Physical Biology of the Cell

Rob Phillips
Jané Kondev
Julie Theriot

illustrated by Nigel Orme

Garland Science

Chapter 2

What and Where: Construction Plans for Cells and Organisms

"Although not everyone is mindful of it all cell biologists have two cells of interest: the one they are studying and *Escherichia coli*." - F. Neidhardt

Chapter Overview: In Which We Consider the Size of Cells and the Nature of Their Contents

Cells come in a dazzling variety of shapes and sizes. Even so, their molecular inventories share many common features, reflecting the underlying biochemical unity of life. In this chapter, we introduce the bacterium *Escherichia coli* (we will abbreviate this cell type as *E. coli* throughout the book) as our biological standard ruler. This cell serves as the basis for a first examination of the inventory of cells and will permit us to get a sense of the size of cells and the nature of their contents. Indeed, using simple estimates, we will take stock of the genome size, numbers of lipids and proteins and the ribosome content of bacteria. With the understanding revealed by *E. coli* in hand, we then take a powers-of-ten journey down and up from the scale of individual cells. Our downward journey will examine organelles within cells, macromolecular assemblies ranging from ribosomes to viruses and then the macromolecules that are the engines of cellular life. Our upward journey from the scale of individual cells will examine a second class of biological structures, namely those resulting from different forms of multicellularity, this time with an emphasis on how cells act together in contexts ranging from bacterial biofilms to the networks of neurons in the brain.

2.1 An Ode to *E. coli*

Scientific observers of the natural world have been intrigued by the processes of life for many thousands of years as evidenced by early written records from Aristotle, for example. Early thinkers wondered about the nature of life and its “indivisible” units in much the same way that they mused about the fundamental units of matter. Just as physical scientists arrived at a consensus that the fundamental unit of matter is the atom (at least for chemical transactions), likewise, observers of living organisms have agreed that the indivisible unit of life is the cell. Nothing smaller than a cell can be shown to be alive in a sense that is generally agreed upon. At the same time, there are no currently known reasons to attribute some higher “living” status to multicellular organisms.

Cells are able to consume energy from their environments and use that energy to create ordered structures. They can also harness energy from the environment to create new cells. A standard definition of life merges the features of metabolism (that is, consumption and use of energy from the environment) and replication (giving offspring that resemble the original organism). Stated simply, the cell is the smallest unit of replication (though viruses are also replicative units, but depend upon their infected host to provide much of the machinery making this replication possible).

The recognition that the cell is the fundamental unit of biological organization originated in the seventeenth century with the microscopic observations of Hooke and van Leeuwenhoek. This idea was put forth as the modern cell theory by Schwann, Schleiden and Virchow in the mid-nineteenth century and was confirmed unequivocally by Pasteur shortly thereafter and repeatedly in the time since. Biologists agree that all forms of life share cells as the basis of their organization. It is also generally agreed that all living organisms on earth shared a common ancestor several billion years ago that would be recognized as a cell by any modern biologist. In terms of understanding the basic rules governing metabolism, replication and life more generally, one cell type as the basis of experimental investigations of these mechanisms should be as good as any other. For practical reasons, however, biologists have focused on a few particular types of cell to try to illuminate these general issues. Among these, the human intestinal inhabitant *E. coli* stands unchallenged as the most useful and important representative of the living world in the biologist’s laboratory.

Several properties of *E. coli* have contributed to its great utility and has made it a source of repeated discoveries. First, it is easy to isolate because it is present in great abundance in human fecal matter. Unlike most other bacteria that populate the human colon, *E. coli* is able to grow well in the presence of oxygen. In the laboratory, it replicates rapidly and can easily adjust to changes in its environment including changes in nutrients. In addition, using molecular biology, the generation of mutants is nearly routine. Mutant organisms are those which differ from their parents and from other members of their species found in the wild because of specific changes in DNA sequence which give rise to biologically significant changes. For example, *E. coli* is normally able to synthesize purines for DNA and RNA on its own from sugar as a nutrient source. However, particular mutants of *E. coli* with enzymatic deficiencies in these pathways have lost the ability to make their own purines and become reliant on being fed precursors for these molecules. A more familiar and frightening example is the way in which mutant bacteria acquire antibiotic resistance. Throughout the book we will be using specific examples of biological phenomena to illustrate general physical principles that are relevant to life. Often, we will have recourse to *E. coli* because of particular experiments that have been performed on this organism. Further, even when we speak of experiments on other cells or organisms, often *E. coli* will be behind the scenes coloring our thinking.

2.1.1 The Bacterial Standard Ruler

The Bacterium *E. coli* Will Serve as Our Standard Ruler

Throughout the book we will discuss many different cells which all share with *E. coli* the fundamental biological directive to convert energy from the environment into structural order and to perpetuate their species. On Earth, it is observed that there are certain minimal requirements for the perpetuation of cellular life. These are not necessarily absolute physical requirements, but in the competitive environment of our planet, all surviving cells share these features in common. These include a DNA-based genome, mechanisms to transcribe DNA into RNA and subsequently, translation mechanisms using ribosomes to convert information in RNA sequences into protein sequence and protein structure. Within those individual cells, there are many substructures with interesting functions. For example, the ribosomes that generate proteins from RNA sequence and the individual proteins that they create are both important classes of substructure. Larger than the cell there are also structures of biological interest that arise because of cooperative interactions between many cells. These include higher organisms such as Redwood trees and sharks. In this chapter, we will begin with the cell as the fundamental unit of biological organization using *E. coli* as the standard reference and standard ruler. We will then look at smaller structures within cells and finally, larger multicellular structures, zooming in and out from our fundamental cell reference frame.

Fig. 2.1 shows several experimental pictures of an *E. coli* cell and its schematization into our standard ruler. In particular, the electron micrograph in fig. 2.1 shows that these bacteria have a rod-like morphology with a typical length between 1 and 2 microns and a diameter between 1/2 and 1 micron. To put the standard ruler in perspective, we note that with its characteristic length scale of 1 micron, it would take roughly fifty such cells lined up end to end in order to measure out the width of a human hair. On the other hand, we would need to divide the cell into roughly five hundred slices of equal width in order to measure out the diameter of a DNA molecule. Note that the average size of these cells depends on the nutrients they are provided, with those growing faster also having a larger size. Our reference growth condition throughout the book will be a chemically defined solution referred to by microbiologists as "minimal media" with glucose as the sole carbon source. "Minimal medium"

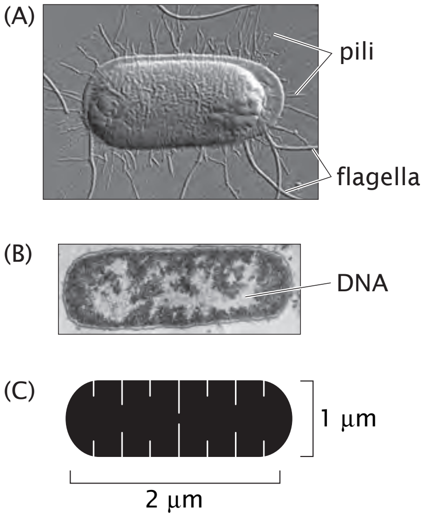


Figure 2.1: *E. coli* as a standard ruler for characterizing spatial scales. (A) Atomic force microscopy image of an *E. coli* cell (courtesy of C. T. Lim), (B) Electron micrograph of *E. coli* bacterium, (C) the *E. coli* ruler.

refers to a completely chemically defined mixture of salts, sugars, amino acids and vitamins that can support the growth of a microorganism. In the laboratory, bacteria are often grown in "rich media", which are poorly defined but nutrient-rich mixtures of extracts from organic materials such as yeast cultures or cow brains. Although microorganisms can grow very rapidly in rich media, they are rarely used for biochemical studies because their exact contents are not known. In minimal media, however, it is easy to simply leave out or add a single compound (for example, a single amino acid such as tryptophan) and measure the effects of that compound on the microorganism's growth.

Because of its central role as the quantitative standard in the remainder of the book, it is useful to further characterize the geometry of *E. coli*. One example in which we will need a better sense of the geometry of cells and their internal compartments is in the context of reconciling *in vitro* (i.e. in test tubes) and *in vivo* (i.e. in living cells) experiments. Results from solution biochemistry are based upon the concentrations of different molecular species. On the other hand, in *in vivo* situations we might know the number of copies of a given molecule such as a transcription factor. To reconcile these two pictures, we will need the cellular volume to make the translation between molecular counts and concentrations. Similarly, when examining the distribution of membrane proteins on the cell surface, to estimate the mean spacing between these proteins, which will tell us about the extent of interactions between them, we will need a sense of the cell area. For most cases of interest in this book, it suffices to attribute a volume $V_{E.coli} \approx 1\ \mu\mathrm{m}^3 = 1$ fL to *E. coli* and an area of roughly $A_{E.coli} \approx 6\ \mu m^2$ (see the problems to actually work out these numbers from known cellular dimensions).

2.1.2 Taking the Molecular Census

In the remainder of this section, we will proceed through a variety of estimates to try and get a grip on the number of molecules of different kinds that are in an *E. coli* cell. Why should we care about these numbers? First, a realistic physical picture of any biological phenomenon demands a precise, quantitative understanding of the individual particles involved (for biological phenomena, this usually means molecules) and the spatial dimensions over which they have the freedom to act. One of the most immediate outcomes of our cellular census will be the realization of just how crowded the cellular interior really is, a subject explored in detail in chap. 14. Our census will paint a very different picture of the cellular interior as the seat of biochemical reactions than is suggested by the dilute and homogeneous environment of the biochemical test tube. Indeed, we will see that the mean spacing between protein molecules within a typical cell is less than 10 nm.

Taking the molecular census is also important because we will use our molecular counts in chap. 3 to estimate the rates of macromolecular synthesis during the cell cycle. How fast is a genome replicated? What is the average rate of protein synthesis during the cell cycle and given what we know about ribosomes, how do they maintain this rate of synthesis? A prerequisite to beginning to answer these questions is the macromolecular census itself.

Ultimately, to understand many experiments in biology, it is important to realize that most experimentation is comparative. That is, we compare “normal” behavior to perturbed behavior to see if some measurable property has increased or decreased. To make these statements meaningful, we need to first understand the quantitative baseline relative to which such increases and decreases are compared. There is another sense in which numbers of molecules are particularly meaningful which will be explored in detail in subsequent chapters that has to do with whether we can describe a cell as having “a lot” or “a few” copies of some specific molecule. If a cell has a lot of some particular molecule, then it is appropriate to describe the concentration of that molecule as the basis for predicting cellular function. However, when a cell has only a few copies of a particular molecule, then we need to consider the influence of random chance (or stochasticity) on its function. In many cases, cells have an interesting medium number of molecules where it is not immediately clear which perspective is appropriate. However, knowing the absolute numbers always gives us a reality check for subsequent assumptions and approximations for modeling biological processes.

Because of these considerations, in recent years much effort among biological scientists has been focused on the development of quantitative techniques for measuring the molecular census of living cells (both bacteria and eukaryotes). In this chapter we will rely primarily on order-of-magnitude estimates based on simple assumptions. These estimates are validated by comparison with measurements. In subsequent chapters, these estimates will be refined through explicit model building and direct comparison to quantitative experiments.

- **Estimate: Sizing Up** *E. coli***.** As already noted in the previous chapter, cells are made up of an array of different macromolecules as well as small molecules and ions. To estimate the number of proteins in an *E. coli* cell we begin by noting that with its 1 fL volume, the mass of such a cell is roughly 1 pg, where we have assumed that the density of the cell is that of water which is 1 g/mL. Measurements reveal that the dry weight of the cell is roughly 30 percent of its total and half of that mass is protein. As a result, the total protein mass within the cell is roughly 0.15 pg. We can also estimate the number of carbon atoms in a bacterium on the grounds that roughly half the dry mass comes from the carbon content of these cells, a figure that implies 10^{10} carbon atoms per cell. Two of the key sources that have served as a jumping off point for these estimates are Neidhardt *et al.* (1990) and Zimmerman and Trach (1991), who describe the result of a molecular census of a bacterium.

  As a first step to revealing the extent of crowding within a bacterium, we can estimate the number of proteins by assuming a mean protein of 300 amino acids with each amino acid having a characteristic mass of 100 Da. These assumptions are further examined in the problems at the end of the chapter. Using these rules of thumb, we find that the mean protein has a mass of 30,000 Da. Using the conversion factor that 1 Da $\approx 1.6 \times 10^{-24}$ g, we have that our typical protein has a mass of 5×10^{-20} g. The number of proteins per *E. coli* cell is estimated as

  $$N_{protein} = \frac{\text{total protein mass}}{\text{mass per protein}} \approx \frac{15 \times 10^{-14}\ \text{g}}{5 \times 10^{-20}\ \text{g}} \approx 3 \times 10^6. \qquad (2.1)$$

  If we invoke the rough estimate that one-third of the proteins coded for in a typical genome correspond to membrane proteins this implies that the number of cytoplasmic proteins is of order 2×10^6 and the number of membrane proteins is 1×10^6, although we note that not all of these membrane-associated proteins are strictly transmembrane proteins.

  Another interesting use of this estimate is to get a rough impression of the number of ribosomes - the cellular machines that synthesize proteins. To be concrete, we need one other fact, which is that roughly 20 percent of the protein complement of a cell is ribosomal protein. If we assume that all of this protein is tied up in assembled ribosomes, then we can estimate the number of ribosomes by noting: a) that the mass of an individual ribosome is roughly 2.5MDa and b) that an individual ribosome is roughly 1/3 by mass protein and 2/3 by mass RNA, facts which can be directly confirmed by the reader by inspecting the structural biology of ribosomes.

As a result, we have

$$N_{ribosome} = \frac{0.2 \times 0.15 \times 10^{-12} g}{830,000 Da} \times \frac{1 Da}{1.6 \times 10^{-24} g} \approx 20,000 \text{ ribosomes.} \tag{2.2}$$

The numerator of the first fraction has 0.2 as the fraction of protein that is ribosomal, 0.15 as the fraction of the total cell mass that is protein and 1pg as the cell mass. 830,000Da is our estimate for that part of the ribosomal mass that is protein. The size of a ribosome is roughly 20nm (in "diameter") and hence the total volume taken up by these 20,000 ribosomes is roughly 10^8 nm^3. This is 10 percent of the total cell volume.

Idealizing an *E. coli* cell as a cube, sphere or spherocylinder yields (see the problems) that the surface area of such cells is $A_{E.coli} \approx 6 \mu m^2$. This number may be used in turn to estimate the number of lipid molecules associated with the inner and outer membranes of these cells as

$$N_{lipid} \approx \frac{4 \times 0.5 \times A_{E.coli}}{A_{lipid}} \approx \frac{4 \times 0.5 \times (6 \times 10^6 \ nm^2)}{0.5 \ nm^2} \approx 2 \times 10^7, \tag{2.3}$$

where the factor of 4 comes from the fact that the inner and outer membranes are each *bilayers*, implying that the lipids effectively cover the cell surface area four times. A lipid bilayer consists of two sheets of lipids with their tails pointing toward each other. The factor of 0.5 is based on the crude estimate that roughly half of the surface area is covered by membrane proteins rather than lipids themselves. We have made the similarly crude estimate that the area per lipid is 0.5 nm^2. The measured number of lipids is of order 2 × 10^7 as well.

In terms of sheer numbers, water molecules are by far the majority constituent of the cellular interior. One of the reasons this fact is intriguing is that during the process of cell division, a bacterium such as *E. coli* has to take on a very large number of new water molecules each second. The estimate we do here will be used to examine this transport problem in the next chapter. To estimate the number of water molecules we exploit the fact that roughly 70% of the cellular mass (or volume) is water. As a result, the total mass of water is 0.7 pg. We can find the approximate number of water molecules as

$$N_{H_2O} \approx \frac{0.7 \times 10^{-12} g}{18 g/mole} \times 6 \times 10^{23} \text{molecules/mole} \approx 2 \times 10^{10} \text{ water molecules.} \tag{2.4}$$

It is also of interest to gain an impression of the content of inorganic ions in a typical bacterial cell. To that end, we assume that a typical concentration of positively charged ions such as K$^+$ is 100 mM resulting in the estimate

$$N_{ions} \approx \frac{(100 \times 10^{-3} \text{moles}) \times (6 \times 10^{23} \text{molecules/mole})}{10^{15} \mu m^3} \times 1 \mu m^3 = 6 \times 10^7. \tag{2.5}$$

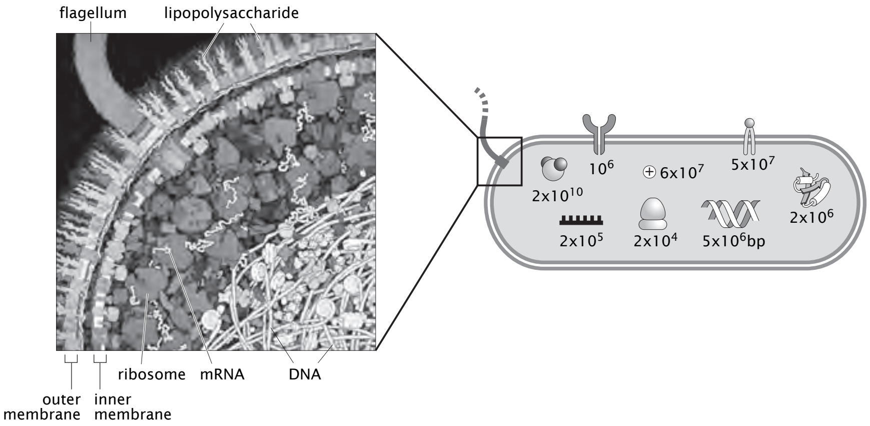


Figure 2.2: Molecular contents of the bacterium *E. coli.* The cartoon on the left shows the crowded cytoplasm of the bacterial cell. The cartoon on the right shows an order-of-magnitude molecular census of the *E. coli* bacterium with the approximate number of different molecules in *E. coli.*

> This result could have been obtained even more easily by noting yet another simple rule of thumb, namely, that one molecule per *E. coli* cell corresponds roughly to a concentration of 2 nM.
>
> The outcome of our attempt to size up *E. coli* is illustrated schematically in summary form in fig. 2.2. A more complete census of an *E. coli* bacterium can be found in Neidhardt *et al.* (1990). The outcome of experimental investigations of the molecular census of an *E. coli* cell is summarized (for the purposes of comparing to our estimates) in Table 2.1.2.

How is the census of a cell taken experimentally? This is a question we will return to a number of different times, but will give a first answer here. For the case of *E. coli*, one important tool has been the use of gels like that shown in fig. 2.3. Such experiments work by breaking open the contents of a cell and keeping only the protein component. By applying electric fields first in one direction and then in a perpendicular direction, it is possible to separate the proteins by both mass and charge. The intensity of the spots on such a gel can then be used as a basis for quantifying each species. Similar tricks are used to characterize the amount of RNA and lipids, for example, resulting in a total census like that shown in Table 2.1.2.

The Cellular Interior Is Highly Crowded With Mean Spacings Between Molecules That Are Comparable to Molecular Dimensions

One of the most intriguing implications of our census of the molecular parts

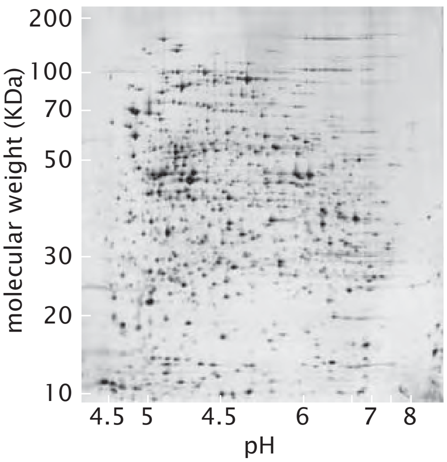


Figure 2.3: Experimental census of the cell. Measurement of protein census using two-dimensional polyacrylamide gel electrophoresis. Figure adapted from the 2DPage database.

Substance	% of total dry weight	Number of molecules
Macromolecule		
Protein	55.0	2.4×10^6
RNA	20.4	
23S RNA	10.6	19,000
16S RNA	5.5	19,000
5S RNA	0.4	19,000
Transfer RNA (4S)	2.9	200,000
Messenger RNA	0.8	1,400
Phospholipid	9.1	22×10^6
Lipopolysaccharide	3.4	1.2×10^6
DNA	3.1	2
Murein	2.5	1
Glycogen	2.5	4,360
Total macromolecules	**96.1**	
Small molecules		
Metabolites, building blocks, etc.	2.9	
Inorganic ions	1.0	
Total small molecules	**3.9**	

Table 2.1: Observed macromolecular census of an *E. coli* cell. Adapted from Neidhardt *et al.* and Schaechter *et al.*.

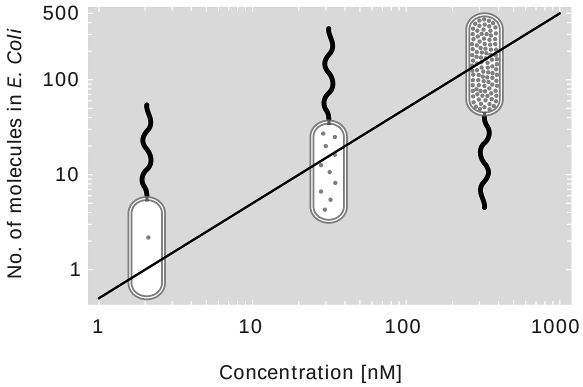


Figure 2.4: Concentration in *E. coli* units. Number of copies of a given molecule in a volume the size of an *E. coli* cell as a function of the concentration.

list of a bacterium is the extent to which the cellular interior is crowded. Because of experiments and associated estimates on the contents of *E. coli*, Goodsell undertook a series of attempts to depict the cellular interior in a way that respects the molecular census. The crowded environs of the interior of such a cell is shown in fig. 2.2. This figure gives a number of different views of the crowding associated with any *in vivo* process. In chap. 14, we will see that this crowding effect will force us to call in question our simplest models of chemical potentials, the properties of water and the nature of diffusion. We have already made an estimate of the typical spacing of ribosomes in bacterial cells. The generic conclusion is that the mean spacing of proteins and their assemblies is comparable to the dimensions of these macromolecules themselves. The cell is a very crowded place!

The quantitative significance of fig. 2.2 can be further appreciated by converting these numbers into concentrations. To do so, we recall that the volume of an *E. coli* cell is 1 fL. The rule of thumb that emerges from this analysis is that 2nM implies roughly one molecule per bacterium. A concentration of 2μM implies roughly 1000 copies of that molecule per cell. Concentration in terms of our standard ruler is shown in fig. 2.4. What is being plotted is the number of copies of the molecule of interest in such a cell as a function of the concentration.

We can use these concentrations directly to compute the mean spacing between molecules. That is, given a certain concentration, there is a corresponding average distance between the molecules. Having a sense of this distance can serve as a guide to thinking about the likelihood of diffusive encounters and reactions between various molecular constituents. If we imagine the molecules at a given concentration arranged on a cubic lattice of points, then the mean spacing between those points is given by

$$d = c^{-1/3}, \tag{2.6}$$

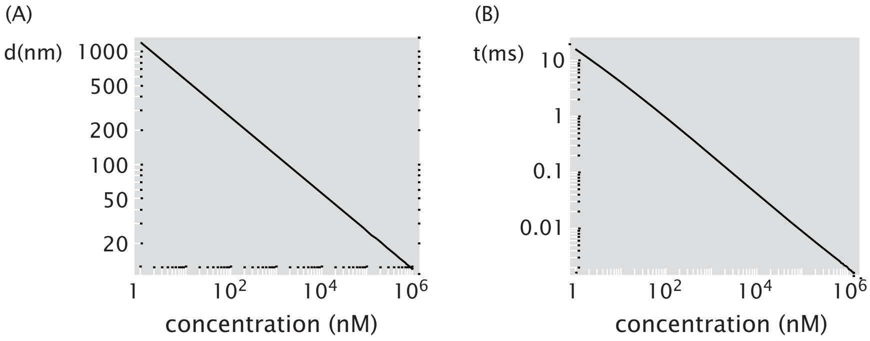


Figure 2.5: Different representations of concentration. (A) Concentration expressed in units of typical distance (d) between neighboring molecules measured in nanometers. (B) Diffusion time over the distance between neighboring molecules as a function of concentration. The diffusion constant $D = 100\ \mu m^2/sec$ is typical for a protein in water.

where c is the concentration of interest (measured in units of number of molecules per unit volume). Larger concentrations imply smaller intermolecular spacings. This idea is formalized in fig. 2.5 which shows the relation between the mean spacing measured in nanometers and the concentration.

2.1.3 Looking Inside of Cells

The remainder of the chapter focuses on the various structures that make up cells and organisms. To talk about these structures, it is helpful to have a sense of how we know what we know about them. Further, model building requires facts. To that end, we periodically take stock of the experimental basis for our models. For this chapter, the "Experiments Behind the Facts" focuses on how we know what we know about biological structures.

- **Experiments Behind the Facts: Probing Biological Structure.** To size up cells and their organelles we need to extract "typical" structural parameters from a variety of experimental studies. Though we leave a description of the design and setup of such experiments to more specialized texts, the goal is to provide at least enough details that the reader sees where some of the key structural facts that we will use throughout the book come from. We emphasize two broad categories of experiments: i) those in which some form of radiation interacts with the structure of interest and ii) those in which forces are applied to the structure of interest.

  Fig. 2.6 shows three distinct experimental strategies which feed into our estimates and all of which reveal different facets of biological structure. One of the mainstays of structural analysis is light microscopy. Fig. 2.6(A) shows a schematic of the way in which light can excite fluorescence in samples that have some distribution of fluorescent molecules within them. In particular, this example shows a schematic of a microtubule which has some distribution of fluorophores along its length. Incident photons of one wavelength are absorbed by the fluorophore and this excitation leads them to emit light of a different wavelength which is then detected. As a result of selective labeling of only the microtubules with fluorophores, when examined in the microscope it is only these structures that are observed. These experiments permit a determination of the size of various structures of interest, how many of them there are and where they are localized. By calibrating the intensity from single fluorophores it has become possible to take a single molecule census for many of the important proteins in cells.

A totally different window on the structure of the cell and its components is provided by tools such as the atomic-force microscope (AFM). As will be explained in chap. 10, the AFM is a cantilever beam with a sharp tip on its end. The tip is brought very close to the surface where the structure of interest is present and is then scanned in the plane. One way to operate the instrument is to move the cantilever up and down so that the force applied on the tip remains constant. Effectively, this demands a continual adjustment of the height as a function of the x-y position of the tip. The nonuniform pattern of cantilever displacements can be used to map out the structure of interest. Fig. 2.6(B) shows a schematic of an atomic-force microscope scanning a typical fibroblast cell.

Fig. 2.6(C) gives a schematic of the way in which x-rays or electrons are scattered off of a biological sample. The schematic shows an incident plane wave of radiation which interacts with the biological specimen and results in the emergence of radiation with the same wavelength but a new propagation direction. Each point within the sample can be thought of as a source of radiation and the observed intensity at the detector reflects the interference from all of these different sources. By observing the pattern of intensity it is possible to deduce something about the structure that did the scattering. This same basic idea is applicable to a wide variety of radiation sources including x-rays, neutrons and electrons.

An important variation on the theme of measuring the scattered intensity from irradiated samples is cryo-electron tomography. This technique is one of the centerpieces of structural biology and is built around uniting electron microscopy with sample preparation techniques which rapidly freeze the sample. The use of tomographic methods has made it possible to go beyond the planar sections seen in conventional electron microscopy images. The basic idea of the technique is indicated schematically in fig. 2.7, and is built around the idea of rotating the sample over a wide range of orientations and then to build up a corresponding three-dimensional reconstruction on the basis of the entirety of these images. These techniques have already revolutionized our understanding of particular organelles and are now being used to image entire cells.

2.1.4 Where Does *E. coli* Fit?

Biological Structures Exist Over a Huge Range of Scales

The spatial scales associated with biological structures run from the nanometer scale of individual molecules, all the way to the scale of the earth itself. Where does *E. coli* fit into this hierachy of structures? Fig. 2.8 shows the different structures that can be seen as we scale in and out from an *E. coli* cell. At each scale, new classes of structure can be seen. A roughly tenfold increase in magnification relative to an individual bacterium reveals the viruses that attack bacteria. These viruses, known as bacteriophage, have a characteristic scale of roughly 100 nm. They are made up of a protein shell (the capsid) which is filled with the viral genome. Continuing our downward descent using yet higher magnification, we see the ordered packing of the viral genome within its capsid. These structures are intriguing because they involve the ordered arrangement of more than 10 μm of DNA in a capsid which is less than 100 nm across. Another rough factor of ten increase in resolution reveals the structure of the DNA molecule itself with a characteristic cross sectional radius of roughly 1 nm and a length of 3.4 nm per helical repeat.

A similar scaling out strategy reveals new classes of structures. As shown in fig. 2.8, a tenfold increase in spatial scale brings us to the realm of eukaryotic cells in general, and specifically, to the scale of the epithelial cells that line the human intestine. We use this example because bacteria such as *E. coli* are a central player as part of our intestinal fauna. Another tenfold increase in spatial scale reveals one of the most important inventions of evolution, namely, multicellularity. In this case, the cartoon depicts the formation of planar sheets of epithelial cells. These planar sheets are themselves the building blocks of yet higher-order structures such as tissues. Scaling out to larger scales would bring us to multicellular organisms and the structures they build.

The remainder of the chapter is devoted to an attempt to take stock of the structures at each of these scales and to provide a feeling for the molecular building blocks that make up these different structures. Our strategy will be to build upon our cell-centered view and to first descend in length scale from that of cells to the molecules they are made of. Once this structural descent is complete, we will embark on an analysis of biological structure in which we zoom out from the scale of individual cells to collections of cells.

