	m_{lipid}	≈ 800 Da