2.2 Cells and Structures Within Them

2.2.1 Cells: A Rogue's Gallery

All living organisms are based on cells as the indivisible unit of biological organization. However, within this general rule there is tremendous diversity among

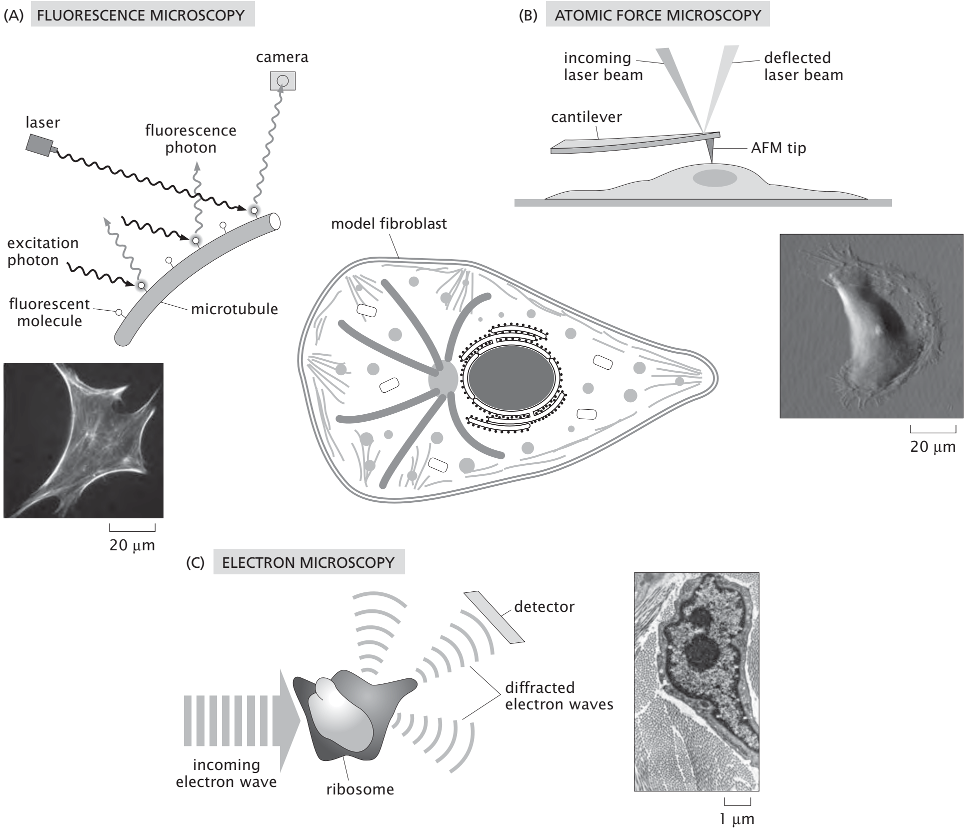


Figure 2.6: Experimental techniques which have revealed the structure of both cells and their organelles. (A) Fluorescence microscopy and associated image of fibroblast with labeled actin, (B) Atomic force microscopy schematic and associated image of surface topography of fibroblast. (C) Electron microscopy schematic and images of a fibroblast.

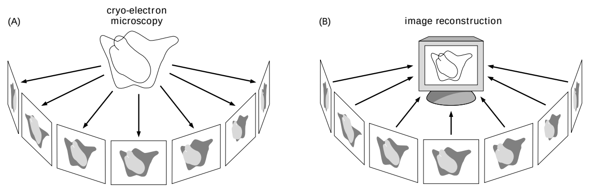


Figure 2.7: Schematic of tomographic reconstruction. (A) The sample is rotated so that radiation is scattered from a series of different orientations, (B) three-dimensional reconstruction of the structure giving rise to the pattern of scattering.

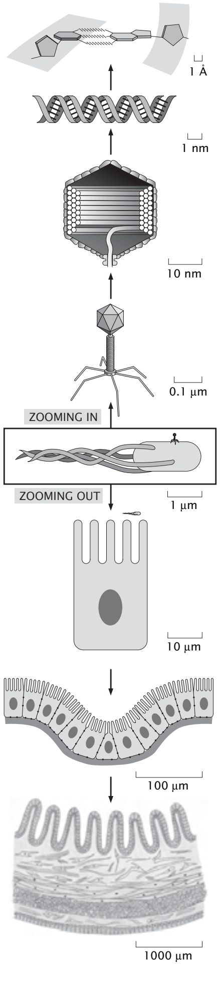


Figure 2.8: Powers of ten representation of biological length scales. The hierarchy of scales is built around the *E. coli* standard ruler. Starting with *E. coli* the first part of the chapter will consider a succession of tenfold increases in resolution as are shown in the figure. The second part of the chapter will zoom out from the scale of an *E. coli* cell.

living cells. Several billion years ago, our last common ancestor gave rise to three different lineages of cells now commonly called Bacteria, Archaea and Eukarya, a classification suggested by similarities and differences in ribosomal RNA sequences. Every living organism on earth is a member of one of these groups. Most bacteria and archaea are small (3 μm or less) and extremely diverse in their preferred habitats and associated lifestyles ranging from geothermal vents at the bottom of the ocean to permafrost in Antarctica. Bacteria and archaea look very similar to one another and it has only been within the last few decades that molecular analysis has revealed that they are completely distinct lineages that are no more closely related to each other than the two are to eukaryotes.

Most of the organisms that we encounter in our everyday life and can see with the naked eye are members of Eukarya (individuals are called eukaryotes). These include all animals, all plants ranging from trees to moss and also all fungi such as mushrooms and mold. Thus far we have focused on *E. coli* as a representative cell although we must acknowledge that *E. coli*, as a member of the bacterial group, is in some ways very different from a eukaryotic or archaeal cell. The traditional definition of a eukaryotic cell is one that contains its DNA genome within a membrane-bound nucleus. Most bacteria and archaea lack this feature and also lack other elaborate intracellular membrane-bound structures such as the endoplasmic reticulum and the Golgi apparatus that are characteristic of the larger and more complex eukaryotic cells.

Cells Come in a Wide Variety of Shapes and Sizes and With a Huge Range of Functions

Cells come in such a wide variety of shapes, sizes and lifestyles that choosing one representative cell type to tell their structural story is misleading. In fig. 2.9 we show a rogue's gallery illustrating the variety of cell sizes and shapes found in the eukaryotic group. This gallery is by no means complete. There is much more variety than we can illustrate, but this covers a reasonable range of eukaryotic cell types that have been well studied by biologists. In this figure we have chosen a variety of examples that represent experimental bias among biologists where more than half of the examples are human cells and the others represent the rest of the eukaryotic group. The vast majority of eukaryotes are members of a group called protists. This poorly-defined group encompasses all eukaryotes that are neither plants nor animals nor fungi. Protists are extremely diverse in their appearance and lifestyles, but they are all small (ranging from 0.002 mm to 2 mm). Some examples of protists include marine diatoms such as *Emiliana Huxleyi*, soil amoeba such as *Dictyostelium discoideum* and the lovely creature *Paramecium* seen in any sample of pond water and familiar from many high-school biology classes. Another notable protist is the pathogen that causes malaria called *Plasmodium falciparum*. Fig. 2.9(A) shows the intriguing protist *Giardia lamblia*, a parasite known to hikers as a source of water contamination.

Although protists constitute the vast majority of eukaryotic cells on the planet, biologists are often inclined to study cells more related to us. This includes the plant kingdom which is obviously important as a source of food and flowers. Plant cells like that shown in fig. 2.9(B) are characterized by a rigid cell

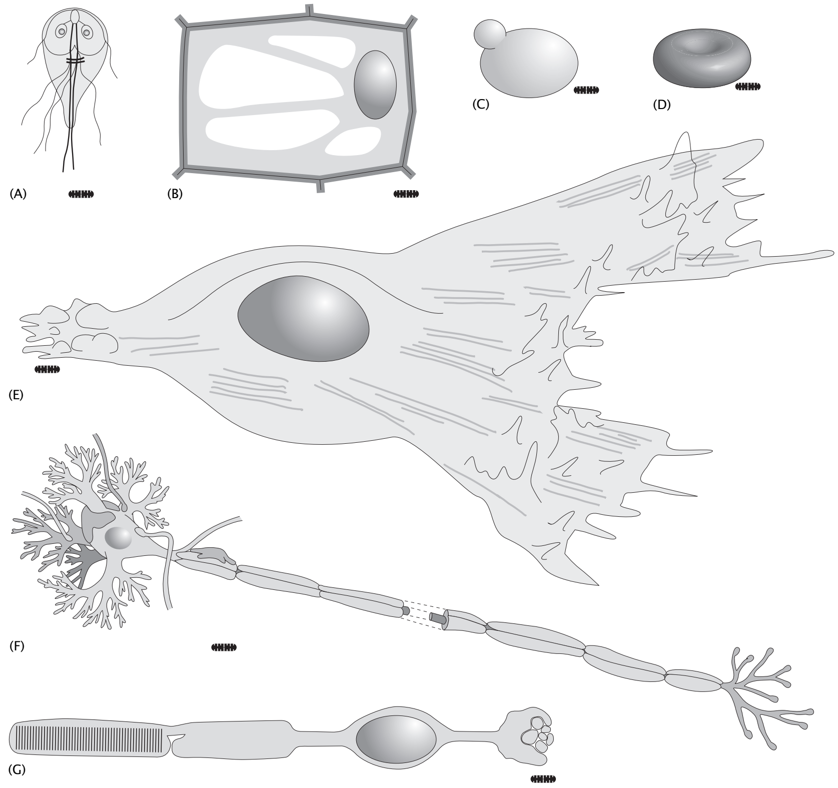


Figure 2.9: Cartoons of several different types of cells all referenced to the standard *E. coli* ruler. (A) the protist *Giardia lamblia*, (B) plant cell, (C) *Saccharomyces cerevisiae*, yeast cell (D) red blood cell, (E) fibroblast cell, (F) eukaryotic nerve cell and (G) rod cell.

wall, often giving them angular structures like that shown in the figure. The typical length scale associated with these cells is often tens of microns. One of the distinctive features of these cell is their large vacuoles within the intracellular space that hold water and contribute to the mechanical properties of plant stems. These large vacuoles are very distinct from animal cells where most of the intracellular space is filled with cytoplasm. Consequently, in comparing a plant and animal cell of similar overall size, the plant cell will have typically tenfold less cytoplasmic volume because most of its intracellular space is filled with vacuoles. Hydrostatic forces matter much more to plants than animals. For example, a wilting flower can be revived simply by application of water since this allows the vacuoles to fill and stiffen the plant stem.

Fungi are even more closely related to us than plants. The representative fungus shown in the figure is the budding yeast *Saccharomyces cerevisiae* (which we will refer to as *S. cerevisiae*). *S. cerevisiae* was domesticated by humans several thousand years ago and continues to serve as a treasured microbial friend that makes our bread rise and provides alcohol in our fermented beverages such as wine. Just as *E. coli* sometimes serves as a key model prokaryotic system, the yeast cell often serves as the model single-celled eukaryotic organism. Besides the fact that humans are fond of *S. cerevisiae* for its own intrinsic properties, it is also useful to biologists as a representative fungus. Of all the other organisms on earth, fungi are closest to animals in terms of evolutionary descent and similarity of protein functions. Although there are no single-celled animals, there are some single celled fungi including *S. cervisiae*. Therefore, many laboratory biological experiments relying on rapid replication of single cells are most easily performed on this organism. Fig. 2.10(A) shows a scanning electron microscope image of a yeast cell engaged in budding. As this image shows, the geometry of yeast is relatively simple compared to many other eukaryotic cells and it is also a fairly small member of this group with a characteristic diameter of roughly 5 microns. Nonetheless, it possesses all the important structural hallmarks of the eukaryotes including, in particular, a membrane bound nucleus, segregating the DNA genome from the cytoplasmic machinery that performs most metabolic function.

Earlier, we estimated the molecular census of an *E. coli* cell. It will now be informative to compare those estimates with the corresponding model eukaryotic cell that will continue to serve as a comparative basis for all our eukaryotic estimates.

- **Estimate: Sizing Up Yeast.** The volume of a yeast cell can be computed in *E. coli* volume units, $V_{E.\ coli}$. In particular, if we recall that $V_{E.\ coli} \approx 1.0\mu\mathrm{m}^3$ and think of yeast as a sphere of diameter 5μm, then we have the relation $V_{yeast} \approx 60 V_{E.\ coli}$, that is, roughly 60 *E. coli* cells would fit inside of a yeast cell. The surface area of a yeast cell can be estimated using a radius of $r_{yeast} \approx 2.5\mu m$ which yields $A_{yeast} \approx 80\mu m^2$. If we treat the yeast nucleus as a sphere with a diameter of roughly $2.0\mu m$, its volume is roughly 4 μm^3. Within this nucleus is housed the 1.2×10^7 bp of the yeast genome which is divided amongst 16 chromosomes. The

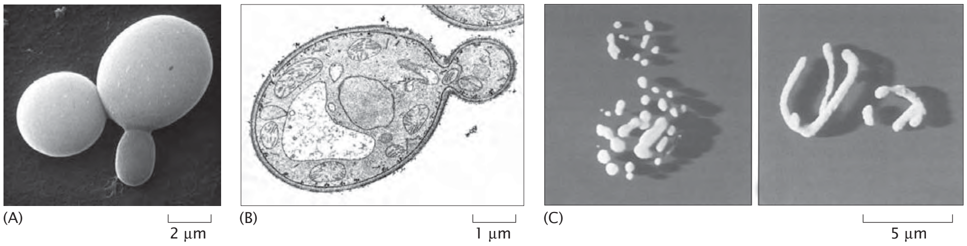


Figure 2.10: Microscopy images of a yeast cell. (A) Scanning electron micrograph of the yeast *Saccharomyces cerevisiae* revealing the overall size scale of these budding yeast. (B) Electron microscopy image of a budding yeast cell. (C) Confocal microscopy images of the mitochondria of *Saccharomyces cerevisiae.*

DNA in yeast is packed into higher order structures mediated by protein assemblies known as histones. In particular, the DNA is wrapped around a series of cylindrical cores made up of eight such histone proteins each, with roughly 150 bp wrapped around each histone octamer, and approximately a 50 bp spacer between. As a result, we can estimate the number of nucleosomes (the histone-DNA complex) as

$$N_{nucleosome} \approx \frac{12 \times 10^6 \text{bp}}{200\text{bp / nucleosome}} \approx 60,000. \tag{2.7}$$

Experimentally, the measured number appears to be closer to 80,000, with a mean spacing between nucleosomes of order 170bp. The total volume taken up by the histones is roughly 150nm^3 per histone (thinking of each histone octamer as a cylindrical disk of radius 3 nm and height 5 nm), resulting in a total volume of 9×10^6 nm^3 taken up by the histones. The volume taken up by the genome itself is comparable at $1.2 \times 10^7 nm^3$, where we have used the rule of thumb that the volume per base pair is 1nm^3. The packing fraction associated with the yeast genomic DNA can be estimated by evaluating the ratio

$$\rho_{pack} \approx \frac{(1.2 \times 10^7 bp) \times (1nm^3/bp)}{4 \times 10^9 nm^3} \approx 3 \times 10^{-3}. \tag{2.8}$$

Note that we have used the fact that the yeast genome is 1.2×10^7 base pairs in length and is packed in the nucleus which has a volume of ≈ 4 μm^3.

These geometric estimates may be used to make corresponding molecular estimates such as the number of lipids and proteins in a typical yeast cell. The number of proteins can be estimated several ways - perhaps the simplest is just to assume that the fractional occupancy of yeast cytoplasm is identical to that of *E. coli* with the result that there will be 60 times as many proteins in yeast as in *E. coli* based strictly on scaling up the cytoplasmic volume. This simple estimate is obtained by *assuming* that the composition of the yeast interior is more or less the same as that of an *E. coli* cell. This strategy results in

$$N_{protein}^{yeast} \approx 60 \times N_{protein}^{E.coli} \approx 2 \times 10^8. \qquad (2.9)$$

The number of lipids associated with the plasma membrane of the yeast cell can be obtained as

$$N_{lipid} \approx \frac{2 \times 0.5 \times A_{yeast}}{A_{lipid}} \approx \frac{2 \times 0.5 \times (80 \times 10^6 nm^2)}{.25 nm^2} \approx 4 \times 10^8, \quad (2.10)$$

where the factor of 0.5 is based on the idea that roughly half of the surface area is covered by membrane proteins rather than lipids themselves and the factor of 2 accounts for the fact that the membrane is a bilayer.

Another interesting estimate suggested by fig. 2.10(C) is associated with the organellar content of these cells. In particular, this figure shows the mitochondria of yeast which are being grown in two different media. These pictures suggest several interesting questions such as what fraction of the cellular volume is occupied by mitochondria and what is the surface area tied up with the mitochondrial outer membranes? The number of mitochondria in the image can be estimated several ways - one of which is to attempt to count them directly, the other of which is to estimate their mean spacing and to compute the corresponding density and number. Using the latter method results in an estimate of roughly 40 mitochondria in the image on the left of fig. 2.10(C). Further, we estimate that the typical mitochondrial size is roughly 3/4 μm, resulting in a total mitochondrial volume of

$$V_{mito} \approx 40 \times \frac{4\pi}{3} (\frac{3}{8})^3 \mu m^3 \approx 9 \mu m^3, \qquad (2.11)$$

which given the total volume of the cell of 60 μm^3 translates into a volume fraction of roughly 15 percent. The total area of the outer membranes of these mitochondria is roughly 70 μm^2, comparable to the entire area of the plasma membrane itself. The analysis of the image on the right is left as an exercise for the reader in the problems.

Our estimates are brought into sharpest focus when they are juxtaposed with actual measurements. The census of yeast cells has been performed in several distinct and fascinating ways in recent years. The key idea is to generate thousands of different yeast strains, each of which has a tag on a different one of the yeast gene products. For example, it is possible to generate strains with a peptide fragment that can then be recognized by antibodies. A second scheme is to construct protein fusions in which the protein of interest is attached to a fluorescent protein such as the green fluorescent protein (GFP). Then, by querying each and every cell either by examining the extent of antibody binding or fluorescence, it is possible to count up the numbers of each type of protein. Fig. 2.11 shows a histogram of the number of proteins that occur with a given

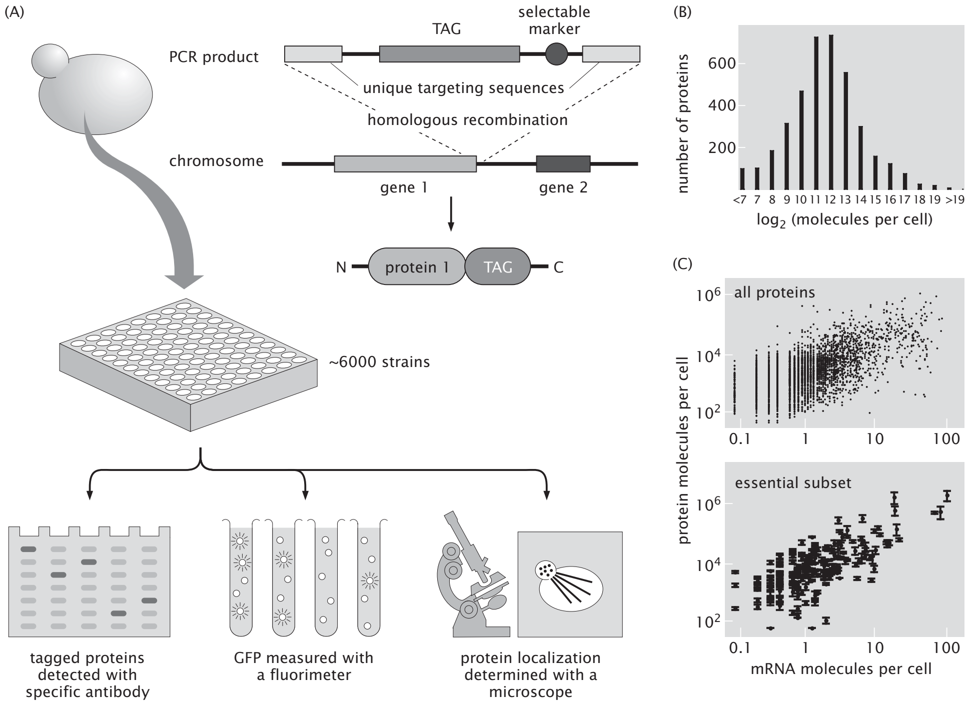


Figure 2.11: Protein copy numbers in yeast. (A) Result of antibody detection of various proteins in yeast showing the number of proteins that have a given copy number. The number of copies of the protein is expressed in powers of 2 as 2^N. (B) The mean number of proteins associated with various processes within cells. (C) The mean number of proteins associated with different spatial compartments in the cell.

protein copy number. By adding up the total number of proteins on the basis of this census, we estimate there are 50×10^6 proteins in a yeast cell, somewhat less than suggested by our crude estimate given above.

The remainder of the cells in fig. 2.9 are all human cells and show another interesting aspect of cellular diversity. To a first approximation, every cell in the human body contains the same DNA genome. And yet, individual human cells differ significantly with respect to their sizes (with sizes varying from roughly 5 microns to 1 meter for the largest neurons), shapes and functions. For example, rod cells in the retina are specialized to detect incoming light and transmit that information to the neural system so that we can see. Red blood cells are primarily specialized as carriers of oxygen and, in fact, are dramatically different from almost all other cells in having dispensed with their nucleus as part of their developmental process. As we will discuss extensively throughout the book, other cells have specializations.

Another important example of the structural diversity of cells, this time from animals, is the red blood cell shown in fig. 2.9(D). Note that the shapes of red blood cells are decidedly not spherical raising interesting questions about the mechanisms of cell-shape maintenance. Despite their characteristic size of order 5 microns, these cells easily pass through capillaries with less than half their diameter as shown in fig. 2.12, implying that their shape is altered significantly as part of their normal life cycle. While in capillaries (either artificial or *in vivo*), the red blood cell is severely deformed to pass through the narrow passage. In their role as the transport vessels for oxygen-rich hemoglobin, these cells will serve as an inspiration for our discussion of the statistical mechanics of cooperative binding. Red blood cells are a target of one of the most common infectious diseases suffered by humans caused by the invasion of a protozoan. Malaria infected red blood cells are much stiffer than normal cells and cannot deform to enter small capillaries. Consequently, people suffering from malaria experience severe pain and damage to tissues because of the inability of their red blood cells to enter those tissues and deliver oxygen.

One of the favorite eukaryotic cells from multicellular organisms is the fibroblast as shown schematically in fig. 2.9(E) and shown in an AFM image in fig. 2.13. These cells will serve as a centerpiece for much of what we will have to say about "typical" eukaryotic cells in the remainder of the book. Fibroblasts are associated with animal connective tissue and are notable for secreting the macromolecules of the extracellular matrix.

Cells in multicellular organisms can be even more exotic. For example, nerve cells (fig. 2.9(F)) and rod cells (fig. 2.9(G)) reveal a great deal more complexity than the examples highlighted above. In these cases, the cell shape is intimately related to their function. In the case of nerve cells, their sinewy appearance is tied to the fact that the various branches (also called "processes") known as dendrites and axons convey electrical signals which permit communication between distant parts of an animal's nervous system. Despite having nuclei with typical eukaryotic dimensions, the cells themselves can extend processes with characteristic lengths of tens of centimeters or even more. The structural complexity of rod cells is tied to their primary function of light detection in the retina of the eye. These cells are highly specialized to perform transduction of light energy into chemical energy that can be used to communicate with other cells in the body and in particular, with brain cells that permit us to be conscious of perceiving images. Rod cells accomplish this task using large stacks of membranes which are the antennas participating in light detection. Fig. 2.9 only scratches the surface of the range of cellular size and shape, but at least conveys an impression of cell sizes relative to our standard ruler.

2.2.2 The Cellular Interior: Organelles

As we descend from the scale of the cell itself, a host of new structures known as organelles come into view. The presence of these membrane-bound organelles is one of the defining characteristics that distinguishes eukaryotes from bacteria and archaea. Fig. 2.14 shows a schematic of a eukaryotic cell and associated

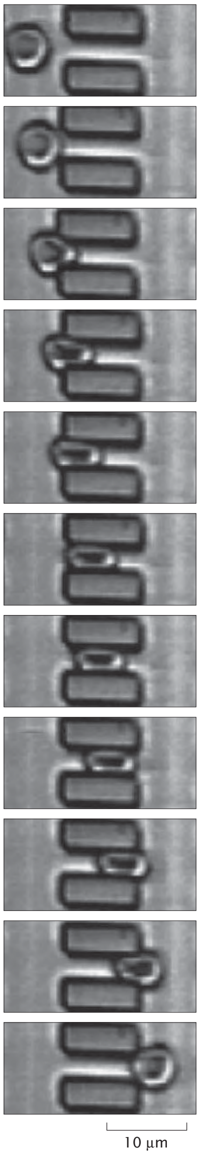


Figure 2.12: Deformability of red blood cells. To measure the deformability of human red blood cells, an array of blocks was fabricated in silicon, each block was 4 × 4 × 12 microns. The blocks were spaced by 4 microns in one direction and 13 microns in the other. A glass coverslip covered the top of this array of blocks. A dilute suspension of red blood cells in a saline buffer was introduced to the system. A slight pressure applied at one end of the array of blocks provided bulk liquid flow, from left to right in the figure. This liquid flow carried the red blood cells throught the narrow passages. Video microscopy captured the results. The figure shows consecutive video fields with the total elapsed time just over one third of a second. (courtesy of James Brody)

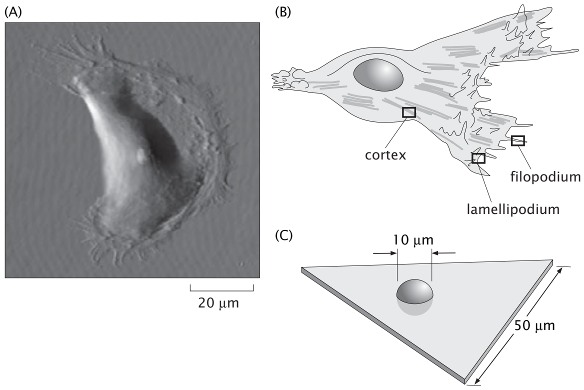


Figure 2.13: Structure of a fibroblast. (A) Atomic-force microscopy image of a fibroblast (courtesy of Manfred Radmacher). (B) cartoon of the external morphology of a fibroblast, (C) characteristic dimensions of the morphology of a fibroblast.

images of some of the key organelles. These organelles serve as the specialized apparatus of cell function, serving in capacities ranging from genome management (the nucleus) to energy generation (mitochondria and chloroplasts) to protein synthesis and modification (endoplasmic reticulum and Golgi apparatus) and beyond. The compartments that are bounded by organellar membranes can have completely different protein and ion compositions. In addition, the membranes of each of these different membrane systems are characterized by distinct lipid and protein compositions.

A characteristic feature of many organelles is that they are compartmentalized structures that are separated from the rest of the cell by membranes. The nucleus is one of the most familiar examples since it is often easily visible using standard light microscopy. If we use the fibroblast as an example, then the cell itself has dimensions of roughly 50 microns, while the nucleus has a characteristic linear dimension of roughly 10 microns as shown schematically in fig. 2.13. From a functional perspective, the nucleus is much more complex than simply serving as a storehouse for the genetic material. Chromosomes are organized within the nucleus forming specific domains as will be discussed in more detail in chap. 8. Transcription occurs in the nucleus as well as several kinds of RNA processing. There is a busy traffic of molecules such as transcription factors moving in and completed RNA molecules moving out through elaborate gateways in the nuclear membrane known as nuclear pores. Portions of the genome involved in synthesis of ribosomal RNA are clustered together forming striking spots that can be seen in the light microscope and called nucleoli.

Moving outward from the nucleus, the next membraneous organelle we encounter is often the endoplasmic reticulum. Indeed, the membrane of the nuclear envelope is contiguous with the membrane of the nuclear envelope. In some cells

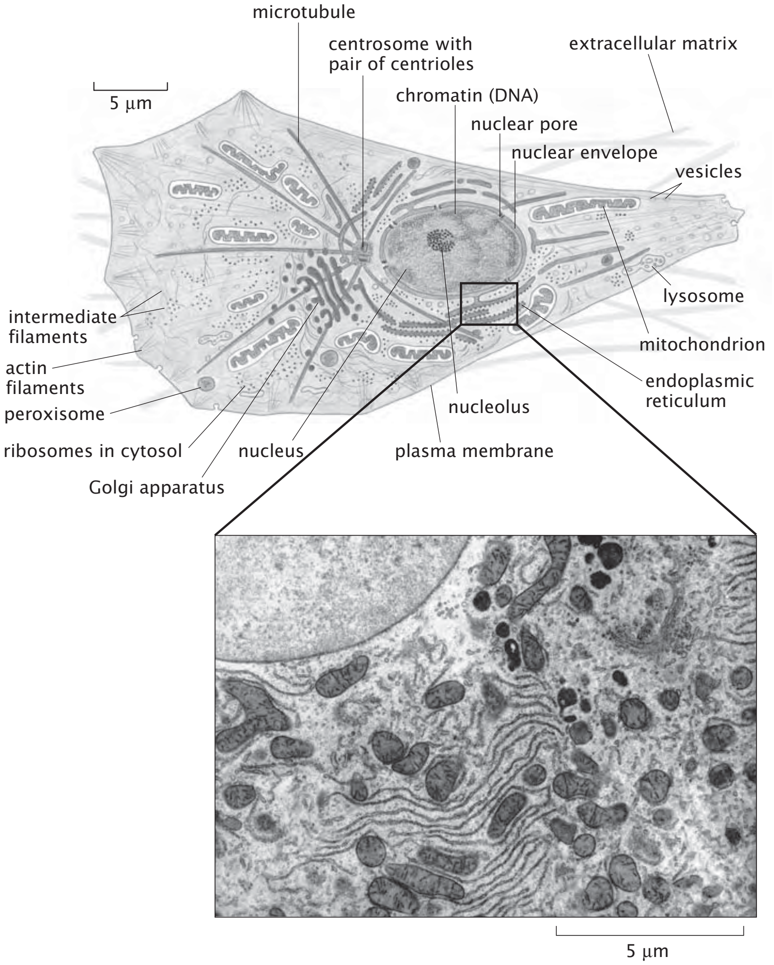


Figure 2.14: Eukaryotic cell and its organelles. The schematic shows a eukaryotic cell and a variety of membrane bound organelles. A thin-section electron microscopy image shows a portion of a rat liver cell approximately equivalent to the boxed area on the schematic. A portion of the nucleus can be seen in the upper left corner. The most prominent organelles visible in the image are mitochondria, lysosomes, the rough endoplasmic reticulum and the Golgi apparatus. (adapted from Fawcett, 1966)

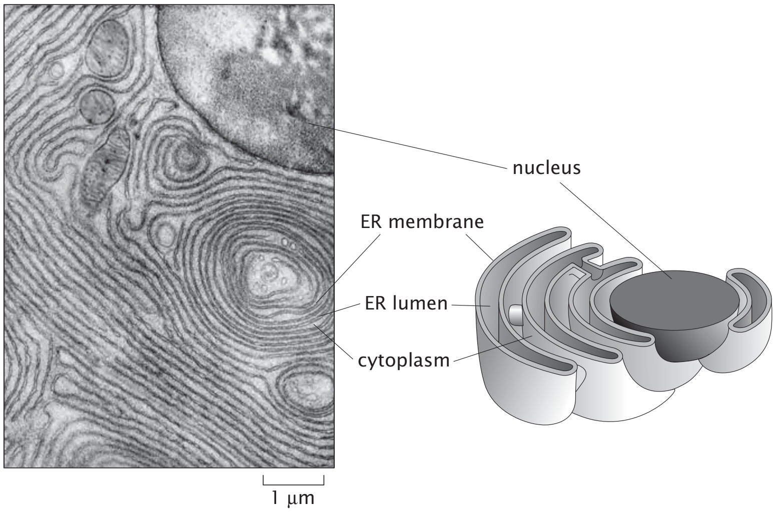


Figure 2.15: Electron micrograph and associated schematic of the endoplasmic reticulum. The left panel shows a thin-section electron micrograph of an acinar cell from the pancreas of a bat. The nucleus is visible at the upper right and the dense and elaborate ER structure is strikingly evident. The right panel shows a schematic diagram of a model for the three-dimensional structure of the ER in this cell. Notice that the size of the lumen in the ER in the schematic is exaggerated for ease of interpretation. Electron micrograph from Fawcett, 1966.

such as the pancreatic cell shown in fig. 2.15, the endoplasmic reticulum takes up the bulk of the cell interior. This elaborate organelle is the site of lipid synthesis and also the site of synthesis of proteins that are destined to be secreted or incorporated into membranes. From images such as those in fig. 2.15 and 2.16 it is clear that the ER can assume different geometries in different cell types and under different conditions. How much total membrane area is taken up by the ER? How strongly does the specific membrane morphology affect the total size of the organelle?

- **Estimate: Membrane Area of the Endoplasmic Reticulum.** One of the most compelling structural features of the endoplasmic reticulum is its enormous surface area. To estimate the area associated with the endoplasmic reticulum, we take our cue from fig. 2.15 which suggests that we think of the ER as a series of concentric spheres centered about the nucleus. We follow Fawcett (1966) who characterizes the ER as forming "lamellar systems of flat cavities, rather uniformly spaced and parallel to one another" as shown in fig. 2.15.

  An estimate can be made by adding up the areas from each of the concentric spheres making up our model ER. This can be done by simply

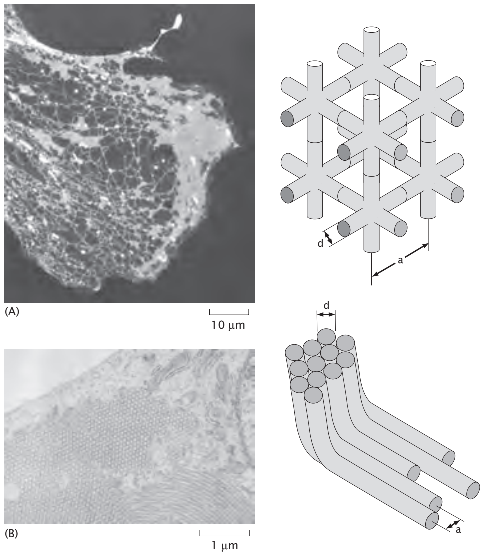


Figure 2.16: Variable morphology of the ER. (A) In most cultured cells, the ER is a combination of a web-like reticular network of tubules and larger flattened cisternae. In this image, a cultured fibroblast was stained with a fluorescent dye called DiOC6 that specifically labels ER membrane. On the right is a schematic of an idealized three-dimensional reticular network. (B) Some specialized cells and those treated with drugs that upregulate the synthesis of lipids reorganize their ER to form tightly-packed, nearly crystalline arrays that resemble piles of pipes.

noticing that the volume enclosed by the ER can be written as

$$V_{\mathrm{ER}} = \sum_i A_i d \ , \tag{2.12}$$

where A_i is the area of the i^{th} concentric sphere and d is the distance between adjacent cisternae. Since two membranes bound each cisterna the total area of the ER membrane is $A_{\mathrm{ER}} = 2 \times \sum_i A_i$. In our model, the total volume of the ER can be written as the difference between the volume taken up by the outermost sphere and the volume of the innermost concentric sphere, which is the same as the volume of the nucleus:

$$V_{\mathrm{ER}} = \frac{4\pi}{3} R_{out}^3 - \frac{4\pi}{3} R_{nucleus}^3 \ . \tag{2.13}$$

Combining the two ways of computing the volume of the ER, eqns. 2.12 and 2.13, we arrive at an expression for the ER area,

$$A_{\mathrm{ER}} = \frac{8\pi}{3d}(R_{out}^3 - R_{nucleus}^3) \ . \tag{2.14}$$

Using the values $R_{\mathrm{nucleus}} = 5\mu$m, $R_{\mathrm{out}} = 10\mu$m and $d = 0.05\mu$m, we get at an estimate $A_{\mathrm{ER}} = 15 \times 10^4 \mu\mathrm{m}^2$. This result should be contrasted with a crude estimate for the area of a fibroblast which can be obtained by using the dimensions in fig. 2.13(c) and which yields an area of $10^4\mu\mathrm{m}^2$ for the cell membrane itself. To estimate the area of the ER when it is in reticular form we describe its structure as interpenetrating cylinders of diameter $d \approx 10$nm separated by a distance $a \approx 60$nm, as shown in fig. 2.16. The completion of the estimate is left to the problems, but results in a comparable membrane area.

The other major organelles found in most cells and visible in fig. 2.14 include the Golgi apparatus, mitochondria and lysosomes. The Golgi apparatus, similar to the ER, is largely involved in processing and trafficking of membrane-bound and secreted proteins. The Golgi apparatus is typically seen as a pancake-like stack of flattened compartments, each of which contains a distinct set of enzymes. As proteins are processed for secretion, for example by addition and remodeling of attached sugars, they appear to pass in an orderly fashion through each element in the Golgi stack. The mitochondria are particularly striking organelles with a smooth outer surface housing an elaborately folded system of internal membrane structures. The mitochondria are the primary site of ATP synthesis for cells growing in the presence of oxygen, and their physiology as well as their structure are fascinating and have been well studied. We will return to the topic of mitochondrial structure in chap. 11 and discuss the workings of the tiny machine responsible for ATP synthesis in chap. 16. Lysosomes serve a major role in the degradation of cellular components. In some specialized cells such as macrophages, lysosomes also serve as the compartment where bacterial invaders can be degraded. These membrane-bound organelles are filled with acids, proteases and other degradative enzymes. Their shapes are polymorphous; resting lysosomes are simple and nearly spherical, whereas lysosomes actively involved in degradation of cellular components or of objects taken in from the outside may be much larger and complicated in shape.

These common organelles are only a few of those that can be found in eukaryotic cells. Some specialized cells have remarkable and highly specialized organelles that can be found nowhere else such as the stacks of photoreceptive membranes found in the rod cells of the visual system and as indicated schematically in fig. 2.9. The common theme is that all organelles represent specialized subcompartments of the cell that perform a particular subset of cellular tasks and represent a smaller, discrete layer of organization one step down from the whole cell.

2.2.3 Macromolecular Assemblies: The Whole is Greater than the Sum of the Parts

Macromolecules Come Together to Form Assemblies (Somes)

Proteins, nucleic acids, sugars and lipids often work as a team. Indeed, as will become clear throughout the remainder of the book, these macromolecules often come together to make assemblies, often dubbed "somes". We think of yet another factor of ten magnification relative to the previous section, and with this increase of magnification we see assemblies such as those shown in cartoon form in fig. 2.17. The genetic material in the eukaryotic nucleus is organized into chromatin fibers which themselves are built up of protein-DNA assemblies known as nucleosomes. The replication complex that copies DNA before cell division is similarly a collection of molecules which has been dubbed the replisome. When the genetic message is exported to the cytoplasm for translation into proteins, the ribosome (an assembly of proteins and nucleic acids) serves as the universal translating machine that converts the nucleic acid message from the RNA into the protein product written in the amino acid alphabet. The production of ATP in mitochondria is similarly mediated by a macromolecular complex known as ATP synthase. When proteins have been targeted for degradation, they are sent to another macromolecular assembly known as the proteasome. The key idea of this subsection is to show that there is a very important level of structure in cells that is built around complexes of individual macromolecules (loosely designated as *somes*) and with a characteristic length scale of 10nm.

Helical Motifs Are Seen Repeatedly in Molecular Assemblies

A second class of macromolecular assemblies, characterized not by function but rather by structure is the wide variety of helical macromolecular complexes. Several representative examples are shown in fig. 2.18. In fig. 2.18(a), we show the geometric structure of microtubules. As will be described in more detail later, these structures are built up of individual protein units called tubulin. A second example shown in fig. 2.18(b) is the bacterial flagellum of *E. coli.* Here too, the same basic structural idea is repeated with the helical geometry built up from individual protein units, in this case flagellin. The third example given

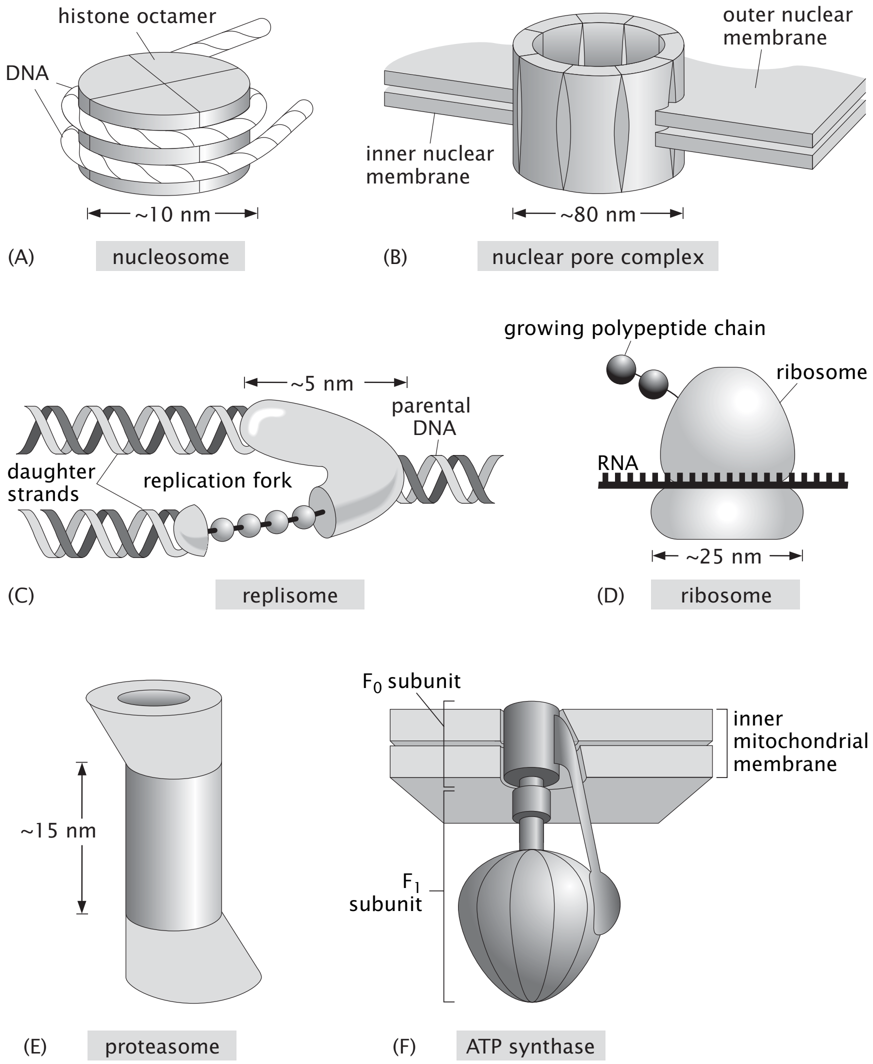


Figure 2.17: The macromolecular assemblies of the cell.

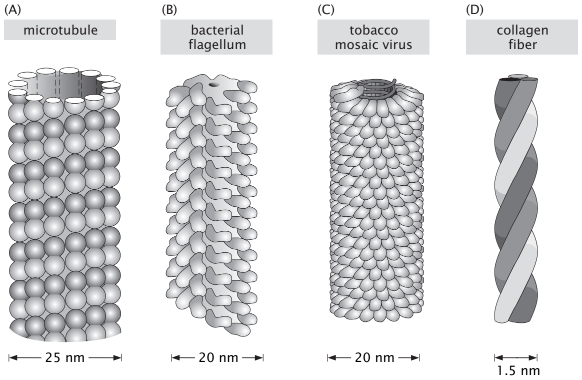


Figure 2.18: Helical assemblies of the cell. Cells have a variety of different helical assemblies, some formed from individual monomeric units (such as (A)-(C)) and others resulting from coils of proteins.

in the figure is that of a filamentous virus, with tobacco mosaic virus (TMV) chosen as one of the most well studied of viruses.

The helical assemblies described above are characterized by individual protein units which come together to form helical filaments. An alternative and equally remarkable class of filaments are those in which alpha helices (chains of amino acids forming protein subunits with a precise, helical geometry) wind around each other to form superhelices. The particular case study which will interest most in subsequent discussions is that of collagen which serves as one of the key components in the extracellular matrix of connective tissues and is one of the majority protein products of the fibroblast cells introduced earlier in the chapter (see fig. 2.9).

Macromolecular Assemblies Are Arranged in Superstructures

Assemblies of macromolecules can interact with each other to create striking instances of cellular hardware with a size comparable to organelles themselves. Fig. 2.19 shows several examples. Fig. 2.19(A) shows the way in which ribosomes are organized on the endoplasmic reticulum with a characteristic spacing which is comparable to the size of the ribosomes ($\approx 20nm$). A second stunning example is the organization of myofibrils in muscles as shown in fig. 2.19(B). This figure shows the juxtaposition of the myofibrils and mitochondria. The myofibrils themselves are an ordered arrangement of actin filaments and myosin motors as will be discussed in more detail in chap. 16. The last example shown in fig. 2.19(C) is of the protrusions of microvilli at the surface of an epithelial cell. These microvilli are the result of collections of parallel actin filaments. The list of examples of orchestration of collections of macromolecules can go on and on.

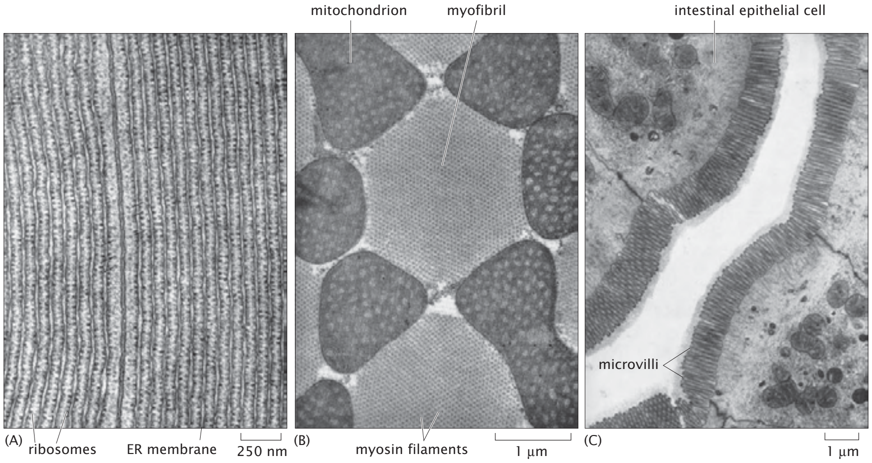


Figure 2.19: Ordered macromolecular assemblies. Collage of examples of macromolecules organized into superstructures. (A) ribosomes on the endoplasmic reticulum, (B) myofibrils in the flight muscle, (C) microvilli at the epithelial surface.

2.2.4 Viruses as Assemblies

Viruses are one of the most impressive and beautiful class of macromolecular assembly. These assemblies are a collection of proteins and nucleic acids (though many viruses have lipid envelopes as well) that form highly ordered and symmetrical objects with characteristic sizes of 10s to 100s of nanometers. The architecture of these viruses is usually a protein shell where the so-called capsid is made up of a repetitive packing of the same protein subunits over and over to form an icosahedron. Within the capsid, the virus packs its genetic material which can be either DNA or RNA depending upon the type of virus. Fig. 2.20 is a gallery of the capsids of a number of different viruses. Different viruses have different elaborations on this basic structure and can include lipid coats, surface receptors, and internal molecular machines such as polymerases and proteases. One of the most amazing features of these viruses is that by hijacking the host cell, the viral genome commands the construction of its own inventory of parts within the host and then in the crowded environment of that host, assembles into these beautiful and subtle killing machines.

HIV (human immunodeficiency virus) is one of the viruses that has garnered the most attention in recent years. Fig. 2.21 shows cryo-EM images of mature HIV virions and gives a sense of both their overall size (roughly 130 nm) and their internal structure. In particular, note the presence of an internal capsid shaped like an ice-cream cone. This internal structure houses the roughly 10kb RNA viral genome. As with our analysis of the inventory of a cell considered earlier in the chapter, part of developing a “feeling for the organism” is to get

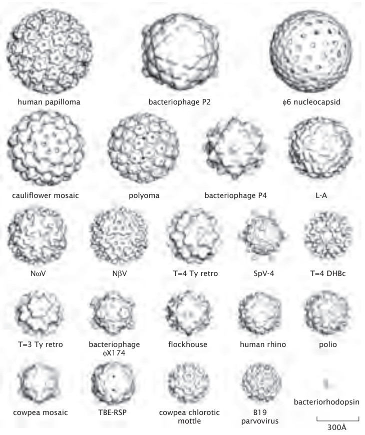


Figure 2.20: Structures of viral capsids. The regularity of the structure of viruses has enabled detailed, atomic-level analysis of their construction patterns. This gallery shows a variety of the different geometries explored by the class of nearly spherical viruses. For size comparison, a large protein bacteriorhodopsin is shown in the bottom right.

a sense of the types and numbers of the different molecules that make up that organism. In the case of HIV, these numbers are interesting for many reasons, including that they say something about the "investment" that the infected cell has to make in order to construct new virions.

For our census of an HIV virion, we need to examine the assembly of the virus. In particular, one of the key products of its roughly 10kb genome is a polyprotein known as Gag and shown schematically in fig. 2.22. The formation of the *immature* virus occurs through the association of the N-terminal ends of these Gag proteins with the lipid bilayer of the host cell and the C-termini pointing radially inward like the spokes of a three-dimensional wheel. As more of these proteins associate on the cell surface, the nascent virus begins to form a bud on the cell surface ultimately resulting in spherical structures like those shown in fig. 2.22. During the process of viral maturation, a viral protease (an enzyme that cuts proteins) clips the Gag protein into its component pieces known as matrix (MA), capsid (CA), nucleocapsid (NC) and p6. The matrix forms a shell of proteins just inside of the lipid bilayer coat. The capsid proteins form the ice-cream cone shaped object that houses the genetic material and the nucleocapsid protein is complexed with the viral RNA.

- **Estimate: Sizing Up HIV.** Unlike many of their more ordered viral counterparts, HIV virions have the intriguing feature that the structure from one to the next is not exactly the same. Indeed, they come in both different shapes and sizes. As a result, our attempt to "size up" HIV

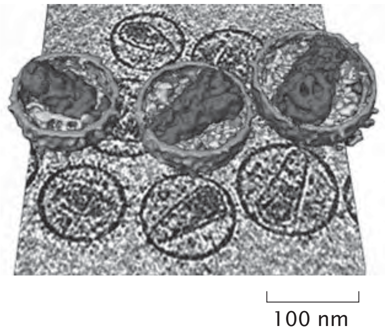


Figure 2.21: Structure of HIV viruses. The planar image shows a single frame from an electron microscopy tilt series. The three-dimensional images show reconstructions of the mature viruses featuring the ice-cream cone shaped capsid on the interior. figure from Briggs, Structure 2006

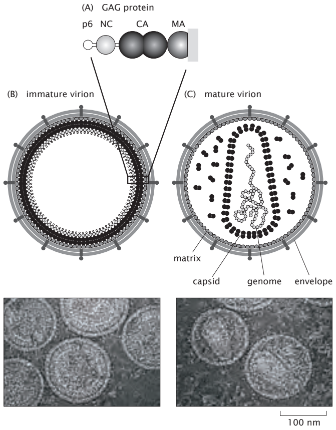


Figure 2.22: HIV architecture. (A) Schematic of the Gag polyprotein, a 41,000 Da architectural building block. (B) Immature virions showing the lipid bilayer coat and the uncut Gag shell on the interior, (C) mature virions in which the Gag protein has been cut by proteases and the separate components have assumed their architectural roles in the virus. The associated electron microscopy images show actual data for each of the cartoons. (adapted from Briggs *et al.*).

will be built around some representative numbers for these viruses, but the reader is cautioned to think of a statistical distribution of sizes and shapes. As shown in the cryo-EM picture of fig. 2.21, the size of the virion is between 120 nm and 150 nm and we take a “canonical” size of 130nm.

We begin with the immature virion. To find the number of Gag proteins within a given virion, we resort to simple geometrical reasoning. Since the radius of the overall virion is roughly 65nm, and the outer 5nm of that radius is associated with the lipid bilayer, we imagine a sphere of radius 60nm that is decorated on the inside with the inward facing spokes of the Gag proteins. If we think of each such Gag protein as a cylinder of radius 2 nm, this means they take up an area $A_{Gag} \approx 4\pi$ nm^2. Using this, we can find the number of such Gag proteins as

$$N_{Gag} = \frac{\text{surface area of virion}}{\text{area per Gag protein}} \approx \frac{4\pi(60\text{ nm})^2}{4\pi\text{ nm}^2} \approx 3500. \tag{2.15}$$

The total mass of these Gag proteins is roughly

$$M_{Gag} \approx 3500 \times 41,000 Da \approx 150 MDa, \tag{2.16}$$

where we have used the fact that the mass of each Gag polyprotein is roughly 40 kDa. This estimate for the number of Gag proteins is of precisely the same magnitude as those that have emerged from recent cryo-electron microscopy observations.

The number of lipids associated with the HIV envelope can be estimated similarly as

$$N_{lipids} \approx \frac{2 \times 4\pi(65\text{ nm})^2}{1/2\text{ nm}^2} \approx 200,000\text{ lipids}, \tag{2.17}$$

where the factor of 2 accounts for the fact that the lipids form a bilayer, and we have used a typical area per lipid of 1/2 nm^2. The lipid census of HIV has been taken using mass spectrometry which permits the measurement of each of the different types of lipids forming the viral envelope. Interestingly, the diversity of lipids in the HIV envelope is enormous with the lipid composition of the viral envelope distinct from that of the host cell membrane. The measured total number of different lipids is roughly 300,000. Further analysis of the parts list of HIV is left to the problems at the end of the chapter.

Ultimately, viruses are one of the most interesting classes of macromolecular assembly. These intriguing machines occupy a fuzzy zone at the interface between the living and the nonliving.

2.2.5 The Molecular Architecture of Cells: From PDB Files to Ribbon Diagrams

If we continue with another factor of ten in our powers of ten descent, we find the individual macromolecules of the cell. In particular, this increase in spatial resolution reveals four broad categories of macromolecule: lipids, carbohydrates, nucleic acids and proteins. As was shown in chap. 1, these four classes of molecule make up the stuff of life and have central status in making up cells both architecturally and functionally. Though often these molecules are highly anisotropic (for example, a DNA molecule is usually many orders of magnitude longer than it is wide), their characteristic scale is between one and ten nanometers. For example, as shown earlier in the chapter, a "typical" protein has a size of several nanometers. Lipids are more anistropic with lengths of 2-3 nm and cross-sectional areas of roughly $1/2\ \ \text{nm}^2$.

The goal of this section is to provide several different views of the molecules of life and how they fit into the structural hierarchy described throughout the chapter.

Macromolecular Structure Is Characterized Fundamentally By Atomic Coordinates

The conjunction of X-ray crystallography, nuclear magnetic resonance and cryo-electron microscopy have revealed the atomic-level structures of a dazzling array of macromolecules of central importance to the function of cells. The list of such structures includes molecular motors, ion channels, DNA-binding proteins, viral capsid proteins and various nucleic acid structures too. The determination of new structures is literally a daily experience. Indeed, as will be asked of the reader in the problems at the end of the chapter, a visit to websites such as the Protein Data Bank or VIPER reveals just how many molecular and macromolecular structures are now known.

Though the word structure can mean different things to different people (indeed, that is one of the primary messages of this chapter and chap. 8), at the level of structural biology, the determination of structure ultimately refers to a list of atomic coordinates for the various atoms making up the structure of interest. As an example, fig. 1.1 introduced detailed atomic portraits of nucleic acids, proteins, lipids and sugars. In such descriptions, the structural characterization of the system amounts to a set of coordinates

$$\mathbf{r}_i = x_i\mathbf{i} + y_i\mathbf{j} + z_i\mathbf{k}, \tag{2.18}$$

where, having chosen some origin of coordinates, the coordinates of the i^{th} atom in the structure are given by (x_i, y_i, z_i). That is, we have some origin of Cartesian coordinates and every atomic position is an address on this three-dimensional grid.

Because the macromolecules of the cell are subject to incessant jiggling due to collisions with each other and the surrounding water, a static picture of structure is incomplete. The structural snapshots embodied in atomic coordinates for a given structure miss the fact that each and every atom is engaged in a constant thermal dance. Hence, the coordinates of eqn. 2.18 are really of the form

$$\mathbf{r}_i(t) = x_i(t)\mathbf{i} + y_i(t)\mathbf{j} + z_i(t)\mathbf{k}, \tag{2.19}$$

where the t reminds us that the coordinates depend upon time and what is

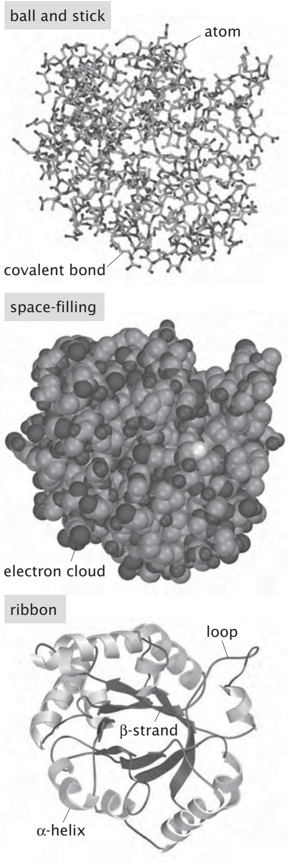


Figure 2.23: Three representations of triose phosphate isomerase. This enzyme is one of the enzymes in the glycolysis pathway.

measured in experiments might be best represented as $\langle \mathbf{r}_i(t) \rangle_{time}$, where the brackets $\langle\rangle_{time}$ signify an average over time.

An example of an atomic-level representation of one of the key proteins of the glycolysis pathway is shown in fig. 2.23. We choose this example because glycolysis will arise repeatedly throughout the book as a canonical metabolic pathway. The figure also shows several alternative schemes for capturing these structures such as using ribbon-diagrams which highlight the ways in which the different amino acids come together to form elements of secondary structure such as alpha helices and beta sheets.

Chemical Groups Allow Us to Classify Parts of the Structure of Macromolecules

When thinking about the structures of the macromolecules of the cell, one of the most important ways to give those structures functional meaning (as opposed to just a collection of coordinates) is through reference to the chemical groups that make them up. For example, the structure of the protein shown in figs. 1.1 and 2.23 is not just an arbitrary arrangement of carbons, nitrogens, oxygens and hydrogens. Rather, this structure reflects the fact that the protein is made up of a linear sequence of amino acids which each have their own distinct identity as shown in fig. 2.24. The physical and chemical properties of these amino acids dictate both the folded shape of the protein as well as how it functions.

Amino acids are but one example of a broader class of nanometer-scale structural building blocks known as "chemical groups". These chemical groups occur with great frequency in different macromolecules and, like the amino acids, each have their own unique chemical identity. We identify such groups with a roughly context-independent chemical behavior. Fig. 2.25 shows a variety of chemical groups that are of interest in biochemistry and molecular biology. These are all biologically important chemical functional groups that can be attached to a carbon atom as shown in fig. 2.25 and are all found in protein structures. The methyl and phenyl groups contain only carbon and hydrogen and are hence hydrophobic (unable to form hydrogen bonds with water). To the right of these are shown two chemically similar groups, alcohol and thiol consisting of oxygen or sulfur plus a single hydrogen. The key feature of these two groups is that they are highly reactive and can participate in chemical reactions forming new covalent bonds. Amino acids containing these functional groups (serine, threonine, tyrosine and cysteine) are frequently important enzyme residues in catalytic reactions. The next row starts with a nitrogen containing amino group which is usually postively charged at neutral pH and a negatively charged carboxylic acid. All amino acids in monomeric form have both of these groups. In a protein polymer, there is a free amino group at the N-terminus of the protein and a free carboxylic acid group at the C-terminus of the protein. Several amino acids also contain these groups as part of their sidechains and the charge-based interactions are frequently responsible for chemical specificity in molecular recognition as well as some kinds of catalysis. An amide group is shown next. This group is not generally charged but is able to participate in a variety of hydrogen bonds. The last group shown is a phosphoryl group which is not part of any amino acid that is incorporated by the ribosome in a polypeptide chain during translation. On the other hand, these groups are frequently added to proteins as a post-translational modification and perform extremely important regulatory functions.

Nucleic acids can similarly be thought of from the point of view of chemical groups. In fig. 1.3, we showed the way in which individual groups can be seen as the building blocks of DNA structures such as that shown in fig. 1.1(a). In particular, we note that the backbone of the double helix is built up of sugars (represented as pentagons) and phosphates. Similarly, the nitrogenous bases which mediate the pairing between the complementary strands of the backbone are represented diagrammatically via hexagons and pentagons, with hydrogen bonds depicted as shown in the figure.

This brief description of the individual molecular units of the machines of

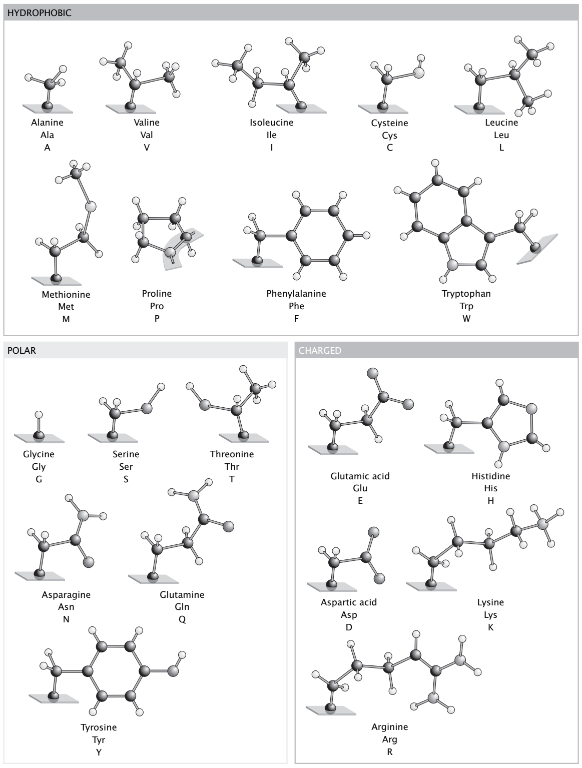


Figure 2.24: Amino acid side chains. The amino acids are represented here in ball and stick form, where a black ball indicates a carbon atom, a small white ball indicates a hydrogen atom and a gray ball, oxygen, nitrogen or sulfur. Only the side chains are shown. The peptide backbone of the protein to which these sidechains are attached is indicated by a flat gray tile. The amino acids are subdivided based upon their physical properties. The group shown at the top are hydrophobic and tend to be found on the interior of proteins. Those at the bottom are able to form hydrogen bonds with water.

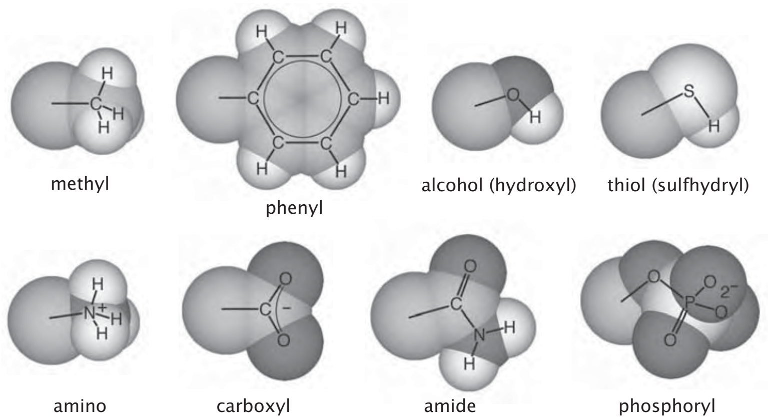


Figure 2.25: Chemical groups. These are some of the most common groups found in organic molecules such as proteins.

the cell brings us to the end of our powers of ten descent which examine the structures of the cell. Our plan now is to zoom out from the scale of individual cells to examine the structures they form together.

2.3 Telescoping Up in Scale: Cells Don't Go It Alone

2.3.1 Multicellularity As One of Evolution's Great Inventions

Our powers of ten journey has thus far shown us the way in which cells are built from structural units going down from organelles to macromolecular assemblies to individual macromolecules to chemical groups, atoms and ions. Equally interesting hierarchies of structures are revealed as we reduce the resolution of our imaginary camera and zoom out from the scale of individual cells. What we see once we begin to zoom out from the scale of single cells is the emergence of communities in which cells do not act independently.

Life has been marked by several different evolutionary events which wrought a wholesale change in the way that cells operate. One important category of such events is the acquisition of the ability of cells to communicate and cooperate with one another to form multicellular communities with common goals. This has happened many times throughout all branches of life and has culminated in extremely large organisms such as redwood trees and giraffes among eukaryotes. In this section, we explore the ways in which cell-cell communication and cooperation have given rise to new classes of biological structures. Fig. 2.26 shows a variety of different examples of cellular communities, some of

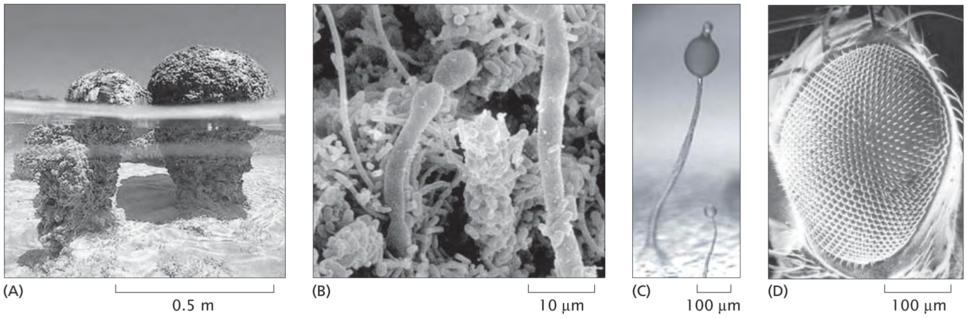


Figure 2.26: Representative examples of different communities of cells. (A) Fossils of the ancient bacterial colonies known as stromatolites, (B) Bacterial biofilm. (C) The social amoeba *Dictyostelium discoideum* forms fruiting bodies. The picture shows a fruiting body with spores - the tall stalk with a bulb at the top is a collection of amoebae. (D) The *Drosophila* eye.

which form the substance of the remainder of the chapter.

Bacteria Interact to Form Colonies Such as Biofilms

The oldest known cellular communities recorded in the fossil record are essentially gigantic bacterial colonies called stromatolites such as those shown in fig. 2.26(A). These fossils have a characteristic size of a meter and reflect collections of bacterial cells held together by an extracellular matrix secreted by these cells. Although most stromatolites were outcompeted in their ecological niches by subsequent fancier forms of multicellular life, a few can still be found today taking essentially the same form as their two-billion year old fossils.

Many interesting kinds of bacterial communities consist of more than one species. Indeed, through a sophisticated system of signaling, detection and organization, bacteria form colonies of all kinds ranging from biofilms to the ecosystems within animal guts. Bacterial biofilms are familiar to us all as the basis of the dentist's warning to floss our teeth every night. These communities are functionally as well as structurally interdependent. Other biofilms are noted for their destructive force when they attach to the surfaces of materials. Fig. 2.26(B) shows a biofilm that grew on a silicon rubber voice prosthesis that had been implanted in a patient for about three months.

Structurally, a biofilm is formed as shown schematically in fig. 2.27. The key building blocks of such structures are a population of bacteria, a surface onto which these cells may adhere and an aqueous environment. The formation of a biofilms results in a population of bacterial cells that are attached to a surface and enclosed in a polymeric matrix built up of molecules produced by these very same bacteria. The early stages of biofilm formation involve the adhesion of the bacteria to a surface followed by changes in the characteristics of these bacteria such as the loss of flagella and the development of pili. At a larger scale, these

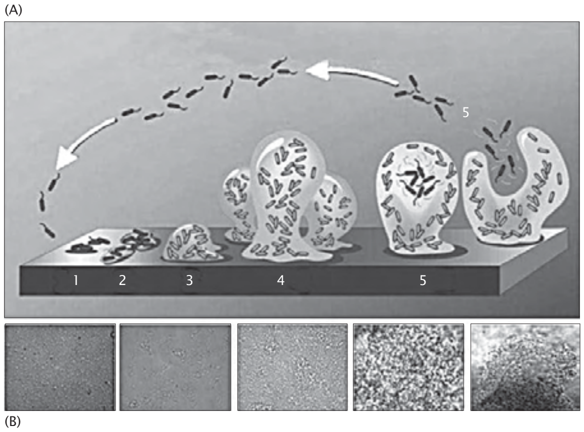


Figure 2.27: Schematic of the formation of a biofilm by bacteria. The various stages in the formation of the biofilm are: 1) attachment to surface, 2) secretion of extracellular polymeric substance (EPS), 3) early development, 4) maturation and 5) shedding of cells from the biofilm. The microscopy images below show biofilms in various stages of the film formation process. (adapted from Stoodley *et al.*, 2002) RP: need scale bars and caption for lower figures.

changes at the cellular level are attended by the formation of colonies of cells and differentiation of the colonies into structures which are embedded in extracellular polysaccharides. Though there are a variety of different morphologies that are adopted by such films, roughly speaking, these biofilms are relatively porous structures (presumably to provide a conduit for import and export of nutrients and waste, respectively) that typically take on mushroom-like structures such as indicated schematically in fig. 2.27. These films have a relative proportion of something like 85 percent of the mass taken up by extracellular matrix while the remaining 15 percent is taken up by cells themselves. A typical thickness for such films ranges from 10-50 μm.

Teaming Up in a Crisis: Lifestyle of *Dictyostelium discoideum*

Although bacteria can form communities, eukaryotes have clearly raised this to a high art. One particularly fascinating example that may give clues as to the origin of eukaryotic multicellularity is the cellular slime mold *Dictyostelium discoideum* as shown in fig. 2.26(C). This small, soil-dwelling amoeba pursues a solitary life when times are good but seeks the comfort of its fellows during times of starvation. *Dictyostelium* is usually content to wander around as an individual with a characteristic size between 5 and 10 μm. However, when deprived of its bacterial diet, these cells undertake a radical change in lifestyle which involves both interaction and differentiation through a series of fascinating intermediate steps as shown in fig. 2.28. A group of *Dictyostelium* amoebae in a soil sample that find themselves faced with starvation, send chemical signals

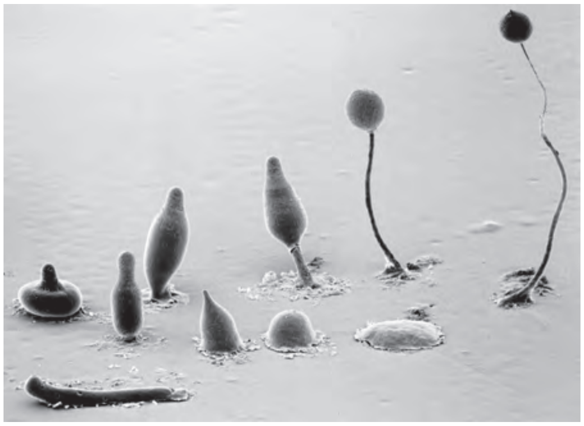


Figure 2.28: Formation of a multicellular structure during starvation. The social amoeba *Dictyostelium discoideum* responds to starvation by forming a structure made up of tens of thousands of cells in which individual cells suffer different fates. Cells near the top of the structure form spores which are resurrected once conditions are favorable. The figure shows the developmental stages that take place on the way to making fruiting bodies, starting in the bottom right and proceeding clockwise. RP: the citation is "Copyright, M.J. Grimson and R.L. Blanton, Biological Sciences Electron Microscopy Laboratory, Texas Tech University." RP: scale bar

to one another resulting in the coalescence of thousands of separate amoebae to form a slug that looks like a small nematode worm. These cells appear to be poised on the brink between unicellular and multicellular lifestyles and can readily convert between them. Ultimately, as shown in the figure, the slug stops moving and begins to form a stalk. At the tip of the stalk is a nearly spherical bulb that contains many thousands of spores, essentially cells in a state of suspended animation. When environmental conditions are appropriate for individual amoeba to thrive, the spores undergo the process of sporulation, with each spore becoming a functional, individual amoeba.

- **Estimate: Sizing Up the Slug and the Fruiting Body.** The relation between the number of cells in a slug and its size is shown in fig. 2.29 where we see that these slugs can range in size between several hundred and several thousand microns with the number of cells making up the slug between tens of thousands and several million. The next visible stage in the development is the sprouting of a stalk with a bulb at the top known as a fruiting body. The stalk is of order a millimeter in length while the size of the fruiting body itself is several hundred microns across. This fruiting body is composed of thousands of spores, amoeba that are effectively in a state of suspended animation. An example of such a fruiting body that has been squished on a microscope slide is shown in fig. 2.30. This structure has functional consequences. In particular, those cells that are part of the spore remain in a sort of suspended animation, remaining poised to respond to a better day, while the cells that formed the stalk have effectively ended their lives for the good of those that survive.

An immediate question of interest concerning the multicellular fruiting bodies shown in fig. 2.28 is how many cells conspire to make up such structures. Fig. 2.30 provides the answer, but it is also of interest to try to reason it out. An estimate of the number of cells in a fruiting body can be constructed by examining the nearly hemispherical colony of cells shown in fig. 2.30. The diameter of this hemisphere is roughly $= 200\mu$m. Our rough estimate for the number of spores in the fruiting body is obtained by evaluating the ratio

$$\text{number of cells} = \frac{V_{body}}{V_{cell}}, \tag{2.20}$$

where we assume that the entirety of the fruiting body volume is made up of cells. If we assume that the cell size is 4 μm in diameter, this yields

$$\text{number of cells} = \frac{\frac{2}{3}\pi(100\mu m)^3}{\frac{4}{3}\pi(2\mu m)^3} \approx 2 \times 10^4 \text{cells}. \tag{2.21}$$

Note that the size of the ball of cells in a fruiting body can vary dramatically from the 200 μm scale shown here to several times larger, and as a result, the estimate for the number of cells in a fruiting body can vary. Also, note that a factor of two error in our estimate of the size of an individual cell will translate into a factor of eight error in our count of the number of cells in the slug or fully formed fruiting body.

Multicellular Organisms Have Many Distinct Communities of Cells

The three branches of life that have most notably exploited the potential of the multicellular lifestyle are animals, plants and fungi. While bacterial and protozoan colonial organisms rarely form communities with characteristic dimensions of more than a few millimeters (with the exception of stromatolites), individuals in these three groups routinely grow to more than a meter in size. Their enormous size and corresponding complexity can be attributed to at least three factors: i) production of extracellular matrix material that can provide structural support for large communities of cells, ii) a predilection towards cellular specialization such that many copies of cells with the same genomic content can develop to perform distinct functions and iii) highly sophisticated mechanisms for the cells to communicate with one another within the organism. We emphasize that these traits are not unique to animals plants and fungi, but they are used more extensively there than elsewhere. A beautiful illustration of these principles is seen in the eye of the fruit fly *Drosophila* as shown in fig. 2.26(D). The eye is made up of hundreds of small units called ommatidia,

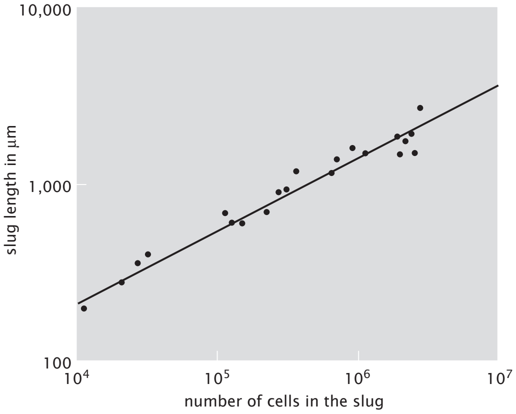


Figure 2.29: Slug size in *Dictyostelium discoideum.* The plot shows a relation between the size of the Dicty slug and the corresponding number of cells. (adapted from Bonner, 20RP)

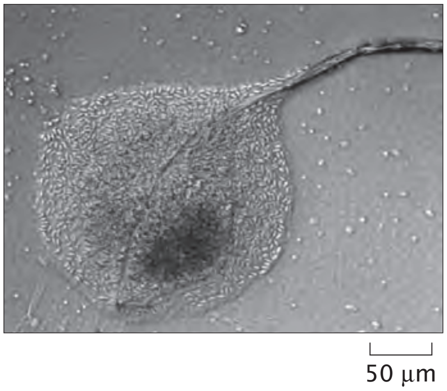


Figure 2.30: Microscopy image of a fruiting body. The fruiting body has been squished on a microscope slide revealing both the size of the spores and their numbers.

each of which contains a group of eight photoreceptor cells, support cells and a cornea. During development of the eye, these cells signal to one another to establish their identities and relative positions to create a stereotyped structure that is repeated many times. The overarching theme of the remainder of this chapter is the exploration of how cells come together to form higher order structures and how these structures fit into the overall hierarchy of structures formed by living organisms.

2.3.2 Cellular Structures From Tissues to Nerve Networks

Multicellular structures are as diverse as cells themselves. Often, the nature of these structures are a reflection of their underlying function. For example, the role of epithelia as barriers dictates their tightly packed, planar geometries. By way of contrast, the informational role of the network of neurons dictates an entirely different type of multicellular structure.

One Class of Multicellular Structures Is the Epithelial Sheets

Epithelial sheets form part of the structural backdrop in organs ranging from the skin to the bladder. Functionally, the cells in these structures have roles such as serving as a barrier to transport of molecules, providing an interface at which molecules can be absorbed into cells and as the seat of certain molecular secretions. Several different views of these structures are shown in fig. 2.31.

The morphology of epithelial sheets are diverse in several ways. First, the morphology of the individual cells can be different (isotropic vs anisotropic). In addition, the assemblies of cells themselves have different shapes. For example, the different structures can be broadly classified into those which are a monolayer sheet (simple epithelium) and those which are a multilayer (stratified epithelium). Within these two broad classes of structures, the cells themselves have different morphologies. The cells making up a given epithelial sheet can be flat, pancake-like cells, denoted as *squamous* epithelium. If the cells making up the epithelial sheet have no preferred orientation, they are referred to as *cuboidal*, while those which are elongated perpendicular to the extracellular support matrix are known as *columnar* epithelia. Epithelial sheets have as one of their functions (as do lipid bilayers) the segregation of different media which can have highly different ionic concentrations, pH, macromolecular concentrations and so on.

Tissues Are Collections of Cells and Extracellular Matrix

We have seen that cells can interact to form complexes. An even more intriguing example of a multicellular structure is provided by tissues such as that shown in fig. 2.32. These connective tissues are built up from a diverse array of cells and materials they secrete. Beneath the epithelial surface, fibroblasts construct extracellular matrix. This connective tissue is built up of three main components: cells such as fibroblasts and macrophage, connective fibers and a structureless supporting substance made up of glycosaminoglycans (GAGs). What is especially appealing and intriguing about these tissues is the orches-

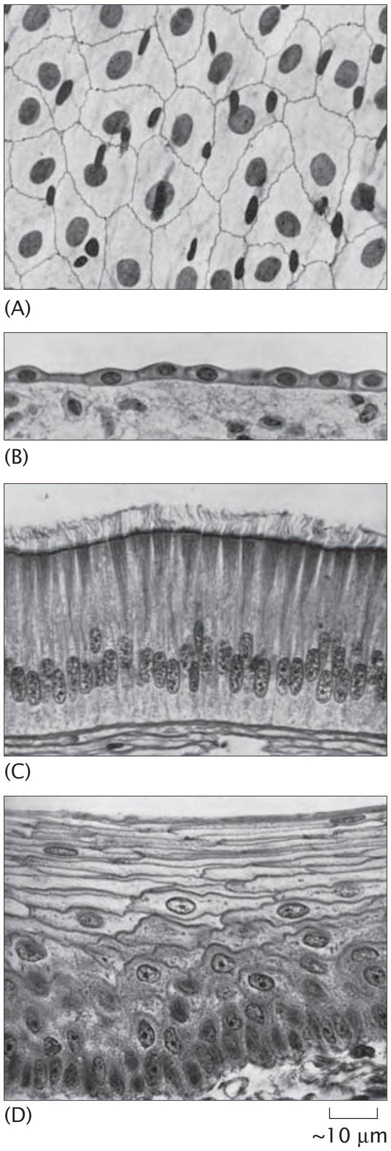


Figure 2.31: Shapes and architecture of epithelial cells. Epithelia are tissues formed by continuous sheets of cells that form tight contacts with one another. (A) Viewed from above, a simple epithelial sheet resembles a tiled mosaic. The dark ovals are the cell nuclei stained with silver. (B) Viewed from the side, simple epithelia such as this from the dog kidney, may form a single layer of flattened cells above a loose, fibrous connective tissue. (C) In some specialized epithelia, the cells may extend upwards forming elongated columns and develop functional specializations at the top surface such as the beating cilia shown here. This particular ciliated columnar epithelium is from the alimentary tract of a freshwater mussel. (D) In other epithelia such as skin or the kitten's gum shown here, epithelial cells may form multiple layers.

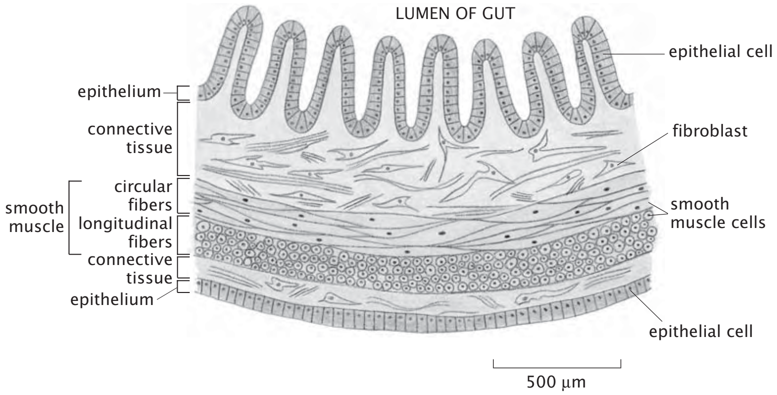


Figure 2.32: Connective tissue. The schematic shows the organization of cells and extracellular matrix that make up connective tissue. The top layer is a planar array of epithelial cells. The volume beneath these cells is made up of fibroblasts and a secretion of extracellular matrix.

tration, both in space and time, leading to the positioning of cells and fibers. The fibroblasts, which were already featured earlier in this chapter, serve as factories for the proteins that make up the extracellular matrix. In particular, they synthesize proteins such as collagen and elastin which, when secreted, assemble to form fibrous structures which can support mechanical loads. The medium within which these fibers (and the cells) are embedded is made up of the third key component of the extracellular matrix, namely, the hydrated gel of glycosaminoglycans.

Nerve Cells Form Complex, Multicellular Complexes

A totally different example of structural organization involving multiple cells is illustrated by collections of neurons. Neurons are the specialized cells in animals that we associate with thinking and feeling. These cells allow for the transmission of information over long distances in the form of electrical signals. Neurons are constructed with many input terminals known as dendrites and a single output path known as the axon. The fascinating structural feature of these cells is that they assemble into complex networks that are densely connected in patterns where dendrites from a given cell reach out to many others, from which they take various inputs. An example of a collection of fluorescently labeled neurons is shown in fig. 2.33. Note that the branches (dendrites and axons) that reach out from the various cells have lengths far in excess of the 10 micron scale characteristic of typical eukaryotes. Indeed, axons of some neurons can have lengths of centimeters and more.

One fascinating example of neuronal contact is offered by the so-called neuromuscular junction as shown in fig. 2.34. These junctions are the point of contact between motor neurons, which convey the marching orders for a given

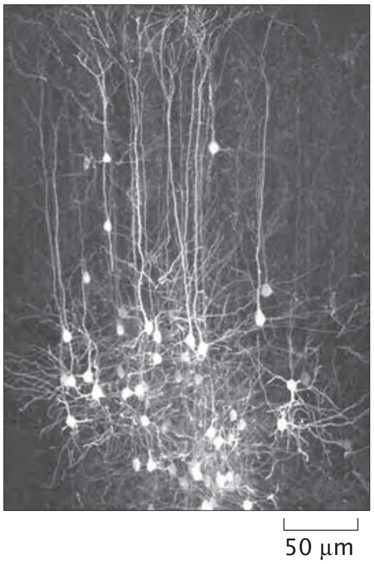


Figure 2.33: Illustration of the complex network of cells formed by neurons. GFP fluorescence observed for a collection of neurons from the brain of a rat. Particular neurons were targeted using lentiviruses. (figure from RP)

muscle, and the muscle fiber itself. As is seen in the figure, the axon from a given motor neuron makes contacts with various muscle fibers. As will be described in more detail in chap. 17, when an electrical signal (action potential) arrives at the contact point known as a synapse, chemical neurotransmitters are released into the space between the nerve and muscle. These neurotransmitters result in the opening of ligand-gated ion channels in the muscle which result in a change in the electrical state of the muscle and lead to motion of the muscle. Contrasting the contacts between epithelial sheets and neurons (or neurons and muscles) reveals the diversity of cell-cell contacts.

2.3.3 Multicellular Organisms

The highest level in the structural hierarchy to be entertained here is individual organisms. The diversity of living multicellular organisms is legendary ranging from roses to hummingbirds, Venus flytraps to the giant squid. What is especially remarkable about this diversity is contained in the simple statement of the cell theory: each and every one of these organisms is a collection of cells and their products. And, each and every one of these organisms is the the result of a long history of evolution resulting in specialization and diversification of the various cells that make up that organism.

Cells Differentiate During Development Leading to Entire Organisms

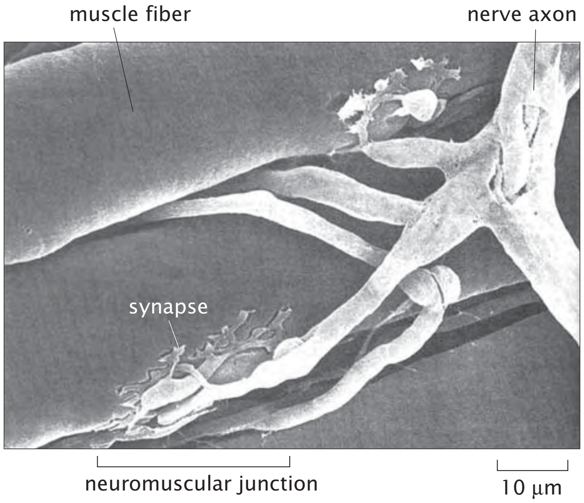


Figure 2.34: Neuromuscular junction. The axon from a nerve cell makes contact with various skeletal muscle fibers. Neurotransmitters secreted by the nerve cell at the synapse initiate contraction of the muscle fibers. (adapted from Bloom and Fawcett, 20RP).

The fruit fly *Drosophila melanogaster* has had a long and rich history as one of the "model" organisms of biology and is a useful starting point for thinking about the size of organisms. As shown in fig. 2.35, the mature fly has a size of roughly 3 mm that can be thought of morphologically as being built up of 14 segments: 3 segments making up the head, three segments making up the thorax and 8 segments making up the abdomen.

Drosophila has attained its legendary status in part because of the way it has revealed so many different concepts about embryonic development. One of the most well-studied features of the development of *Drosophila* is the way in which it lays down its anterior-posterior architecture during early development. The pattern of expression of the so-called even-skipped genes is shown in fig. 2.36. The gene even-skipped (eve) is expressed in seven stripes corresponding to seven of the fourteen *Drosophila* segments. Another intriguing feature of this figure is the sense in which different species of fly maintain the same overall relative position of different morphological features, despite their up to tenfold difference in overall size.

- **Estimate: Sizing Up Stripes in** ***Drosophila*** **Embryos.** To get a feeling for the scales associated with the gradients in transcription factors that dictate developmental decisions and the features they engender, we idealize the *Drosophila* embryo as a spherocylinder. As shown in fig. 2.36, the geometry is characterized by two parameters, the length of the cylindrical region, L, and its radius R, where we use approximate values of 300 μm

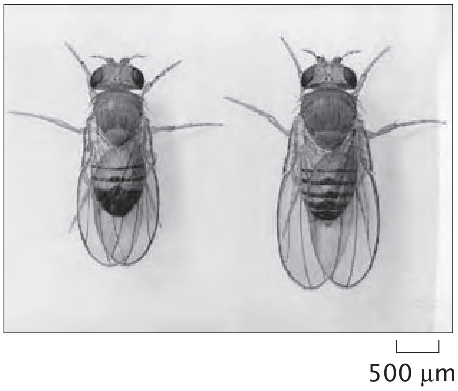


Figure 2.35: Mature *Drosophila* flies (a) male, (b) female. RP: gotten from the flybase. We need to take our own pictures. http://flybase.bio.indiana.edu/.bin/fbidq.html?FBrf0004865

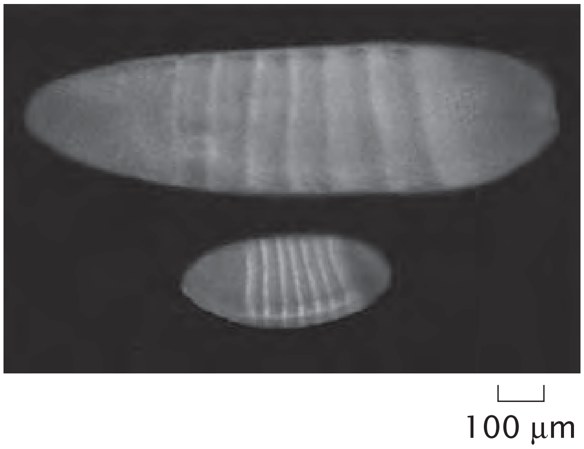


Figure 2.36: Pattern of gene expression in the *Drosophila* embryo. Image of *Drosophila* embryos from different species after RP hours of development, (Scale bars: 100 microns.) RP: here is caption from Gregor paper: Immunofluorescence stainings for products of the gap and pair-rule genes in higher diptera. (A) Immunofluorescence staining of L. sericata (upper embryos) and D. melanogaster (lower embryos) for Hunchback (green) and Giant (red) in the left column, and for Paired (green) and Runt (red) in the right column. (B) Anti-Hunchback (green) and anti-Runt (red) immunofluorescence staining of D. melanogaster (upper embryo) and D. busckii (lower embryo). (Scale bars: 100 microns.)

for the length of the cylindrical region and 100 μm for the radius. Since the early embryo is a syncytium with all of the cells forming a surface layer, our estimates will depend upon having a rough estimate of the areal density of the cells. The area of the embryo in this simple model is given by

$$A = 4\pi R^2 + 2\pi RL. \tag{2.22}$$

If we consider the embryo after 13 generations of cell division and before the gastrulation which folds the developing embryo in, there are roughly 8000 cells. Hence, the resulting density is given by

$$\sigma = \frac{N}{4\pi R^2 + 2\pi RL}. \tag{2.23}$$

Using the numbers described above, this leads to an areal density of 0.025 cells /μm^2. As seen in fig. 2.36(c), the stripe patterns associated with the *Drosophila* embryo are very sharp and reflect cells making decisions at a very localized level.

Inspection of fig. 2.36(c) reveals that the stripes are roughly 30 μm wide. As a result, we can estimate the total number of cells participating in these stripes as

$$n = \sigma 2\pi R l_{stripe} \approx 0.026\text{cells}/\mu m^2 \times 2 \times 3.14 \times 100\mu m \times 30\mu m \approx 500. \tag{2.24}$$

Note that the average area per cell is given by $1/n \approx 36\mu\text{m}^2$, suggesting that the radii of these cells is roughly 3.5 μm. Our main purpose in carrying out this exercise is to demonstrate the length scale of the structures that are put down during embryonic development. In this case, what we have seen is that out of the roughly 8000 cells that characterize the *Drosophila* embryo at the time of gastrulation, groups of roughly 500 cells have begun to follow distinct pathways as a result of differential patterns of gene expression.

The Cells of the Nematode Worm *C. elegans* Have Been Charted

A more recently popularized model organism for studying the genetics and development of multicellular animals is the nematode worm *Caenorhabditis elegans*. Two factors that make this worm particularly attractive in its capacity as a model multicellular eukaryote are that a) its complete genome has been determined and b) the identity of each and every of its 959 cells has been determined (see fig. 2.38). Amazingly, *all* of the cells of this organism have had their lineages traced from the single ancestral cell which is present at the moment of fertilization. What this means precisely is that all of the roughly 1000 cells making up this organism can be assigned a lineage of the kind cell A begat cell B which begat cell C and so on. As shown in fig. 2.38, these worms are roughly 1 mm in length and 0.05 mm across. Like *Drosophila*, they too have been subjected to a vast array of different analyses including, for example, how

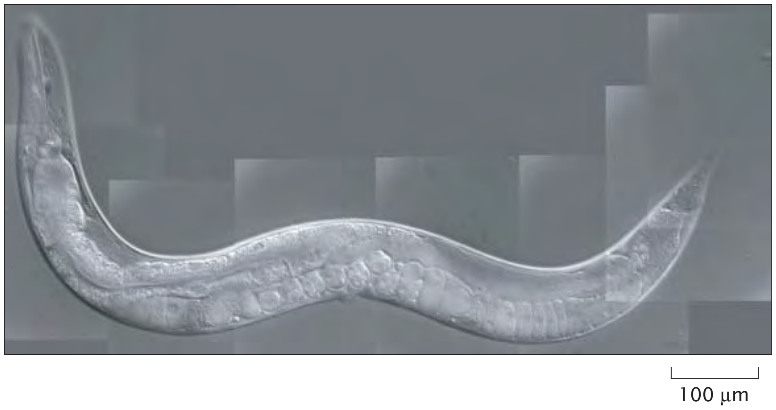


Figure 2.37: *C. elegans*. This DIC image of a single adult worm was assembled from a series of high-resolution micrographs. The worm's head is at the top left corner and its tail is at the right. Its gut is visible as a long tube going down the animal's body axis. Its egg cells are also visible as giant ovals towards the bottom of the body.

their behavior is driven by the sensation of touch. One of the most remarkable outcomes of the series of experiments leading to the lineage tree for *C. elegans* was the determination of the connectivity of the 302 neurons present in this nematode. By using serial thin sections from electron microscopy, it was possible to map out the roughly 7000 neuronal connections in the nervous system of this tiny organism. The various nerve cells are typically less than 5 microns across.

- **Estimate: Sizing Up** *C. Elegans***.** As a simple estimate of the cellular content of a "simple" organism, we contemplate one of the key model organisms of modern biology, *Caenorhabditis elegans*. For our present purposes, we think of these small worms as cylinders of length 1 mm and with a width of .05 mm. The total volume of such a worm is computed simply as roughly $5 \times 10^6 \mu m^3$. A reasonable estimate for the volume of eukaryotic cells is somewhere between 2000 - 10000 μm^3. If we consider a characteristic volume of 5000 μm^3 per cell, this results in an order of magnitude estimate that there are 1000 cells per worm.

Many different species of nematodes share a common body plan involving a small number of well defined cells, however they may vary greatly in size. One of the largest nematodes *Ascaris*, is a common parasite of pigs and humans. It has been estimated that up to one billion people on the planet carry *Ascaris* in their intestine. These worms closely resemble *C. elegans* except that they may be up to 15 cm in length. This kind of observation is not unusual throughout the animal kingdom. Many species have close relatives that differ enormously in size. The reasons and mechanisms that determine the overall size of multicellular organisms remain poorly understood.

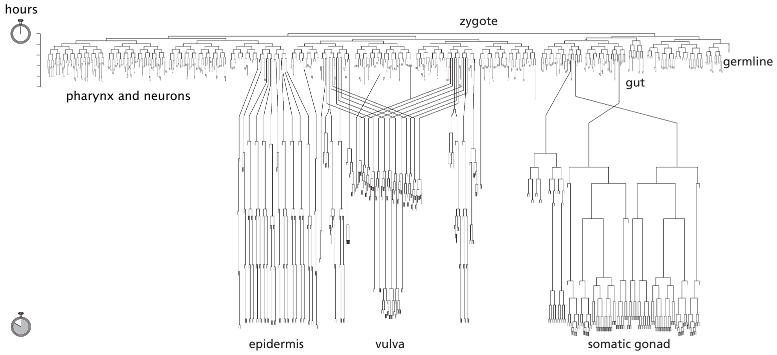


Figure 2.38: Cell lineages in *C. elegans.* The developmental pattern of every cell in the worm is identical from one animal to the other. Because of this feature it has been possible for developmental biologists to determine the family tree of every cell present in the entire animal by patient direct observation. In this schematic representation, the cell divisions that occur during embryogenesis are shown in the band across the top. The later cell divisions of the epidermis, vulva and somatic gonad all take place after the animal has hatched. The Xs represent cells that die; death is a normal developmental fate.

Higher Level Structures Exist as Colonies of Organisms

Organisms do not exist in isolation. Every organism on the planet is part of a larger ecological web that features both cooperation and competition with other individuals and species. Corral reefs represent a vivid example of the interdependence of huge varieties of species living together in close quarters. An equally vivid and diverse community, though one frequently less appreciated is found closer to home within our own intestines, which are inhabited by a teeming variety of bacteria. It has been estimated that the human body actually harbors nearly ten times more microbial cells than it does human cells, and at least one hundred times more bacterial genes than human genes. Thus we should think of ourselves not really as individuals but rather as complex ecosystems of which the human cells form only a small part. Thus, our story of the hierarchy of structures that make up the living world began with the bacterium *Escherichia coli* and will end there now. It is a great irony that we as humans, it might be argued, have been colonized by our standard ruler *E. coli*.

Though we have examined biological structures over a wide range of spatial scales, our powers of ten journey still falls far short of being comprehensive. Ultimately, what we learn from this is that the handiwork of evolution has resulted in biological structures ranging from the nanometer scale of molecular machines all the way to the scale of the planet itself. Trying to understand what physical and biological factors drive the formation of these structures will animate much of the remainder of the book.

2.4 Summary and Conclusions

Biological structures range from the scale of nanometers (individual proteins) to hundreds of meters (redwood trees) and beyond (ecological communities). In this chapter, we have explored the sizes and numbers of biological things starting with the unit of life, the individual cell and working our way down and up. We have found that sometimes biological things show up in very large numbers of identical copies such that averages, for example, concentration of ribosomes in the *E. coli* cytoplasm, are reasonable approximations. On the other hand, sometimes biological things show up in very few copies such that the exact number can make a big difference in the behavior of the system. We have found that cells are crowded, that there really is a world in a grain of sand (or in a biological cell). We have also explored some of the ways that biological units such as proteins or cells may interact with one another to form larger and more complex entities. Having developed an intuitive feeling for size and scale will better enable us to realistically envision the biological processes and problems described throughout the remainder of the book.

2.5 Further Reading

Boeke K., **Cosmic View The Universe in 40 Jumps**, The John Day Company, New York: New York, 1957. Boeke starts with a picture of a child holding a cat and proceeds to view her situation both by decreasing (26 orders of magnitude) and increasing (13 orders of magnitude) the resolution of the view.

Goodsell D., **The Machinery of Life**, Springer-Verlag, New York: New York, 1998. Following on the work of Minton and Zimmermann (RP: check), Goodsell had the very clever idea of trying to represent the cell as it really is, crowded and to reveal the connection between structure and function, explicitly and visually. His books make for fascinating reading.

Kornberg A., **For the Love of Enzymes**, Harvard University Press, Cambridge: Massachusetts, 1991. Kornberg's beautiful book, though featured here because of its constant appeal to cartoons, is hugely fascinating. It is a great pleasure to see his passion for science, with this book in particular reading like one long ode to enzymes. Part of the relevance to the current chapter is his Figure 2-2 where he gives a biological view of the structures that are seen as one telescopes through powers of ten. In addition, Kornberg's book is amongst the most thoughtful we have seen with respect to intelligent figures which center on illustrating particular modeling ideas.

Stryer RP
Fawcett D. W., **The Cell, Its Organelles and Inclusions**, W. B. Saunders and Company, Philadelphia, Pennsylvania, 1966. We imagine our reader comfortably seated with a copy of Fawcett right at his or her side. Fawcett's electron microscopy images are stunning.

Fawcett D. W. and Jensh R. P., **Bloom and Fawcett's Concise Histology**, Arnold Publishers, London, England, 1997. This book is eye candy for for those who wish to see some of the wonderful and beautiful diversity of cells both and their organelles.

Nelson D. L. and Cox M. M., **Lehninger Principles of Biochemistry**, Worth Publishers, New York: New York, 2000. Just as Alberts *et al.* serves as a representative example of the representation of molecular biology via cartoons, the present book gives a similar impression of the way in which biochemical ideas are represented.

Gilbert S. F., **Developmental Biology**, Sinauer Associates, Sunderland: Massachusetts, 2003. Gilbert's book is a beautiful source for learning about the architecture of a host of different organisms during early development. Chap. 9 is especially relevant for the discussion of this chapter.

Wolpert L., Beddington R., Jessell T., Lawrence P., Meyerowitz E., and Smith J., **Principles of Development**, Oxford University Press, New York: New York, 2002. Wolpert's book is full of useful cartoons and schematics that illustrate many of the key ideas of developmental biology. We admire the use of scale bars in some of the photographs and wish the practice were universal.

Levine A. J., **Viruses**, Scientific American Library, New York: New York,

1992. Levine's book is full of interesting cartoons in which the various stages of viral infection are represented by provocative cartoons.

Harold F., **The Way of the Cell**, Oxford University Press, New York: New York, 2001. Harold's book is an ode to emergence. He celebrates with glee the insights that have been garnered on the basis of reductionist thinking in molecular biology. But he point also points out that the emergent properties exhibited by cellular phenomena defy description in terms of such purely reductionist notions. Can we quantify and perfect this line of argument. We need models of emergence. RP: Harold has great quote we should use: "A physics that has no place for life is as impoverished as would be a biology not informed by chemistry."

McMahon T. A. and Bonner J. T., **On Size and Life**, Scientific American Library, New York: New York, 1983.

2.6 Problems

1. **A feeling for the numbers: microbes as the unseen majority**

(a) Justify the assumption that a typical (i.e. *E. coli*) bacterial cell has a volume of 1 μm^3. Also, express this volume in femtoLiters. The claim is made (see Whitman *et al.*, PNAS, 95(12):6578 (1998)) that in the top 200 m of the world's oceans, there are roughly 10^{28} prokaryotes. Work out the total volume taken up by these cells in m^3 and km^3. Compute their mean spacing.

(b) Recall that roughly 2-3 kg of bacteria are to be found in the waste factory of your large intestine. Make an estimate of the total number of bacteria inhabiting your intestine and then all of the intestines of all of the humans currently on the Earth.

(c) Look at the fascinating paper by Zimmerman and Trach (JMB 222(3):599 (1991)) in which they attempt to measure the crowding in the cellular interior. In table 3 they tell us their estimated macromolecular concentrations in the cellular interior. Use these numbers to make an estimate of the mean protein spacing.

2. **A Feeling for the Numbers: Molecular volumes, masses, numbers and charges**

(a) Estimate the volumes of the side chains of the various amino acids, again in $\mathring{A}^3$ units. (RP: see the sheet from Michael Levitt in which he has these estimates.)

(b) Generate an estimate for the size of a "typical amino acid" in daltons. Justify your estimate by explaining how many of each type of atom you chose. Compare your estimate to several key amino acids such as glycine, proline, arginine and tryptophan.

(c) On the basis of your result for part (b), deduce a rule of thumb for converting mass of a protein (reported in kD) into a corresponding number of residues. Apply this rule of thumb to myosin, G-actin, hemoglobin and hexokinase and compare it to the actual number of residues in each of these proteins.

3. **Atomic-Level Representations**

of Biological Molecules

(a) Obtain coordinates for ATP, phosphatidylcholine, B-DNA, G-actin, the lambda repressor/DNA complex or Lac repressor/DNA complex, myoglobin, green fluorescent protein (GFP) and RNA polymerase. You can do this by searching in the Protein Data Bank and various other Internet resources.

(b) Download a structural viewing code such as VMD (University of Illinois), Rasmol (University of Massachusetts) or DeepView (Swiss Institute of Bioinformatics) and create a plot of each of the molecules you downloaded above. Experiment with the orientation of the molecule.

(c) By looking at phosphatidylcholine justify the value of the area per lipid used in the chapter.

(d) Phosphoglycerate kinase is a key enzyme in the glycolysis pathway. One intriguing feature of such enzymes is their enormity in comparison with the sizes of the molecules upon which they act (their "substrate"). Obtain the coordinates for both phosphoglycerate kinase and glucose (for example, at Molecules R Us) and examine the relative size of these molecules. First, use your graphics programs to plot both molecules simultaneously. Next, treat each of these molecules as a sphere and characterize them both in terms of their linear dimensions and also in terms of their relative volumes.

4. Packing of DNA in viruses Visit the VIPER website and examine the viral capsids of the double stranded DNA viruses herpes and T7. Estimate the lengths of the genomes of these viruses by assuming that the packing of the DNA is such that the effective diameter of a DNA molecule is 25 $\mathring{A}$ and that the space is completely filled. That is, each base pair occupies a volume of a stub cylinder with length 3.4 $\mathring{A}$ and radius 12.5$\mathring{A}$. How accurate are your estimates of genome lengths? Compare the packing fraction for different bacterial and eukaryotic viruses. Deduce an expression relating the radius of the capsid and the length of the genome in number of base pairs, N_{bp}.

5. HIV estimates

a) Estimate the total mass of an HIV virion by comparing it to *E. coli.*

b) The maturation process involves proteolytic clipping of the Gag polyprotein so that the capsid protein CA can form the shell surrounding the RNA genome and nucleocapsid NC can complex with the RNA itself. Estimate the number of CA proteins that are used up to make up the capsid and to see how this number compares with the total number of Gag proteins.

6. Areas and volumes of organelles and cells

(a) Consider reconstructions of the Golgi complex such as those shown in fig. 4 of McIntosh, 2001 and estimate the area of these membrane-bound organelles as well as the number of lipid molecules. Compare this area to that of the entire plasma membrane.

(b) Estimate the membrane surface area of a fibroblast and the area for a reticular model of the ER.

(c) By looking at figure 2.19(a) estimate the total number of ribosomes associated with the ER.

(c) To take stock of the density of ribosomes in eukaryotic cells we can examine their mean spacing in the rough endoplasmic reticulum and use this to estimate the total number of ribosomes associated with the rough ER. Examination of electron micrographs like that shown in fig. 2.19(a) suggests a mean spacing between ribosomes of roughly twice their diameter, or 40 nm. This implies an areal density of

$$\sigma \approx \frac{1\text{ribosome}}{1600nm^2} \approx 6{\times}10^{-4} ribosomes/nm^2. \tag{2.25}$$

Given our earlier estimate of the area of the rough ER, namely, $A_{ER} \approx 1.5 \times 10^{11} nm^2$, this implies the number of ribosomes tied to the ER in a eukaryote is of order 90×10^6.

7. Minimal media and *E. coli* Look up the formula for "minimal media" for growing bacteria. What are the carbon sources in this media? Make an estimate of the number of carbon atoms it takes to make up the macromolecular contents of a bacterium such as *E. coli.*

Chapter 3

When: Stopwatches at Many Scales

"Dost thou love life? Then do not squander time, for that is the stuff life is made of." - Benjamin Franklin

Chapter Overview: In Which Various Stopwatches Are Used to Measure the Rate of Biological Processes

Just as biological structures exist over a wide range of spatial scales, biological processes take place over time scales ranging from much faster than microseconds to the time scales that characterize the history of Earth itself. Using the cell cycle of *E. coli* as a standard stopwatch, this chapter develops a feel for the rates at which different biological processes occur. With this "feeling for the numbers" in hand, we then explore several different views of the passage of biological time.

3.1 The Hierarchy of Temporal Scales

One of the defining features of living systems is that they are dynamic. The time scales associated with biological processes run from the nanosecond (and faster) scale of enzyme action to the more than 10^9 years that cover the evolutionary history of life itself. The inexorable march of biological time is revealed over many orders of magnitude difference in time scale, as illustrated in fig. 3.1. If we are to watch biological systems unfold with different stopwatches in hand, the resulting phenomena will be different - at very fast time scales we will see the molecular dance of different biochemical species as they interact and change identity. At much slower scales we see the unfolding of the lives of individual cells. If we slow down our stopwatch even more, what we see is the trajectories of entire species. To some extent, there is a coupling between the temporal scales described in this chapter and the spatial scales described in the previous chapter; small things such as individual molecules tend to operate at fast rates, and large things such as elephants tend to operate at slow rates.

The aim of this chapter is to describe the time scales of biological phenomena from a number of different perspectives. In section 3.1, we develop a feeling for biological time scales by examining the range of different time scales seen in cell biology and evolutionary biology. This discussion is extended by describing the experimental basis for what we know about time scales in biology. As in chap. 2, we once again invoke *E. coli*, this time by using the cell cycle of our "reference cell" as the standard stopwatch. The remainder of the chapter is built around viewing time in biology from three distinct perspectives. In section 3.2, we show how the time scale of certain biological processes is dictated by how long it takes some particular procedure (such as replication) to occur. We will refer to this as *procedural time*. Section 3.3 explores time from a different angle. In this case, we consider a broad class of biological processes whose timing is of the "socks before shoes" variety. That is, processes are linked in a sequential string and in order for one process to begin, another must have finished. We will refer to this kind of time keep as *relative time*. Finally, section 3.4 reveals a third way of viewing time in biological processes, as a commodity to be manipulated. In this case, we show how cells and organisms find ways to either speed up or slow down key processes such as replication and metabolism.

3.1.1 Biological Processes: A Rogue's Gallery

Biological Processes Are Characterized By a Huge Diversity of Time Scales

A range of different processes associated with individual organisms, and their associated time scales, is shown in fig. 3.2 (we leave a discussion of evolutionary processes for the next section). Broadly speaking, the aim of this figure is to show a loose powers of ten representation of different biological processes. As we will see later in the chapter, an *absolute* measure of time in seconds or minutes is sometimes not the most useful way to think about the passage of time within cells. For example, embryonic development for humans takes drastically longer than for chickens, but the relative timing of common events is meaningfully compared. For the moment, our discussion of fig. 3.2 is intended to give a feeling for the numbers how long do various key biological processes actually take in absolute terms as measured in seconds, minutes and hours?

We begin (fig. 3.2(A) and (B)) with some of the processes associated with the development of the fruit fly *Drosophila melanogaster*. *Drosophila* has been one of the key workhorses of developmental biology, and much that we know about embryonic development was teased out of watching the processes which take place over the roughly ten days between fertilization of the egg and the emergence of a fully functioning fly. If we increase our temporal resolution by a factor of ten, we see the processes in the development of the fly embryo itself. Over the first ten hours or so after fertilization as shown in fig. 3.2(B), a single cell is turned into more than 5000 cells with particular spatial positions and

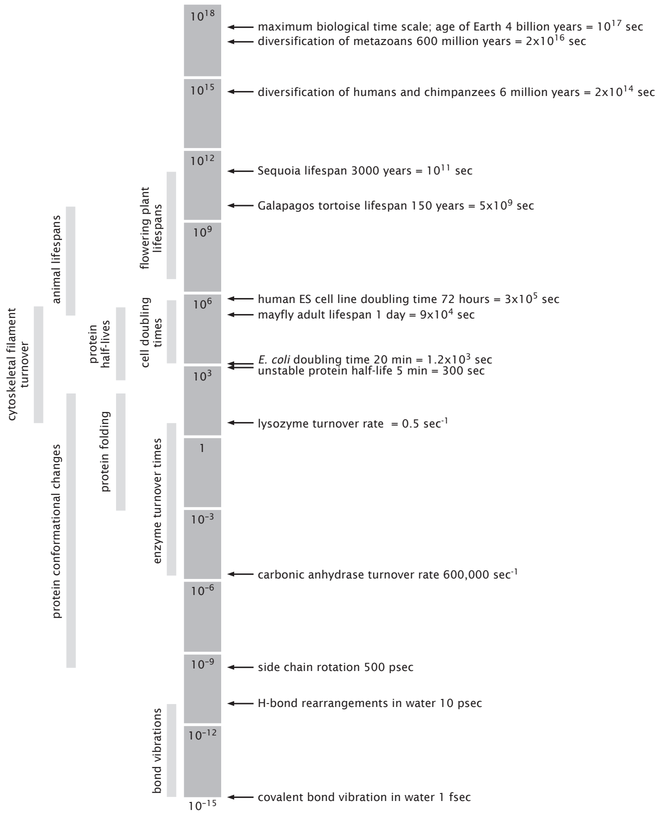

10^18
10^15
10^12
10^9
10^6
10^3
1
10^-3
10^-6
10^-9
10^-12
10^-15
maximum biological time scale; age of Earth 4 billion years = 10^17 sec
diversification of metazoans 600 million years = 2x10^16 sec
diversification of humans and chimpanzees 6 million years = 2x10^14 sec
Sequoia lifespan 3000 years = 10^11 sec
Galapagos tortoise lifespan 150 years = 5x10^9 sec
human ES cell line doubling time 72 hours = 3x10^5 sec
mayfly adult lifespan 1 day = 9x10^4 sec
E. coli doubling time 20 min = 1.2x10^3 sec
unstable protein half-life 5 min = 300 sec
lysozyme turnover rate = 0.5 sec^-1
carbonic anhydrase turnover rate 600,000 sec^-1
side chain rotation 500 psec
H-bond rearrangements in water 10 psec
covalent bond vibration in water 1 fsec
cytoskeletal filament turnover
animal lifespans
protein conformational changes
protein half-lives
protein folding
flowering plant lifespans
cell doubling times
enzyme turnover times
bond vibrations

Figure 3.1: Logarithmic scale showing range of times scales associated with various biological processes.

functions. One of the most dramatic parts of this embryonic development is the process of gastrulation when the future gut forms as a result of a series of folding events in the embryo. This process is indicated schematically in fig. 3.2(B).

Individual cells have a natural developmental cycle as well. The *cell cycle* refers to the set of processes whereby a single cell, through the process of cell division, becomes two daughter cells. The time scales associated with the cell cycle of a bacterium such as *E. coli* are shown in fig. 3.2(C), with a characteristic scale of several thousand seconds. The lives of individual cells are fascinating and complex. If we are to dissect the activities of an individual cell as it goes about its business between cell divisions, we would find a host of processes taking place over a range of different time scales. If we stare down a microscope at a swimming bacterium for several seconds, we will notice episodes of directed motion, punctuated by rapid directional changes. Fig. 3.2(D) shows the time scales over which an individual bacterium such as *E. coli* exercises its random excursion during movement. If our stopwatch now runs a factor of ten faster we are now operating at the scale of deciseconds, a scale which characterizes the rate of amino acid incorporation during protein synthesis, a process represented in fig. 3.2(E). Macromolecular synthesis is one of the most important sets of processes which any cell must undertake to make a new cell. Another key part of the macromolecular synthesis required for cell division is the process of transcription, which is the intermediate step connecting the genetic material as contained in DNA and the readout of that message in the form of proteins. Transcription refers to the synthesis of messenger RNA molecules as faithful copies of the nucleotide sequence in the DNA, a polymerization process catalyzed by the enzyme RNA polymerase. The rate of incorporation by RNA polymerase of nucleotides onto the messenger RNA during transcription, as depicted schematically in fig. 3.2(F), happens roughly ten times as fast as does the rate of amino acid incorporation by ribosomes during protein synthesis.

In the moment to moment life of the cell, proteins do most of the work. Many proteins are able to operate at time scales much faster than the relatively stately machinery carrying out the central dogma operations. For example, a great number of biological processes are dictated by the passage of ions across ion channels, with a characteristic time scale of milliseconds as shown in fig. 3.2(G). A factor of thousand speed up of our stopwatch brings us to the world of enzyme kinetics at the microsecond time scale (fig. 3.2(H)) and faster. It is important to note that these time scales merely represent a general rule of thumb. For example, turnover rates for individual enzymes may range from 0.5 sec^{-1} to 600,000 sec^{-1}.

Before proceeding, one of the questions we wish to consider is how the time scales depicted in fig. 3.2 are actually known. As with much of our story, the stopwatches associated with each of the cartoons in that figure have been determined as the results of many kinds of complementary experiments.

- **Experiments Behind the Facts.** Broadly speaking, the experiments which characterize the dynamics of cells and the molecules that populate them are ultimately based on tracking transformations. We can divide

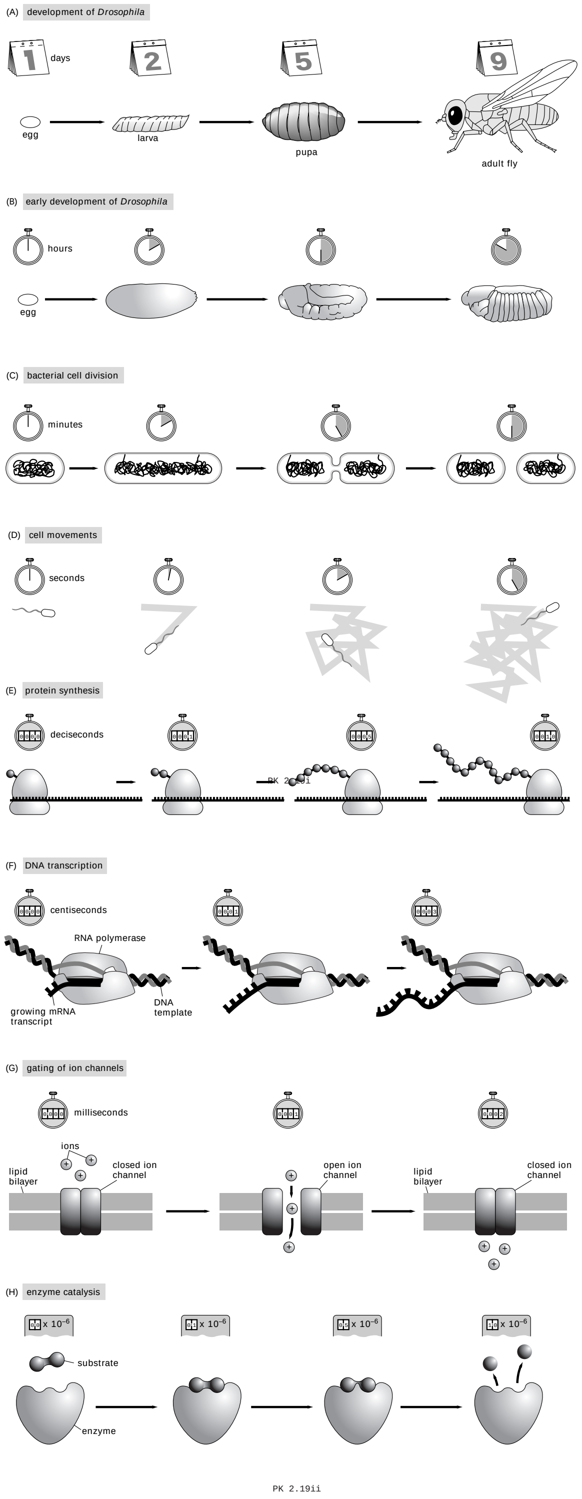


Figure 3.2: Hierarchy of biological time scales. Cartoon showing range of time scales associated with different biological processes.

these experiments into four broad categories that can be applied across all levels of spatial scale from molecular to ecological. These methods are summarized in fig. 3.3.

Direct Observation. The first and most obvious way to characterize time in a biological process is simply to observe the process unfold and to record the absolute time at which transformation occurs. An example of this strategy is shown in fig. 3.3. For example, looking down a microscope at a mammalian cell in tissue culture it is possible to observe many of the steps in the cell cycle unfolding over real time, including condensation of the chromosomes, alignment of the chromosomes through the action of the mitotic spindle, their segregation into daughter nuclei and finally cytokinesis when the cell is pinched into two fully formed daughter cells. Although this is easy to do for processes that take minutes to hours and occur over spatial scales that can be observed with the light microscope or the unaided human eye, it is extremely difficult to measure time simply by observation for events that are very fast, very slow, very small or very large. Over the past few decades there have been vast experimental improvements in direct or near-direct observation of single molecules such that this naturalistic approach to "observing a lot just by watching" can be applied all the way down to the molecular scale. We will see many examples of this approach throughout the book.

Fixed time points. When events of interest cannot be directly observed, there are other ways to probe their duration. Rather than continuously observing an individual over time, one can draw individuals from a population at given time intervals and examine their properties at this series of fixed time points. For example, a bacterial population in a liquid culture started from a single cell will grow exponentially and then plateau and eventually die off over a period of several days. Rather than staring at the tube continuously for several days, the essential kinetics of this process can be measured simply by examining cell density at some fixed interval such as every hour as shown in fig. 3.3. Similarly, the events of embryonic development for useful model organisms such as flies and frogs unfold over a period of days to weeks. However, under a given set of environmental conditions, the sequence and timing of these events is stereotyped from one individual to another. Therefore the investigator can accurately describe the sequence of events in frog development by examining one dish of embryos an hour after fertilization, a second dish of embryos two hours after fertilization, etc. This is useful when the methods used to examine the embryos result in their death. For example, fixing them and staining for a particular protein of interest or preparing them for electron microscopic examination. At a much smaller spatial scale and faster time, the method of stop-flow kinetics enables investigators to follow enzymatic events by mixing together an enzyme and its substrate and then squirting the mixture into a denaturing acid bath after fixed intervals of time. These methods are all more indirect than direct observation, but in many cases are technically easier and different kinds of complementary information can be gleaned by comparing both for a single process.

Pulse-chase. Many biological processes operate in a continuous fashion. For example, bacteria constantly take in sugar from their medium for energy and to generate the molecular building blocks to synthesize new constituents. The process of glycolysis converts a molecule of glucose into two molecules of pyruvate. Because glucose is continuously taken up and pyruvate is continuously generated, it is extremely difficult to ask how long the conversion process takes. The set of methods used to tackle these kinds of problems are generally called pulse-chase experiments. In this particular example, a bacterial cell may be fed glucose tagged with radioactive carbon for a very brief period of time, for example one minute. This is followed by feeding with nonradioactive glucose. Cells can then be removed from the bacterial culture at various time intervals and their metabolites can be examined to see which contain the radioactive carbon. Over time, the amount of labeled glucose will decrease and the major radioactive species will pass through a series of intermediates until finally most of the radioactive carbon will be found in pyruvate. Thus, a pulse-chase experiment can be used to determine the order of intermediates in a metabolic pathway and also the amount of time it takes for the cell to perform each transformation. A classic example of this strategy to examine transport in neurons is shown in fig. 3.3. Essentially the same method is used by naturalists examining dispersion times of birds and other animals by tagging individuals with a band or radio-transmitter and releasing them back into their natural population to see where they are and when.

Product accumulation. The final type of experiment used to determine biological rates is exemplified by an assay with a purified enzyme where a colorless substrate is converted into a colored product over time. By measuring the concentration of the colored product as a function of time, the investigator can extrapolate the average turnover rate given the known concentration of enzymes in the test tube. Similar experiments where observation of the accumulation of a product can be used as a surrogate for rate measurements can also be performed in living cells. A particularly useful example is expressing the green fluorescent protein (GFP) downstream of a promoter of interest as shown in fig. 3.3. When the promoter is induced (i.e. by exposing the cells to some molecule that turns the gene of interest on) GFP begins to accumulate and the amount of fluorescence can be directly measured and converted into numbers of GFP molecules. Because GFP is remarkably stable, its accumulation can often represent a more accurate reporter for promoter activity than the promoter's natural product which may be subject to other layers of regulation including rapid degradation.

METHOD	direct observation	fixed time points	pulse-chase	product accumulation
TYPICAL TIMESCALES	milliseconds to hours	microseconds to years	minutes to days	minutes to days
TYPES OF PROCESSES	individual transformations	population changes	continuous (e.g., metabolism, transport)	biosynthetic or enzymatic
EXAMPLE	cell crawling	bacterial growth curve	axonal transport	GFP expression

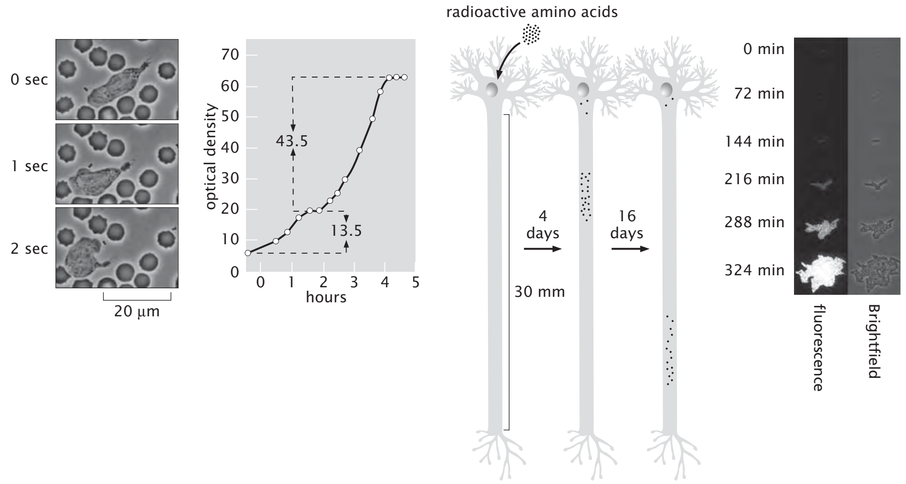


Figure 3.3: Experiments to measure the timing of biological processes. The figure summarizes four strategies for measuring biological rates. For *direct observation* the example shows three frames from a video sequence of a single white blood cell (neutrophil) pursuing a bacterium through a forest of red blood cells. The movement of the cell is sufficiently fast that it can be directly observed by the human eye. For *fixed time points*, the experiment shown is a classic performed by Monod who tracked the growth of *E. coli* in a single culture when two different nutrient sugars were mixed together. The bacteria initially consumed all of the available glucose and then their growth rate slowed as they switched over into a new metabolic mode enabling them to use lactose. For *pulse-chase*, labeling proteins at their point of synthesis in a neuron cell body with a pulse of radioactive amino acids followed by a chase of unlabeled amino acids was used to measure the rate of continuous axonal transport. *Product accumulation* is illustrated by the expression of GFP under a regulated promoter in a bacterial cell. The rate of gene transcription can be inferred by measuring the amount of GFP present as a function of time.

3.1.2 The Evolutionary Stopwatch

The general rule that all biological processes are dynamic and undergo change over time applies to molecules, cells, organisms and species. The evolutionary clock started more than three billion years ago with the appearance of the first cellular life forms on Earth. It is generally accepted that there were complex life-like processes occurring prior to the emergence of the first recognizable cells, though we cannot learn anything about what they were like either from the fossil record or comparative studies among organisms living today.

All of the astonishing diversity of cellular life currently existing on the planet ranging from archaea living in geothermal vents deep in the ocean to giant squid to redwood trees to the yeast that make beer were all descended from a universal common ancestor (probably a population of cells rather than an individual). This last universal common ancestor (LUCA) would have been clearly recognizable as a cell: it contained DNA as a genetic material, it transcribed its DNA into mRNA and translated mRNA into proteins using ribosomes. It also processed sugar to make energy through the process of glycolysis and contained a rudimentary cytoskeleton consisting of an actin-like molecule and a tubulin-like molecule. We can attribute all of these features to LUCA because they are universally shared among all existing branches of cellular life. However, the demonstrable differences between redwood trees and giant squids accumulated slowly over evolutionary time as individual cellular populations became genetically isolated from one another and underwent change and divergence to fill different ecological niches. As the planet Earth is constantly being reshaped and remodeled by the uncounted legion of organisms that inhabit it, environmental niches are always unstable and can be changed either by geological processes, global climate alterations or the actions of competing organisms.

We can fruitfully think of evolution as the process of change in the genetic information carried by a population of related organisms. Sometimes a single lineage can be seen as altering over time as its environment changes. More commonly, a single population will subdivide into populations that will become isolated and suffer different fates. Some will die off, some will remain similar to the ancestral population and some will undergo significant biochemical, morphological or behavioral alterations over time that are ultimately recognized as new species. These basic ideas were beautifully articulated by Charles Darwin in *The Origin of Species* and illustrated by the single figure in that book reproduced as fig. 3.4.

How long does this evolutionary process take and how can we measure the passage of evolutionary time? It is unsatisfying to rely on the extrapolation of mutation rates measured in artificial laboratory experiments to the evolution of species over time in the real world. Real world conditions are much less stable or controlled than laboratory conditions, and furthermore the time scales of greatest interest for studying the evolution of species are much longer than can be achieved in the laboratory by even the most patient experimentalist. Traditionally, our understanding of evolutionary alterations depended upon two kinds of observations: comparisons among currently living species and examination of

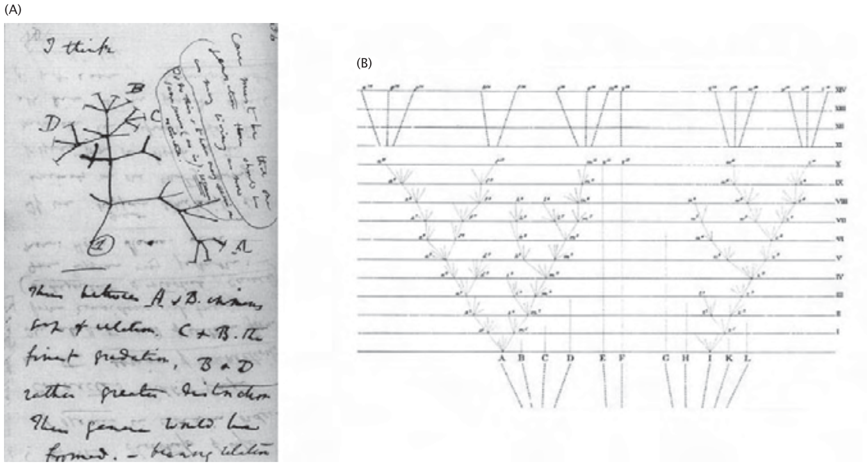


Figure 3.4: Two versions of Darwin's phylogenetic tree. (A) In his notebooks, Darwin drew the first version of what we now recognize as a common schematic demonstrating the relatedness of organisms. He introduced this speculative sketch with the words "I think" as his theory was beginning to take form. (B) In the final published version of *The Origin of Species*, the tree had assumed more detail showing the passage of time and explicitly indicating that most species have gone extinct.

the fossil record. Information about the age of particular fossils can be inferred from identification of the geological strata in which they are found, and also by examining the proportions of different radioisotopes which decay at a regular rate and thereby provide information about when the rock was formed.

Comparison of living species to ascertain degree of relatedness was carried out for many hundreds of years before the modern theory of evolution was first described. It is immediately obvious that some organisms are more closely related than others. For example, horses and donkeys are clearly more similar to each other than either is to a dog, but horses and dogs are more similar to each other than either is to a squid. These obvious morphological differences have been the basis of the science of systematics going back to Linnaeus in the 1700s. In the modern era of molecular genetics, we can more easily ascribe a universal metric for genetic similarity among organisms based on similarities and differences in DNA sequence. As a population evolves over time, its DNA complement will change by several mechanisms. First, small scale mutations or large scale rearrangements of its genome may occur (an illustration of the consequences of this kind of rearrangement is shown in fig. 3.5). Second, it may acquire new genes or even entire groups of genes by horizontal transfer from other organisms. And third, it may simply lose large chunks of DNA. Thus different organisms contain different complements of genes as well as sequence differences between homologous copies of the same gene. The term homologous refers to descent from a common ancestor. For example, ribosomal RNAs are homologous in all cells. In chap. 18, we will give some examples of ribosomal RNA sequences and show how they can be used to build a universal phylogenetic tree. One example of a tree based on ribosomal RNA sequences that attempts to show the relatedness among all branches of existing life is shown in fig. 3.6.

Phylogenetic trees established by molecular methods tend to be in excellent agreement with analogous trees of similarity based on morphological or biochemical criteria as have been established by botanists, zoologist and microbiologists over the past several hundred years. We will examine statistical methods for constructing such trees in chap. 18.

What does any of this have to do with the determination of evolutionary time? In the laboratory, we can observe that certain types of changes in DNA sequences within a population happen frequently (for example, single point mutations changing a C to a T) while others happen more rarely (crossover events reversing the order of all the genes within a segment of a chromosome). We can even measure the time constants that characterize such events. If we assume that these kinds of mutational events happen with the same frequency in wild populations as they do in the laboratory, then we can estimate divergence times for organismal populations based on calculating how long on average it would take to achieve the observed number of sequence alterations given known rates of sequence alteration events. In a few cases, these time estimates can be anchored by reference to the fossil record. In reality, inferring evolutionary time from sequence similarity is fraught with peril because not all sequence alterations are equally likely to be randomly incorporated into the genetic heritage of a population of organisms. Some mutations will prove to be unfavorable

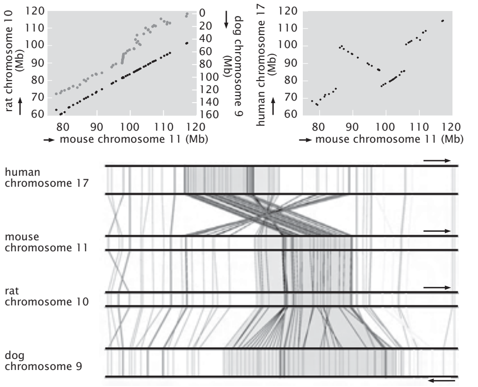
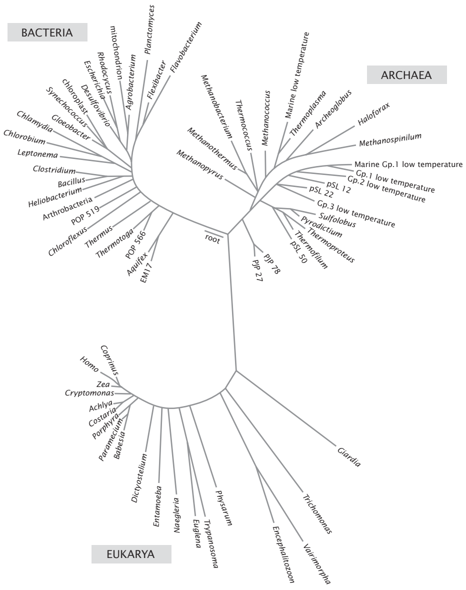


Figure 3.5: Inferring evolutionary relatedness by chromosome alignment. Equivalent regions of four chromosomes from mouse, rat, dog and human were compared to find the location of homologous genes. The graphs at top show the position of each gene in the rat, dog and human sequences as a function of their positions on the mouse sequence. Because little change has occurred in chromosomal structure between the mouse and the rat, the points representing the locations of homologous genes form a nearly perfectly straight line. On the equivalent chromosomal segment from the dog, the genes are again mostly in the same order, but the spacing between them has changed substantially. Comparing the human to the mouse, a large inversion can be detected. The same data is shown in a different form in the chart at the bottom. Each vertical line on the chromosome represents a particular gene and the diagonal lines between the chromosomes link up homologs between human and mouse, mouse and rat and rat and dog.



pk 3.4

Figure 3.6: Universal phylogenetic tree. This diagram shows the similarity among 16S ribosomal RNA sequence for representative organisms from all major branches of life on Earth.

for a given organism's lifestyle and individuals carrying those mutations will be eliminated from the population by natural selection. Other mutations will prove to be advantageous and organisms carrying those mutations will quickly outcompete other members of their species. These selection effects can make the sequence-determined evolutionary clock appear to run too slow or too fast. Biologists face challenges similar to those faced by astronomers. In the astronomical setting, continual refinements in cosmological distance scales based on various types of standard candle (light sources of known absolute intensity) have led to increasingly refined measurements of astronomical distance. Similarly, biologists have a number of different standard stopwatches that can be used to calibrate the flow of evolutionary time.

3.1.3 The Cell Cycle and the Standard Clock

The *E. coli* Cell Cycle Will Serve as Our Standard Stopwatch

In fig. 2.1 we used the size of an *E. coli* cell as our standard measuring stick. Similarly, we now invoke the time scale of the *E. coli* cell cycle as our standard stopwatch. The goal of fig. 3.2 was to illustrate the variety of different processes that occur in cell biology and the time scales over which they are operative. As with our discussion of structural hierarchies, we once again use the trick of invoking *E. coli* as our reference, this time with the several thousand seconds of its cell cycle as our reference time scale.

As shown in fig. 3.27, the bacterial cell cycle will be defined as the time between the "birth" of a given cell resulting from division of a parental cell to the time of its own subsequent division. This cell cycle is characterized structurally by the segregation of the duplicated bacterial chromosome into two separate clumps and the construction of a new portion of the cell wall, or septum, that separates the original cell into two daughters. Because *E. coli* is a roughly cylindrical cell that maintains a nearly constant cross-sectional area as it grows longer, the total cell volume can be easily estimated simply from measuring the length and this also provides a guide as to the point in the cell cycle. As cell division proceeds, *E. coli* doubles in length and hence also doubles in volume. The time scale associated with the binary fission process of interest here is of order an hour (to within a factor of two), though division can take place in under 20 minutes under optimal growth conditions.

In the previous chapter, we argued that having a proper molecular inventory of a cell is a prerequisite to building models of many problems of biological interest. Here we argue that a similar "feeling for the numbers" is needed concerning biological time scales. How long does it take for an *E. coli* cell to copy its genome and is this rate consistent with the speed of the molecular machine (DNA polymerase) that does this copying? On what time scale do newly formed proteins in neurons reach the ends of their axons and can this be explained by diffusion? Often, the time scale associated with a given process will provide a clue about what physical mechanisms are in play. In addition, one of our biggest concerns in coming chapters will be to figure out under

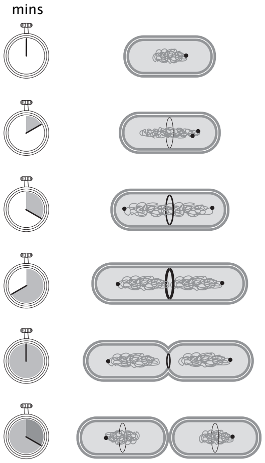


Figure 3.7: Schematic of an idealized bacterial cell cycle. A newborn cell shown at the top has a single chromosome with a single origin of replication marked by the dot. The cell cycle initiates with the duplication of the origin and DNA replication then proceeds in an orderly fashion around the circular chromosome. At the same time, a group of cell division proteins beginning with the tubulin analog FtsZ form a ring at the center of the cell that will dictate the future site of septum formation. As DNA replication proceeds and the cell elongates, the two origins become separated from each other with one traveling the entire length of the cell to take up residence at the opposite pole. As the septum begins to close down, the two chromosomal masses are physically separated into the two daughter cells where the cycle can begin anew.

what conditions we are justified in using the ideas from equilibrium physics (as opposed to nonequilibrium physics). The answer to this question will be determined by whether or not there is a separation of time scales and the only way we can know that is by having a feeling for what time scales are operative in a given problem. To that end, we begin by taking stock of the processes that an *E. coli* cell must make to copy itself.

For estimates in this book we will choose a standard for bacterial growth in a minimal defined medium with glucose as the sole carbon source. As mentioned previously, the rate of cell division can vary by more than tenfold depending upon nutrient availability and temperature, so we must define the terms under which we will proceed with our estimates. The choice of minimal media with glucose at 37 degrees Celsius is a practical one since many quantitative experiments have been performed under this condition. With sufficient aeration, *E. coli* in this medium typically double in the range of 40-50 minutes and we will use 3000 seconds as our canonical cell cycle time. In general, time scales for biological processes are much more variable than spatial scales, although it is true that rapidly growing *E. coli* are slightly larger than slowly growing *E. coli.* The difference in size may be an order of magnitude less than the difference in cycle time.

- **Estimate: Timing *E. coli*.** In chap. 2, we sized up *E. coli* by giving a series of rough estimates of its parts list. We now borrow those estimates to gain an impression of the rates of various processes in the *E. coli* cell cycle. The simple idea behind these estimates is to take the total quantity of material that must be used to make a new cell and to divide by the time (≈ 3000 seconds) of the cell cycle. When *E. coli* is grown on minimal media with glucose as the sole carbon source, six atoms of carbon are added to the cellular inventory for each molecule of glucose taken up. In the previous chapter, we estimated that the number of carbon atoms it takes to double the material in a cell so that it can divide in two (just the construction material) is of order 10^{10}. For this estimate we ignored the material released as waste products and the reader will have the opportunity to estimate this contribution in the problems at the end of the chapter. We are also deliberately ignoring the glucose molecules that must be consumed to generate energy for the synthesis reactions - this topic will be taken up in chap. 5. At this point, we can estimate the rate of sugar uptake required simply to deliver the 10^{10} carbon molecules necessary for building the material of the new cell. 10^{10} carbons must be captured over 3000 seconds with 6 carbons per glucose molecule, giving an average rate of roughly 5×10^5 glucose molecules every second.

  Of course, having the carbon present is not the same as the macromolecular synthesis required to make a new cell. One of the most important processes in the cell cycle is replication. Given that the complete *E. coli* genome is about 5×10^6 base pairs (bp) in size, we can estimate the required rate of replication as

$$\frac{dN_{bp}}{dt} \approx \frac{N_{bp}}{\tau_{cell}} \approx \frac{5 \times 10^6 \text{ bp}}{3000 \text{ sec}} \approx 2000 \text{ bp/sec.} \tag{3.1}$$

Similarly, the rate of protein synthesis can be estimated by recalling from the previous chapter that the total number of proteins in *E. coli* is roughly 3×10^6, implying a protein synthesis rate of

$$\frac{dN_{protein}}{dt} \approx \frac{N_{protein}}{\tau_{cell}} \approx \frac{3 \times 10^6 \text{proteins}}{3000 \text{sec}} \approx 1000 \text{ proteins/sec.} \tag{3.2}$$

Note that we have rounded to the nearest thousand. A similar estimate can be performed for the rate of lipid synthesis resulting in

$$\frac{dN_{lipid}}{dt} \approx \frac{N_{lipid}}{\tau_{cell}} \approx \frac{5 \times 10^7 \text{lipids}}{3000 \text{sec}} \approx 20,000 \text{ lipids/sec.} \tag{3.3}$$

Yet another intriguing aspect of the mass budget associated with the cell cycle is the control of water content within the cell. Recalling our estimate from the previous chapter that an *E. coli* cell has roughly 10^{11} water molecules, results in the estimate that the rate of water uptake during the cell cycle is

$$\frac{dN_{H_2O}}{dt} \approx \frac{N_{H_2O}}{\tau_{cell}} \approx \frac{10^{11} \text{waters}}{3000 \text{sec}} \approx 3 \times 10^7 \text{ waters/sec.} \tag{3.4}$$

This rate of water uptake can be considered slightly differently by working out the average mass flux across the cell membrane. The flux is defined as the amount of mass crossing unit area per unit time and in this instance is given by

$$j_{water} \approx \frac{dN_{H_2O}/dt}{A_{E.coli}} \approx \frac{3 \times 10^7 \text{waters/sec}}{6 \times 10^6 nm^2} \approx 5 \text{ waters/nm}^2 \text{ sec,} \tag{3.5}$$

though we also note that this mass transport is mediated primarily by proteins which are distributed throughout the membrane.

We argue that each of these estimates tells us something about the nature of the machinery that mediates the processes of the cell. In remaining sections, these estimates will serve as our jumping off point for estimating the rate at which individual molecular machines carry out the processes of synthesis and transport needed to support metabolism and the cell cycle.

3.1.4 Three Views of Time in Biology

Modern humans have built much of the activity of our societies around an obsession with absolute time. This obsession is revealed by the propensity for events to occur at a certain time of day, for example, class starts at 9am, or scheduling our activity by measured blocks of time, for example, you must practice the piano for half an hour. It is not clear, however, that other organisms relate to time in this manner. In the remainder of the chapter we will discuss three different views of time that seem to be important to life and we will term them *procedural time*, *relative time* and *manipulated time*.

In the previous chapter we explored the question of why biological things are a certain size and the ultimate reason is the finite extent of the atoms that make up biological molecules. Here we are trying to understand why biological processes take a certain amount of time, a difficult task. For the most part, the size of things does not strongly depend on environment and external conditions, but the time scale of processes often does. For example, bacteria growing in leftover potato salad will replicate rapidly when the salad is left on a picnic table in full sun but much more slowly in a refrigerator. The fundamental reason for the difference in replication rates as a function of temperature can be attributed to the slowing of the many individual enzymatic steps that must take place for the cell to double in size and divide. In this sort of context, it appears that organisms pay attention to *procedural time* rather than absolute time: they do something for as long as it takes to get it done since there is some procedure such as DNA replication dictated by an enzymatic rate. A particularly interesting class of procedural time mechanisms are those that organisms use to build clocks that are extremely good at keeping track of absolute time without regard to perturbation by external conditions. One fascinating example of this that we will explore in more detail later in the chapter is the diurnal clock that enables an organism to perform different acts at different times of the day, even in the absence of external signals such as the rising and setting of the sun. For these clocks to work, organisms must have a way to convert procedural time into absolute time so as to ignore external conditions, including temperature.

Although calculating procedural time for a process of interest can often put a lower limit on how fast that process can occur, cells often seem to put as much effort into making sure that processes occur in the correct order as in making sure that they occur quickly. In the context of cell division, for example, it would be disastrous for a cell to try to segregate its chromosomes into the two daughters until the process of DNA replication is complete. The result would be that at least one daughter would lack the full genetic complement of the mother cell. We will refer to processes where one must be complete before another can start under the category of *relative time* (i.e. before or after rather than how long).

Third, and perhaps most interestingly, it appears that living organisms are rarely content to accept time as it is. In some cases, they seem to be impatient, demanding that their life processes occur more quickly than permitted by the underlying chemical and physical mechanisms. Rate acceleration by enzyme catalysis is a prime example. In other cases, they seem to delay the intrinsic proceeding of events, freezing time in suspended animation as in formation of bacterial spores that can survive for hundreds or thousands of years, only to be reanimated when conditions become favorable. In section 3.4, we will argue that these processes are examples of what we will refer to as *manipulated time*.

3.2 Procedural Time

The underlying idea of measurements of procedural time is simply that the chemical and physical transformations characteristic of life do not happen instantaneously. Complex processes can be thought of as being built up from many small steps, each of which takes a finite amount of time. For many biological processes that are intrinsically repetitive such as the replication of DNA or the synthesis of proteins, the same step is used over and over again; addition of single nucleotides to a growing daughter strand or addition of single amino acids to a growing polypeptide chain. In this section on procedural time, we will begin by making some estimates about these processes of the central dogma as an example of the general issues of computing procedural time for multi-step biological processes. Then we will move on to the interesting special examples of clocks and oscillators where procedural times are calibrated so that cell cycles and diurnal cycles can follow the constant ticking of a reliable clock.

3.2.1 The Machines (or Processes) of the Central Dogma

The Central Dogma Describes the Processes Whereby the Genetic Information Is Expressed Chemically

One of the most important classes of processes in cellular life are those associated with the so-called Central Dogma of molecular biology. The suite of processes associated with the Central Dogma are those related to the polymerization of the polymer chains that make up the nucleic acids and proteins that are at the heart of cellular life. The fundamental processes of replication, transcription and translation and their linkages are shown in fig. 3.8. The basic message of this "dogma" in its least sophisticated form is that DNA leads to RNA which leads to proteins. From the standpoint of cellular timing, the processes of the Central Dogma will serve as a prime example of procedural time. A typical circular bacterial genome, for example, is replicated by just two DNA polymerase complexes that take off in opposite directions from the origin of replication and each travel roughly halfway around the bacterial genome to meet on the opposite side. The time to replicate a bacterial genome is governed by the rate at which these polymerase motors travel along to copy the roughly 5×10^6 bp of the bacterial genome. Similarly, the time to synthesize a new protein is governed by the rate of incorporation of amino acids by the ribosome.

The Processes of the Central Dogma Are Carried Out By Sophisticated Molecular Machines

One of the primary processes shown in fig. 3.8 is the copying of the genome, also known as replication. DNA replication must take place before a cell divides. As shown in fig. 3.8, the process of DNA replication is mediated by a macromolecular complex (the replisome) which has a variety of intricate parts such as the enzyme DNA polymerase which incorporates new nucleotides onto the nascent DNA molecule, and helicases and primases which tear open the DNA

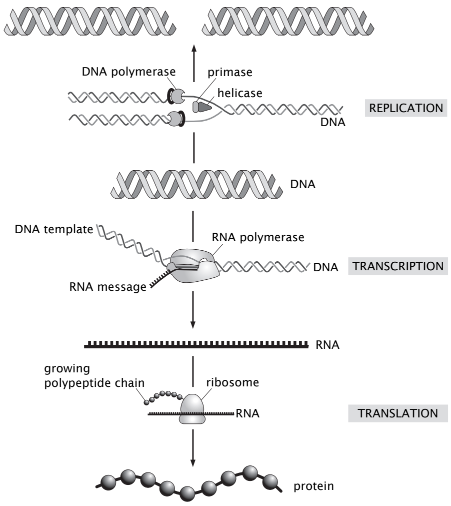


Figure 3.8: The processes of the central dogma. DNA is replicated to make a second copy of the genome. Transcription refers to the process when RNA polymerase makes a mRNA molecule. Translation refers to the synthesis of a polypeptide chain whose sequence is dictated by the arrangement of nucleotides on mRNA.

at the replication fork and prime the polymerization reaction.

The DNA molecule serves as a template in two different capacities. As described above, a given DNA molecule serves as a template for its own replication. However, in its second capacity as the carrier of the genetic material, a DNA molecule must also dictate the synthesis of proteins (the expression of its genes). The first stage in this process of gene expression is the synthesis of a messenger RNA molecule (mRNA) with a nucleotide sequence complementary to the DNA strand from which it was copied, which will serve as the template for protein synthesis. This transcription process is carried out by a molecular machine called RNA polymerase that is shown schematically in fig. 3.8. In eukaryotes, transcription takes place in the nucleus while subsequent protein synthesis takes place in the cytoplasm so there must be an intermediate step of mRNA export.

Once the messenger molecule (mRNA) has been synthesized, the translation process can begin in earnest, (RNA → Protein). As already described in section 2.1.4, translation is mediated by one of the most fascinating macromolecular assemblies, namely, the ribosome. The ribosome is the apparatus that speaks both of the two great polymer languages and in particular, forms a string of amino acids (a polypeptide chain) which are dictated by the codons (collections of three letters) on the mRNA molecule. The structure of the ribosome is indicated in cartoon form in fig. 3.8. As might be expected for a bilingual machine, the ribosome contains structural components of both RNA and protein. The two halves of the ribosome clamp a messenger RNA and then the ribosome moves processively down the length of the mRNA. As the ribosome moves along, successive triplets of nucleotides are brought into registry with active sites in the ribosomal machinery that align special RNA molecules (tRNA), charged with various amino acids, to recognize the complementary triplet codon. Subsequently, the ribosome catalyzes transfer of the correct amino acid from the tRNA onto a growing polypeptide chain and releases the now empty tRNA. As shown in fig. 3.9, the nascent mRNA molecules in bacteria are immediately tackled by ribosomes so that protein translation can occur before transcription is even finished.

The timing of all three of these processes is dictated by the intrinsic rate at which these machines carry out their polymerization reactions. All of them can be thought of in the same framework as repetitions of N essentially identical reactions, each of which takes a time Δt to perform. We will now estimate total times for each of the three central processes of the central dogma.

- **Estimate: Timing the Machines of the Central Dogma.** The estimates concerning the mass budget of dividing cells from chap. 2 can be used as a springboard for contemplating the rates of the machines that mediate the processes of the central dogma. In our first estimate, we expand upon the estimate of the rate at which the genome of an *E. coli* cell is copied performed earlier in the chapter with the aim of learning more about the speed of the DNA replication complex. DNA replication in bacteria such as *E. coli* is undertaken by two replication complexes which

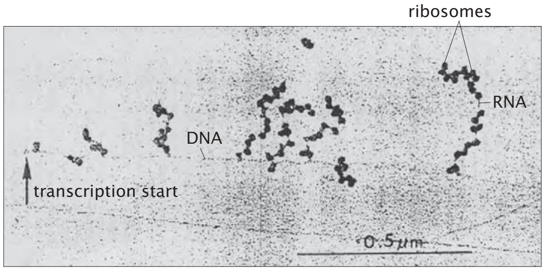


Figure 3.9: Electron microscopy image of simultaneous transcription and translation. The image shows bacterial DNA and its associated mRNA transcripts, each of which is occupied by ribosomes.

travel in opposite directions away from the origin of replication on the circular chromosome.

Given that the complete *E. coli* genome is about 5×10^6 base pairs (bp) in size and it is copied in the 3000 seconds of the cell cycle, we already found that the rate of DNA synthesis is roughly 2000bp/sec, or 1000bp/sec per DNA replication complex (replisome). Biochemical studies have found rates for the DNA polymerase complex in the 250-1000 bp/s range. As we have mentioned, *E. coli* are actually capable of dividing in much less than 3000 seconds, in fact, as little as 1000 seconds, although their DNA replication machinery cannot proceed any faster than this absolute speed limit. How do they pull it off? For now we will leave this as an open mystery and will return to the question in the final section of the chapter on manipulating time.

For a bacterial cell, transcription involves the synthesis of messenger RNA molecules with a length of roughly 1000 bases. Our reasoning is that the typical protein has a length of 300 amino acids, with 3 bases needed to specify each such amino acid. Both bulk and single-molecule studies have revealed that a characteristic transcription rate is tens of nucleotides per second. Using 40 nucleotides/sec, we estimate the time to make a typical transcript is roughly 25 seconds.

Yet another process of great importance in the central dogma is protein synthesis by ribosomes. Recall from our estimates in the previous chapter that the number of proteins in a "typical" bacterial cell like *E. coli* is of order 3×10^6. This suggests, in turn, that there are of order 9×10^8 amino acids per *E. coli* cell which are produced over the roughly 3000 seconds of the cell cycle. We have made the assumption that each protein has 300 amino acids. This implies that the mean rate of amino acid incorporation per second is given by

$$\frac{dN_{aa}}{dt} \approx \frac{9 \times 10^8 \text{amino acids}}{3000 \text{seconds}} \approx 3 \times 10^5 aa/sec. \tag{3.6}$$

The number of ribosomes at work on synthesizing these new proteins is roughly 20,000 which implies that the rate per ribosome is 15 aa/sec, while the measured value is 25 amino acids incorporated per second. These numbers also imply that the mean time to synthesize a typical protein is roughly 20 seconds.

One of our conclusions is that the rate of protein synthesis by the ribosome is slower than the rate of mRNA synthesis by RNA polymerase. However, as shown in fig. 3.9, multiple ribosomes can simultaneously translate a single mRNA by proceeding in a linearly, orderly fashion and indeed, multiple RNA transcripts may exist in different degrees of completion, being transcribed from the same genetic locus. Thus when considering the net rates of processes in cells, the number of players is clearly as important as the intrinsic rate.

3.2.2 Clocks and Oscillators

In the context of the central dogma, we have described measurements of procedural time for processes that essentially happen once and run to completion such as the synthesis of a protein molecule. However, many cellular processes run in repeated regular cycles. These cyclic or oscillatory processes frequently represent control systems where procedural times of some subprocess can be used to set the oscillation period. Two widely studied examples are the oscillators used to drive the cell division cycle and the mechanisms governing behavioral switches between day time and night time which will be explored in detail below. These daily clocks are called circadian or diurnal oscillators. Other everyday oscillators run the beating of our hearts and the pattern of our breathing.

Developing Embryos Divide on a Regular Schedule Dictated by an Internal Clock

One of the best understood examples of an oscillatory clock used by cells is seen in the early embryonic cell cycle of many animals. The best studied example is the South African clawed frog, *Xenopus laevis.* After the giant egg (1mm) is fertilized, a cell division cycle proceeds roughly every twenty minutes until the egg has been cleaved into approximately four thousand similar sized cells. The regularity and synchrony of these cell divisions reflects an underlying oscillatory clock based on a clever manipulation of procedural time. The clock starts each cell division with the synthesis of a protein called cyclin. Cyclin is made from a relatively rare mRNA. As a result, the protein accumulates slowly. The biological function of cyclin is to activate a protein kinase, an enzyme which covalently attaches phosphate groups to amino acid sidechains on other proteins. This process, known as phosphorylation, is one of the key ways that proteins are controlled after translation. Essentially, the protein is inactive in the absence of its phosphate group. Kinase activation cannot begin until the cyclin protein has accumulated to a certain threshold level. After the kinase is activated, one of its targets is an enzyme which in turn catalyzes the destruction of the cyclin protein. All cyclin in the cell is quickly destroyed within RP seconds, resetting the clock to its zero position.

The regularity of this oscillatory clock depends upon several measurements of procedural time. First, accumulation of the cyclin protein to its threshold level depends upon the rate of ribosomal synthesis of that protein. Second, activation of the cyclin-dependent protein kinase kicks off a second procedural time measurement which reflects the length of time required by the kinase to encounter and phosphorylate its enzymatic substrates. Third, the degradation of the cyclin protein also requires a fixed, but brief amount of time. The sum of these three procedural times gives the total period of the clock. The outcome of these molecular events in terms of molecular concentrations is illustrated in fig. 3.10. Just as for all the examples of procedural time described above, the amount of absolute time in seconds, minutes or hours may change depending upon external conditions such as temperature.

This cyclin driven cell cycle oscillator is one example of a very general category of two-component oscillatory systems found throughout biology. A simplified idealized representation of such an oscillator is shown in fig. 3.11(A), while a more accurate representation of the real cell cycle control system from the yeast *S. cerevisae* is shown in fig. 3.11(B). The mathematics of these oscillators is explored in the problems at the end of the chapter.

Diurnal Clocks Allow Cells and Organisms to Be on Time Everyday

A second example of the use of procedural time to build a clock is when cells arrange a series of molecular processes in such a way that they can measure an absolute time. Unlike the cell cycle clock, it is critical that the diurnal clock not change its period when the temperature changes such as during the change of seasons. Many organisms perform some specific task at the same time everyday. A spectacular example is shown in fig. 3.12 where an animal alters the light sensitivity of its eyes in anticipation of sundown. While we might imagine that these kind of daily changes are triggered by, for example, the intensity of sunlight, it has been demonstrated for many organisms that they continue to perform their diurnal cycle even when kept in the dark. Direct observation of these cycles over long periods of time in cyanobacteria have demonstrated that they can operate with tight precision over a week time scale without any external cues about absolute time.

Different organisms use information about the time of day for vastly different purposes. Nevertheless, as illustrated in fig. 3.13, the molecular circuitry governing their circadian rhythms conserves certain common features. Generally, these systems include positive elements which activate transcription of so-called clock genes which drive rhythmic biological outputs as well as promoting the expression of negative elements that inhibit the activities of the positive elements. Phosphorylation of the negative elements leads to their degradation allowing them to restart the cycle. Although the circadian oscillators are capable of continuing to measure time in constant light or constant darkness, they can nevertheless accept inputs from environmental signals such as the sun to reset their phase. Humans commonly experience the inefficiency of the phase resetting mechanisms as the phenomenon of jetlag.

The circadian oscillator known to function with the fewest components is

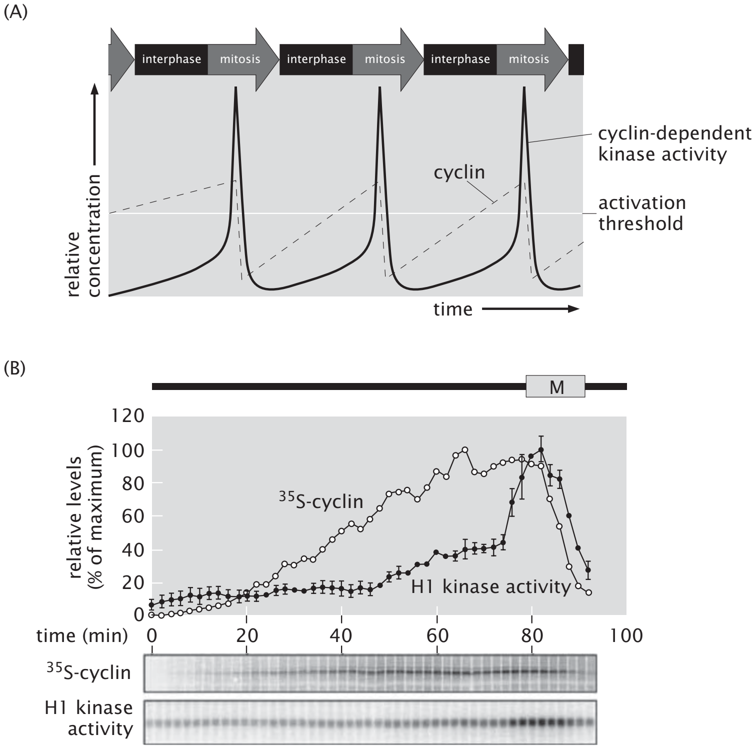


Figure 3.10: The oscillatory cell-cycle clock. This diagram shows the procedural events that underlie the regular oscillations of the cell-cycle clock in the *Xenopus laevis* embryo. Cyclin protein concentration rises slowly over time until it reaches a threshold at which point it activates cyclin-dependent kinase. Cyclin-dependent kinase activity increases sharply at this threshold and in turn activates enzymes involved in cyclin protein degradation. Once the degradation machinery is turned on, cyclin protein levels quickly fall back to zero. Cyclin-dependent kinase activity also falls and the degradation machinery inactivates. This oscillatory cycle is repeated many times.

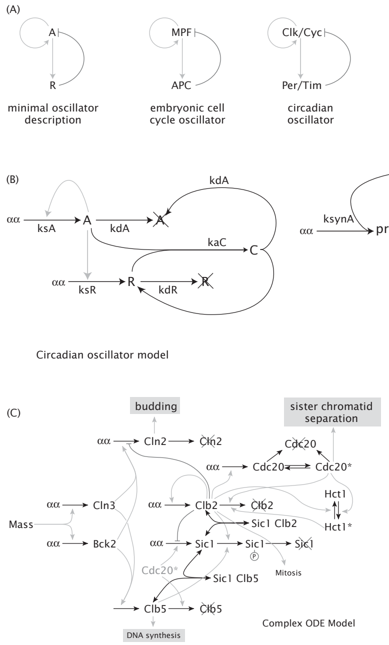


Figure 3.11: Logic diagrams for construction of cell cycle oscillators. (A) The minimal oscillator requires only two components. The first component activates the second component, for example by catalyzing its synthesis. The second component inhibits the first, for example by catalyzing its degradation. (B) A biochemically realistic representation of the cell cycle oscillator in yeast is outrageously more complicated. This is because the real oscillator must work under a wide variety of conditions, be insensitive to fluctuations in the concentrations or activities of its components, and be subject to multiple kinds of regulatory inputs.

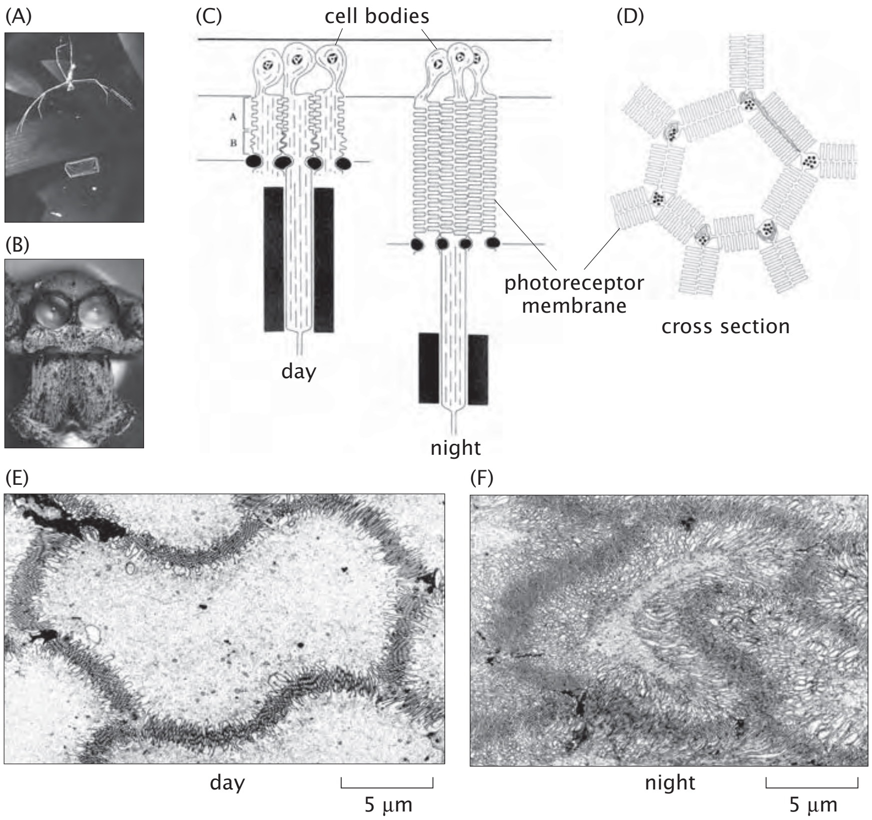


Figure 3.12: An extreme example of a structural change driven by the diurnal clock. (A) The net-casting spider, *Deinopis subrufa*, is a nocturnal hunter with an unusual strategy. It spins a small net which it holds with its legs and tosses to entangle unwary prey passing by. (B) In order to see the prey and know when to toss its net, the spider must have excellent night vision. Two of its eight eyes are extremely enlarged and exquisitely light sensitive. (C) The light sensitivity of the spider's eyes change by a factor of approximately one thousand between daytime and nighttime. During the day, the photoreceptor cell processes are short and fairly disorganized. At night, the total amount of membrane containing light sensors increases both by lengthening of the cells and by the construction of convoluted membrane folds, all packed with photoreceptor molecules. (D) In cross section, the photosensitive membranes of neighboring cells abut each other forming a regular tile-like pattern. (E) A cross section through the photoreceptor cells of a spider sacrificed during the day shows relatively modest thickening of the boundary membranes. (F) An equivalent section taken from a spider sacrificed at night shows a vast increase in the number and size of the membrane folds. At dawn, these membranes will all be degraded only to be resynthesized the following dusk.

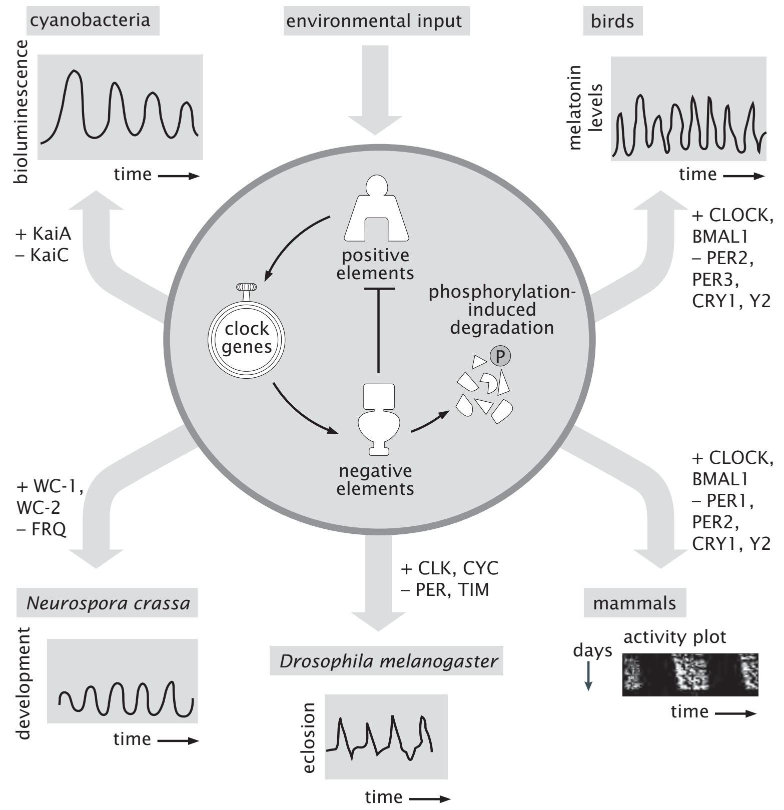


Figure 3.13: Schematic showing generic features of circadian clocks. Circadian clock mechanisms are autonomously driven oscillators that can be modulated by external inputs. Different organisms ranging from cyanobacteria to fungi to insects, birds and mammals use their circadian timers to regulate different kinds of biological outputs and also use very different kinds of protein components in the internal circuitry. The names of some of the genes involved in circadian oscillation for each of the species are shown, with a plus sign indicating positive elements and a minus sign indicative negative elements.

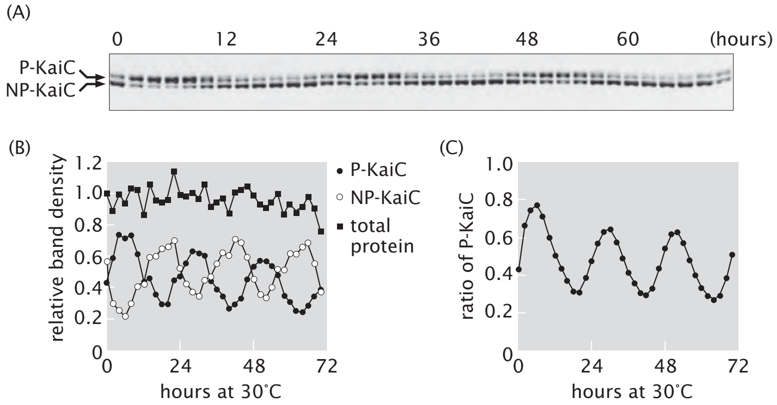


Figure 3.14: Reconstitution of the circadian oscillator. (A) In a mixture of purified KaiC protein together with KaiA, KaiB and ATP, the molecular weight of KaiC can be seen to shift up and down slightly over a twenty-four hour period. The upper band on this gel shows the phosphorylated form of KaiC protein and the lower band is the nonphosphorylated form. (B) Quantitation of the density of these two bands over time reveals that their concentration oscillates in a reciprocal manner such that the total amount of KaiC protein remains roughly constant. (C) The ratio of the amount of phosphorylated KaiC to total KaiC oscillates with a regular period of slightly under twenty-four hours.

the one from the photosynthetic cyanobacterium *Synechoccus elongatus*. This organism's clock requires just three proteins called KaiA, KaiB and KaiC and remarkably, it appears that neither gene transcription nor protein degradation is necessary for this clock to function. A purified mixture of just these three proteins together with ATP is capable of sustaining an oscillatory cycle of KaiC protein phosphorylation over periods of at least several days. KaiC is able to catalyze both its own phosphorylation and its own dephosphorylation. KaiA enhances KaiC auto-phoshorylation and KaiB inhibits the effect of KaiA. Fig. 3.14 shows the data supporting this remarkable finding.

3.3 Relative Time

The examples in the previous section on procedural time have emphasized the ways that cells set and measure the time that it takes to accomplish specific tasks. Some processes are rapid and others are slow because of intrinsic features or environmental circumstances. In the well-regulated life of the cell, it is frequently important that fast and slow processes not be permitted to run independently of one another, but instead be linked in a logical sequence that depends upon the cell's needs. In this context, we now turn to what we will call *relative time* which includes the governing mechanisms that ensure that related processes can be strung together in a "socks before shoes" fashion in which event A must be completed before event B can begin. Event C dutifully awaits the completion of event B before it begins, and so on.

3.3.1 Checkpoints and the Cell Cycle

In our initial discussion of the eukaryotic cell division cycle in the context of clocks and oscillators, we used the example of early embryonic divisions in the frog *Xenopus laevis* and asserted that the underlying driver was a simple two-component oscillator. Once past the earliest stages of embryonic development, the cell cycle becomes much more complex and in particular, becomes sensitive to feedback control from the cell's environment. The points in the cell cycle which are subject to interruption by external signals are referred to as checkpoints. These checkpoints ensure, for example, that chromosomes are not segregated until the DNA replication process is complete.

The Eukaryotic Cell Cycle Consists of Four Phases Involving Molecular Synthesis and Organization

Fig. 3.15 shows the key features of the eukaryotic cell cycle with an emphasis on the regulatory checkpoints that ensure that all processes will occur in the correct order. There is not a single universal time scale for the eukaryotic cell cycle, which can vary greatly from one cell type to the next. In the human body, some cells in the intestinal lining can divide in as little as 10-12 hours while others such as some tissue stem cells have cell cycles measured in days or weeks. The eukaryotic cell cycle is usually described in terms of four stages denoted as G_1, S, G_2 and M, with the M phase including the most recognized features, namely, nuclear division (mitosis) and cell division (cytokinesis), the two G phases (gap) as periods of growth and the S phase (synthesis) during which the nuclear DNA is replicated. Together, the phases other than the M phase constitute interphase. During interphase, the mass content of the cell increases as does its size.

If we use a cultured animal cell such as a fibroblast as our standard, G_1 is roughly 10 hours long and is characterized by a significant increase in cell mass and culminates in a checkpoint to insure sufficient cell size and appropriate environmental conditions to pass to the next stage. At this point, the cell examines itself for DNA damage. Any signs of damage such as double-strand breaks will trigger a checkpoint control that prevents the cell from initiating DNA replication until the damage is completely repaired. This checkpoint also ensures a critical aspect of the regulation of relative time for the cell's replication, specifically, that it must have grown to approximately twice its prior size before it is allowed to begin to divide. If this checkpoint is successfully passed then DNA replication can begin. In S phase, the eukaryotic DNA is replicated over a period of about 6 hours.

Following S phase and a shorter gap phase called G_2, another checkpoint verifies that every chromosome has been completely replicated before the initiation of assembly of the mitotic spindle, the microtubule-based apparatus that

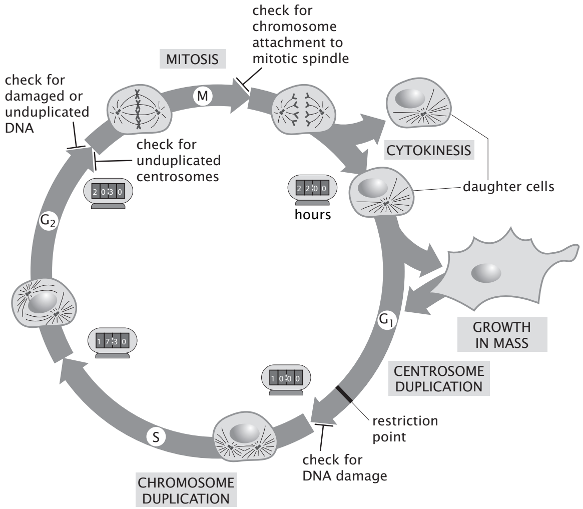


Figure 3.15: The eukaryotic cell cycle. This cartoon shows some of the key elements of the process of cell division. (adapted from Pollard and Earnshaw)

physically separates the chromosomes into the two daughter cells. This enforcement of relative time is particularly critical because if a cell were to try to segregate the chromosomes before replication was complete, then at least one of the daughters would inherit an incomplete copy of the genome. After passing this checkpoint, M phase begins. This relatively brief period of the cell cycle (of order one hour) involves most of the spectacular events of cell division that can be directly observed in a light microscope. Within M phase again, it is critical that events occur in the proper order. The bipolar mitotic spindle built from microtubules forms symmetric attachments to each pair of replicated sister chromosomes. When they have all been attached to the spindle, the chromosomes all suddenly and simultaneously release their sisters and are pulled to opposite poles. A spindle-assembly checkpoint ensures that every chromosome is properly attached before segregation begins. The molecular mechanisms governing the enforcement of relative time in the cell cycle involve protein phosphorylation and degradation events as well as gene transcription. In order to delve deeper into the principles governing the measurement and enforcement of relative time, we will now turn to a different example where gene transcription is the principal site of regulation.

3.3.2 Measuring Relative Time

There is great regulatory complexity involved in orchestrating the cell cycle. That is, the time ordering of the expression of different genes follows a complex program with certain parts clearly following a progression in which some pro-

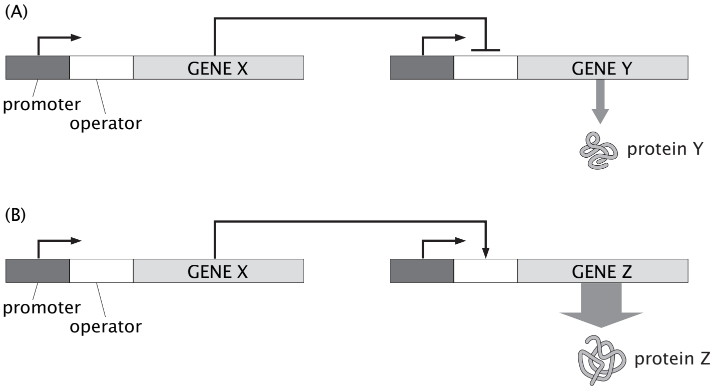


Figure 3.16: Network of interacting genes. Representation of a hypothetical genetic network where the output of the first gene represses or activates the expression of the second. (A) Repression process in which the output of gene X represses the expression of gene Y. (B) Activation process in which the output of gene X activates the expression of gene Z.

cesses must await others before beginning. By measuring the pattern of gene expression, it becomes possible to explore the relative timing of events in the cell cycle. To get a better idea of how this might work, we need to examine how networks of genes are coupled together.

Genetic Networks Are Collections of Genes Whose Expression Is Interrelated

Sets of coupled genes are shown schematically in fig. 3.16. For simplicity, this diagram illustrates how the product of one gene can alter the expression of some other gene. Perhaps the simplest regulatory motif is direct negative control in which a specific protein binds to the promoter region on DNA of a particular gene and physically blocks binding of RNA polymerase and subsequent transcription. This protein is itself the result of some other gene which can in turn be subject to control by yet other proteins (or perhaps the output of the gene that it controls). The second broad class of regulatory motif is referred to as activation and results when a regulatory protein (a transcription factor called an activator) binds in the vicinity of the promoter and "recruits" RNA polymerase to its promoter.

One way to measure the extent of gene expression is using a technique known as a DNA microarray. The idea is that a surface is decorated with fragments of DNA in an orderly arrangement, and the sequence of each spot on this array is different. To take a census of the current mRNA contents of a cell (which gives a snapshot of the current expression level of all genes), the cell is broken open and the mRNA contents are hybridized to the DNA on the microarray. DNAs on the surface which are complementary to RNAs in the cell lysate will

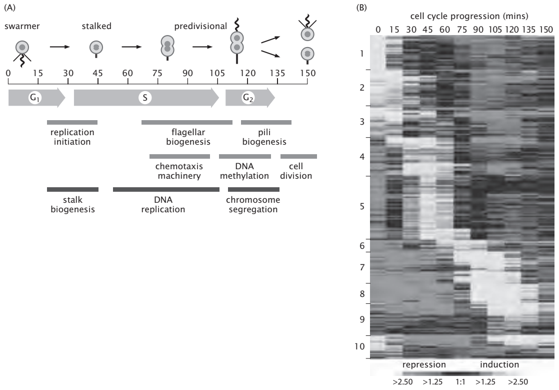


Figure 3.17: Gene expression during the cell cycle for *Caulobacter crescentus*. The figure shows a distinct time ordering for the expression of different genes over the course of the cell cycle. The microarray data shows how different batteries of related genes are expressed in a precise order.

hybridize with their complementary fragments. There is a bit more subtlety in the procedure than we describe since really the mRNA is turned into DNA first, but we focus on the concept of the measurement rather than its practical implementation. The intensity of the spots on the microarray report the extent to which each gene of interest was expressed. By repeating this measurement again and again at different time points, it is possible to profile the state of gene expression for a host of interesting genes at different times in the cell cycle. These measurements yield a map of the *relative* timing of different genes.

One of the key model systems for examining the bacterial cell cycle is *Caulobacter crescentus*. In a beautiful set of experiments roughly 20% of the *Caulobacter* genome was implicated in cell cycle control as a result of time varying mRNA concentrations which were slaved to the cell cycle itself. The idea of the experiment is to break open synchronized cells every fifteen minutes and to harvest their messenger RNA. Then by using a DNA microarray to find out which genes were being expressed at that moment, it was possible to put together a profile of which genes were expressed when. The outcome of this experiment is shown in fig. 3.17. What these experiments revealed is the relative timing of a series of events associated with the cell cycle.

The Formation of the Bacterial Flagellum Is Intricately Organized in Space and Time

A higher resolution look at the relative timing of cellular events is offered by the macromolecular synthesis of one of the key organelles for cell motility, namely, the bacterial flagellum. Fig. 3.18 shows the various gene products (FlgK, MotB, etc.) that are involved in the formation of the bacterial flagellum. Essentially, each of these products corresponds to one of the protein building

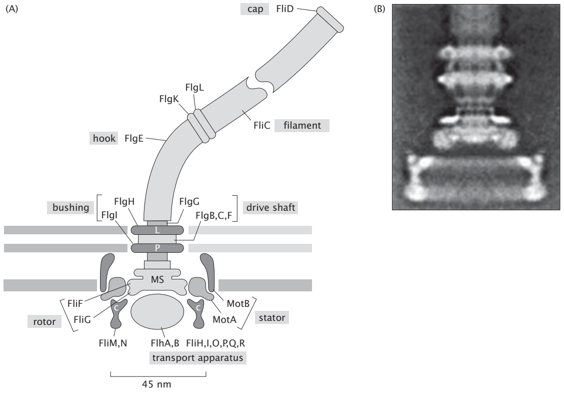


Figure 3.18: Molecular architecture of the bacterial flagellum. The schematic shows both the membrane-bound parts of the flagellar apparatus as well as the flagellum itself. The labels refer to the various gene products involved in the assembly of the flagellum.

blocks associated with flagellar construction. Once the flagella are assembled, the cell propels itself around by spinning them. The dynamical question posed in the experiment is to what extent is the expression of the genes associated with these different building blocks orchestrated in time.

The basic idea of the experiment is to induce the growth of flagella in starved *E. coli* cells and to use a reporter gene, namely, a gene leading to the expression of green fluorescent protein, to report on when each of the different genes associated with the flagellar pathway are being expressed. This experiment permits us to peer directly into the dynamics of assembly of the bacterial flagellum which reveals a sequence of events that are locked into succession in exactly the sort of "socks before shoes" way that we argued is characteristic of relative time. Fig. 3.19 shows the results of this experiment. To deduce a time scale from this figure, we consider the band of expression shown for "Condition A" and note that the roughly 15 genes are turned on over a period of roughly 180 minutes. This implies an approximate delay time between each product of roughly 12 minutes.

3.3.3 Killing the Cell: The Life Cycles of Viruses

Cells are not the only biological entities that care about relative timing. Once viruses have infected a host cell, they are like a ticking time bomb with an ever shortening fuse of early, middle and late genes. Once these genes have been expressed and their products assembled, as many as hundreds of new viruses emerge from the infected (and now defunct) cell to repeat the process elsewhere.

Viral Life Cycles Include a Series of Self-Assembly Processes

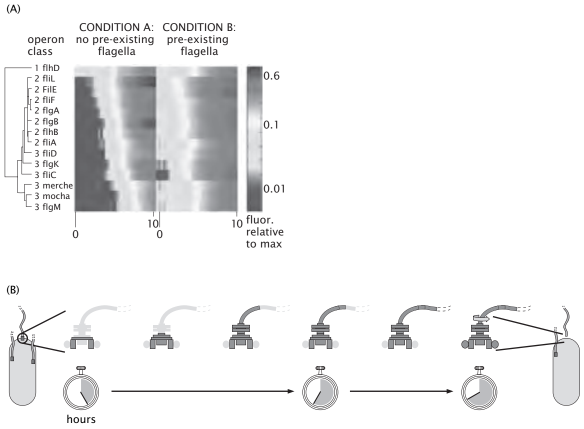


Figure 3.19: Timing of gene expression during the process of formation of the bacterial flagellum. (A) The extent of gene expression (as measured by fluorescence intensity) of each of the gene products as a function of time. (B) The cartoon show the timing of synthesis of different parts of the flagellum. (adapted from Kalir et al., Science, 292, 2080, 2001)

We have already described the cell cycle as a master process characterized by an enormous number of subprocesses. A more manageable example of an entire "life cycle" is that of viruses, which illustrate the intricate relative timing of biological processes. An example of the viral life cycle of a bacteriophage (introduced in the last chapter as a class of viruses that attack bacterial cells) is shown in fig. 3.20 which shows the key components in the life history of the virus. The key stages in this life cycle are captured in kinetic verbs such as infect, transcribe, translate, assemble, package and lyse. Infection is the process of entry of the viral DNA into the host cell. Transcription and translation refer to the hijacking of the cellular machinery so as to produce viral building blocks (both nucleic acid and proteins). Assembly is the coming together of these building blocks to form the viral capsid. Packaging, in turn, is the part of the life cycle when the viral genome is enclosed within the capsid. Finally, lysis refers to the dissolution of the host cell and the emergence of a new generation of phage to go out and repeat their life cycle elsewhere.

As illustrated by the cartoon in fig. 3.20 and in particular, by our use of the stopwatch motifs, the time between the arrival of the virus at the bacterial surface and the destruction of that very same membrane during the lysis phase when the newly formed viruses are released seems very short at 30 minutes. Indeed, one of our charters in the chapters that follow will be to come to terms with the 30 minute characteristic time scale of the viral life cycle and the various processes that make it up. On the other hand, though the absolute units (30 minutes) are interesting, it is important to emphasize that this set of processes is locked together sequentially (as with the synthesis of the bacterial flagellar apparatus).

Because of their stunning structures and rich lifestyles, we now examine a second class of viruses, namely, RNA animal viruses such as HIV. As shown in fig. 3.21, the infection process for these viruses is quite distinct from that in bacterial viruses. In particular, we note the presence of a membrane coat on the virus which allows the entire virus to attach to membrane-bound receptors on the victim cell. As a result of this interaction between the virus and the host cell, the virus is swallowed up by the cell which is under attack in a process of membrane fusion. Once the virus has entered the embattled cell, the genetic material (RNA) is released and reverse transcriptase creates a DNA molecule encoding the viral genome which is then delivered to the host nucleus and incorporated into its genome. After the genetic material has been delivered to the nucleus, a variety of synthesis processes are undertaken which result in copies of the viral RNA as well as fascinating polyproteins (the Gag proteins described in chap. 2) which are exported to the plasma membrane where they undergo an intricate process of self-assembly at the membrane of the infected cell. Once the newly formed virus is exported, it undergoes a maturation process resulting in new, fully infectious, viral particles. Each of these processes is locked in succession in a pageant of relatively timed events.

One of the intriguing features of viral life cycles is the variety of different dynamical mechanisms they exploit in order to produce fully infectious progeny. In this brief section, we have seen processes such as DNA translocation, transcription and reverse transcription, diffusion-based self-assembly, molecular-motor-assisted transport and membrane fusion and budding.

3.3.4 The Process of Development

One of the most compelling, mysterious and visually pleasing processes in biology is the development of multicellular organisms. Development refers to the orchestrated (both in space and time) division and differentiation of cells to construct the full organism and, like the cell cycle of individual cells, depends upon a fixed, relative ordering of events. Perhaps the most studied organism from the developmental perspective is the fruit fly *Drosophila melanogaster*. The process of *Drosophila* development was already schematized in fig. 3.2(A) and (B). The process of development refers to the disciplined outcome of an encounter between an egg and its partner sperm. In the hours that follow this encounter for the fruit fly, the nascent larva undergoes a series of nuclear divisions and migrations as shown in fig. 3.22. In particular, as the nuclei divide to the tenth generation (512 nuclei), they also start to collect near the surface of the developing larva forming the synctial blastoderm, a football shaped object with all of the cells localized to the surface. At the 13th generation, the individual nuclei are enclosed by their own membranes to form the cellular blastoderm. The structural picture by the end of this process is a collection of roughly 5000 cells which occupy the surface of a football shaped object (roughly) which is 500 μm in length and roughly 200 μm in cross-sectional diameter.

Accompanying these latter stages of the developmental pathway is the be-

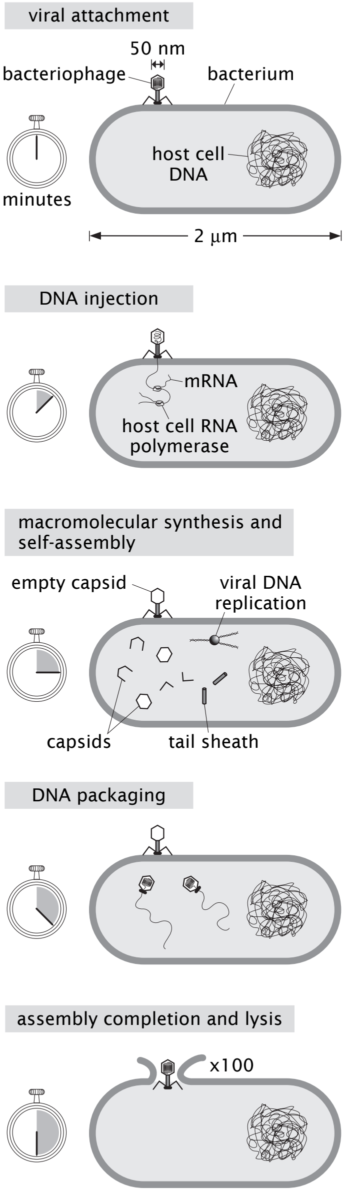


Figure 3.20: Timing the life cycle of a bacteriophage. The cartoon shows stages in the life cycle of a bacteriophage and roughly how long after infection these processes occur. Note that the head and the tail follow distinct assembly pathways and join only after they have separately assembled.

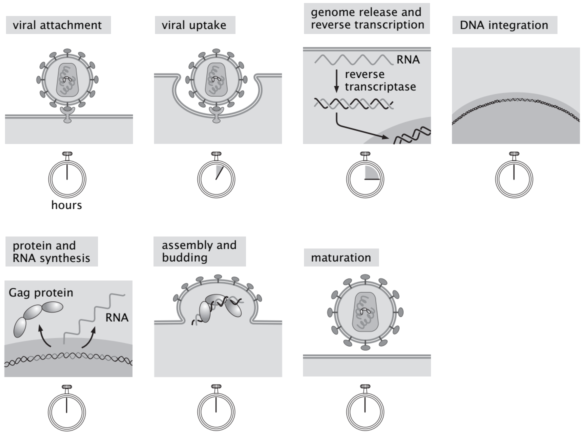


Figure 3.21: Stages in the life cycle of HIV. RP: Nigel needs to make a multiple snapshot picture with stopwatches - use this as the basis of that figure. This figure needs tons of work - the uptake needs binding, Gag insertion into membrane, etc.

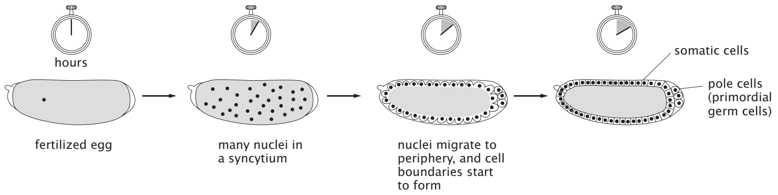

hours
somatic cells
pole cells (primordial germ cells)
fertilized egg
many nuclei in a syncytium
nuclei migrate to periphery, and cell boundaries start to form

Figure 3.22: Early development of the *Drosophila* embryo. After fertilization, the single nucleus undergoes a series of eight rapid divisions producing 256 nuclei that all reside in a common cytoplasm. At this point the nuclei begin to migrate towards the surface of the embryo while continuing to divide. After reaching the surface, cell boundaries form by invaginations of the plasma membrane. At this early stage, the cells that are destined to give rise to sperm or eggs segregate themselves and cluster at the pole of the embryo. All these events happen within roughly two hours.

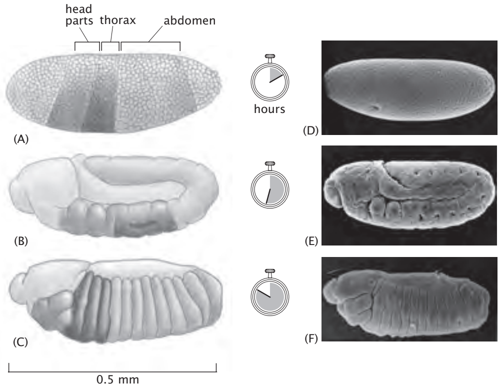


Figure 3.23: Early pattern formation in the *Drosophila* embryo. (A - C) show schematics of the shape of the *Drosophila* embryo at two hours, six hours and ten hours after fertilization, respectively. At two hours, no obvious structures have yet begun to form but the eventual fates of the cells that will form different parts of the animal's body have already been determined. By six hours, the embryo has undergone gastrulation and the body axis of the embryo has lengthened and curled back on itself to fit in the egg shell. By ten hours, the body axis has contracted and the separate segments of the animal's body plan have become clearly visible. On the right are scanning electron micrographs of embryos at each of these stages. Remarkably, all of this pattern formation takes place without any growth.

ginning of a cellular dance in which orchestrated cell movements known as gastrulation lead to the visible emergence of the macroscopic structures associated with the nascent embryo. Snapshots from this process are shown in fig. 3.23 with a time scale associated with the process of gastrulation is of order hours. We already proclaimed the importance and beauty of the temporal organization of gene expression associated with the cell cycle. We now add to that compliment by noting that during the development of the *Drosophila* body plan, there is an ordered spatial pattern of expression of genes with colorful names such as hunchback and giant, which determine the spatial arrangement of different cells.

These developmental processes make their appearance here because they too serve as an example of relative time. In particular, an example of the "socks before shoes" time ordering is the cascade of genes associated with the segmentation of the fly body plan into its anterior and posterior parts. The long axis of the *Drosophila* embryo is subject to increasing structural refinement as a result of a cascade of genes known collectively as segmentation genes. This collection of genes acts in a cascade which is a code word for precisely the kind of sequential processes that are behind *relative time* as introduced in this section. The first set of genes in the cascade are known as the gap genes. These genes divide the embryo into three rough regions, the anterior, middle and posterior. The gap genes have as protein products transcription factors that control the next set of genes in the cascade which are known as pair-rule genes. The pair-rule genes begin to form the identifiable set of seven stripes of cells. Finally, the segment polarity genes are expressed in 14 stripes.

- **Estimate: Timing Development.** A simple estimate of the number of cells associated with a developing organism can be obtained by assuming perfect synchrony from one generation to the next,

  $$\text{number of cells} \approx 2^N, \tag{3.7}$$

  where N is the number of generations. Further, if we assume that the cell cycle is characterized by a time τ_{cc}, then the number of cells as a function of time can be written simply as

  $$\text{number of cells} \approx 2^{t/\tau_{cc}}. \tag{3.8}$$

  Interestingly, in the early stages of *Drosophila* development, since it is only nuclear division (and hence, largely DNA replication) which is taking place, the mean doubling time is eight minutes. Thus after 100 minutes, roughly 10 generations worth of nuclear division will have occurred with the formation of the approximately 1000 cells which form the syncytial blastoderm.

3.4 Manipulating Time

Sometimes the cell is not satisfied with the time scales offered by the intrinsic physical rates of processes and has to find a way to beat these speed limits. For example, in some cases the bare rates of biochemical reactions are prohibitively slow relative to characteristic cellular time scales and as a result, cells have tied their fate to enzymatic manipulation of the intrinsic rates. In a similar vein, diffusion as a means of intracellular transport is ineffective over large distances. In this case, cells have active transport mechanisms involving molecular motors and cytoskeletal filaments which can overcome the diffusive speed limit. There are even more tricky ways in which cells manipulate time such as in the case of beating the bacterial replication limit. These examples and others will serve as the basis of our discussion of the way cells manipulate time.

3.4.1 Chemical Kinetics and Enzyme Turnover

Some chemical reactions proceed much more slowly than necessary for them to be biologically useful. For example, the hydrolysis of the peptide bonds that make up proteins would take times measured in years in the absence of proteases, which are the enzymes that cleave these bonds. Triose phosphate isomerase, one of the enzymes in the glycolysis pathway featured in fig. 2.23, is responsible for a factor of 10^9 speed up in the glycolytic reaction it catalyzes. What these numbers show is that even if a given reaction is favorable in terms of free energy, the energy barrier to that reaction can make it prohibitively slow. As a result, cells have found ways to manipulate the timing of reactions using enzymes as catalysts. Indeed, the whole of biochemistry is in some ways a long tale of catalyzed reactions many of which take place on time scales much shorter than milliseconds, whereas, in the absence of these enzymes, they might not take place in a year! The individual players in the drama of glycolysis such as hexokinase, phosphofructokinase, triose phosphate isomerase and pyruvate kinase reveal their identity as enzymes with the ending *ase* in their names. Enzymes are usually denoted by the ending *ase* and are classified according to the reactions they catalyze. There are six broad classes of enzymes: i) oxidoreductases, which catalyze oxidation-reduction reactions, ii) transferases, which transfer groups from one molecule to another, iii) hydrolases, which catalyze hydrolysis reactions, iv) lysases, which catalyze reactions where a group is removed from a substrate to form a double bond, v) isomerases, which catalyze isomerization reactions and vi) ligases, which are responsible for joining two molecules together.

The basic idea of enzyme action is depicted in fig. 3.24. For concreteness, we consider an isomerization reaction where a molecule starts out in some high energy state A and we interest ourselves in the transitions to the lower energy state B. The key point about the reaction rate is that, as shown in the figure, it depends upon the energy barrier separating the two states according to

$$\Gamma_{A\to B} = \nu_0 e^{-\Delta E/k_B T}, \tag{3.9}$$

where $\Gamma_{A\to B}$ is the transition rate with units of sec^{-1} and ν_0 is a frequency prefactor, also with units of sec^{-1}. Even though the energy of state B might be substantially lower than state A, the transitions can be exceedingly slow because of large barrier heights (i.e. $\Delta E >> k_B T$). The presence of an enzyme does not alter the end states or their energies, but it suppresses the barrier between the two states.

3.4.2 Beating the Diffusive Speed Limit

A second example of the way in which cells manipulate time is offered by the question of transport and trafficking. Organelles, proteins, nucleic acids, etc. are often produced in one part of the cell only to be transported to another part where they are needed. For example, the messenger RNA molecules produced in the nucleus need to make their way to the ribosomes which are found in the endoplasmic reticulum. One physical process that can move material around is passive diffusion.

Diffusion Is the Random Motion of Microscopic Particles in Solution

Ions, molecules, macromolecular assemblies and even organelles, wander around aimlessly as a result of diffusion. Diffusion refers to the random motions suffered by microscopic particles in solution, and is sometimes referred to

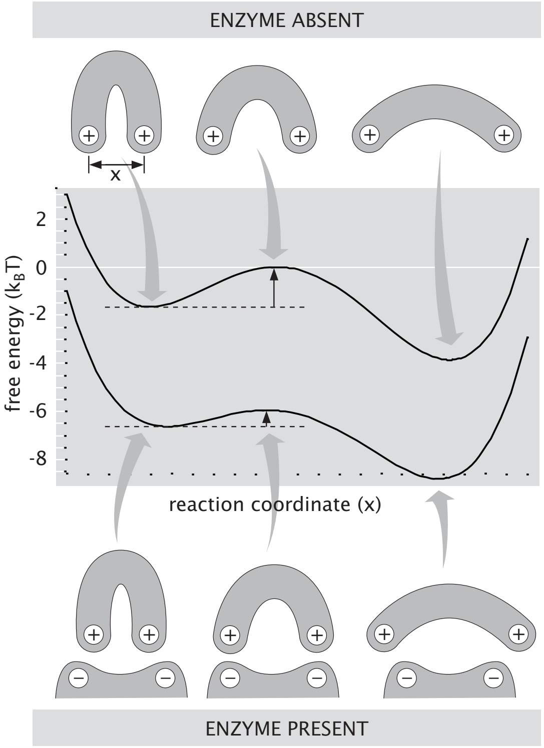


Figure 3.24: Enzymes and biochemical rates. A simple one-dimensional representation of the energy landscape for a biochemical reaction in the absence and presence of an enzyme to catalyze the reaction. The presence of the enzyme lowers the energy barrier and increases the rate of the reaction.

as Brownian motion in honor of the systematic investigations made by Robert Brown in the 1820s. Brown noticed the random jiggling of pollen particles suspended in solution, even for systems that are ostensibly in equilibrium and have no energy source. Indeed, so determined was he to find out whether or not this was some effect intrinsic to living organisms, he even examined exotic suspensions using materials such as the dust from the Sphinx and found the jiggling there too. The effects of Brownian motion are palpable for particles in solution which are micron size and smaller, exactly the length scales that matter to cells. Diffusion results from the fact that in the cell (and for microscopic particles in solution), deterministic forces are on nearly an equal footing with thermal forces. Thermal forces result from the random collisions between particles that can be attributed to the underlying jiggling of atoms and molecules. This fascinating topic will dominate the discussion of chap. 8.

- **Estimate: The Thermal Energy Scale.** One way to quantify the relative importance of the energy scale of a given process and thermal energies is by measuring the energy of interest in k_BT units. At room temperature, the thermal energy scale is

  $$k_BT = 1.38 \times 10^{-23}\text{J/K} \times 300K \approx 4.1 \times 10^{-21}J = 4.1pNnm. \tag{3.10}$$

  One way to see the importance of this energy scale is revealed by eqn. 3.9 (and will also be revealed by the Boltzmann distribution that says that the probability of a state with energy E_i is proportional to e^{-E_i/k_BT}). These expressions show that when the energy is comparable to k_BT, barriers will be small (and probabilities of microstates high). The numerical value ($k_BT \approx 4$ pN nm) is especially telling since many of the key molecular motors relevant to biology act with piconewton forces over nanometer distances, implying a competition between deterministic and thermal forces. This discussion tells us that for many problems of biological interest, thermal forces are on nearly an equal footing with deterministic forces arising from specific force generation.

Diffusion Times Depend Upon the Length Scale

One simple and important biological example of diffusion is the motion of proteins bound to DNA which can be described as one-dimensional diffusion along the DNA molecule. Another example is provided by the arrival of ligands to their specific receptors. The basic picture is that of molecules being battered about and every now and then ending up in the same place at the same time. To get a feeling for the numbers, it is convenient to consider one of the key equations that presides over the subject of diffusion, namely,

$$t_{\text{diffusion}} = \frac{x^2}{D}, \tag{3.11}$$

where D is the diffusion constant. This equation tells us that the time scale for a diffusing particle to travel a distance x scales as the square of that distance.

- **Estimate: Getting Proteins from Here to There.** For molecules and assemblies that move passively within the cell, the time scale can be estimated using eqn. 3.11. For a protein with a 5nm diameter the diffusion constant in water is roughly $100\mu m^2$/s; this estimate can be obtained from the Stokes-Einstein equation (to be discussed in more detail in chap. 12) which gives the diffusion constant of a sphere of radius R moving through a fluid of viscosity η at temperature T, as $D = k_B T/6\pi\eta R$. The time scale for such a typical protein to diffuse a distance of our standard ruler (i.e. across an *E. coli*) is

$$t_{\text{E. coli}} \approx \frac{L^2_{\text{E. coli}}}{D} \approx \frac{1\mu m^2}{100\mu m^2/sec} \approx 0.01s. \tag{3.12}$$

  This should be contrasted with the time scale required for diffusion to transport molecules from one extremity of a neuron to the other as shown in fig. 3.3. In particular, the diffusion time for the squid giant axon which has a length of the order of 10cm is $t_{\text{diffusion}} \approx 10^8$s! The key conclusion to take away from such an estimate is the impossibly long time scales associated with diffusion over large distances. Nature's solution to this conundrum is to exploit *active* transport mechanisms in which ATP is consumed in order for motor molecules to carry out directed motion.

Molecular Motors Move Cargo Over Large Distances in a Directed Way

In many instances, diffusion is too slow to be of any use for intracellular transport. To beat the diffusive speed limit, cells manipulate time with a sophisticated array of molecular machines (usually proteins) that result in directed transport. These processes are collectively powered by the consumption of some energy source (usually ATP). Broadly construed, the subject of active transport allows us to classify a wide variety of molecules as *molecular motors*. We have already seen the existence of such motors in a number of different contexts, with both DNA polymerase and RNA polymerase introduced in the previous section satisfying the definition of active transport.

Concretely, the class of motors of interest here are those that mediate transport of molecules from one place in the cell to another. Often, such transport takes place as vesicular traffic, with the cargo enclosed in a vesicle (a flexible spherical shell made up of lipid molecules in the form of a lipid bilayer) which is in turn attached to some molecular motor. These molecular motors travel in a directed fashion on the cytoskeletal network which traverses the cell. For example, traffic on microtubules runs in both directions as a result of two translational motors, kinesin and dynein. Molecular motor mediated transport on actin filaments is shown in cartoon form in fig. 3.25. Note that this cartoon is meant to indicate a rough idea of the relative proportions of the motors and the actin filaments on which they move and to convey the overall structural features, such as two heads, of the motors themselves. In addition, fig. 3.25(B) shows a time trace of the position of a fluorescently labeled myosin motor which

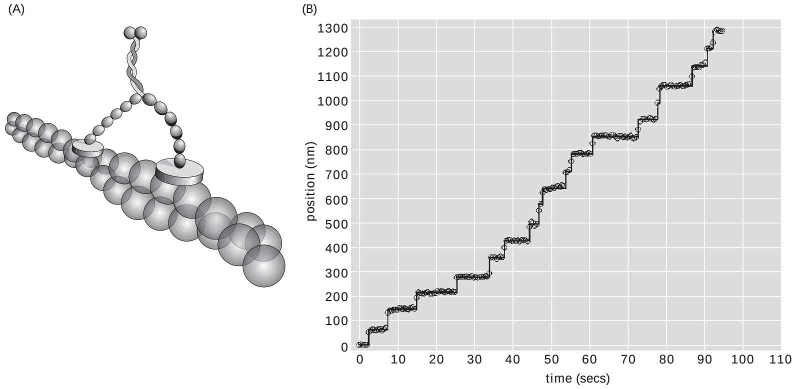


Figure 3.25: Motion of myosin on an actin filament. (A) Schematic of the motor on an actin filament. Note that the step size is determined by the periodicity of the filament. (B) Position as a function of time for the motor myosin V as measured using single-molecule techniques.

illustrates the discrete steps of the motor, also permitting a measurement of the mean velocity.

- **Estimate: Getting Proteins from Here to There, Part 2.** We have already noted that biological motility is based in large measure on diffusion. On the other hand, there are a host of processes that cannot wait the time required for diffusion. In particular, recall that our estimate for the diffusion time for a typical protein to traverse an axon was a whopping 10^8 seconds, or roughly three years. For comparison we can estimate the transport time for kinesin moving on a microtubule over the same distance. As the typical speed of kinesin in a living cell is 1μm/s, the time for it to transport a protein over a distance of 10cm is 10^5 seconds, or just over a day.

  To see these ideas play out more concretely, we can return to fig. 3.3(C). The classic experiment highlighted there traces the time evolution of radioactively labeled proteins in a neuron. What the figure shows is that the radiolabeled proteins travel roughly 18mm in 12 days, which translates into a mean speed of roughly 20 nm/sec. Observed axonal transport speeds for single motors are a factor of ten or more larger, but we can learn something from this as well. In particular, motors are not perfectly "processive" - that is, they fall off of their cytoskeletal tracks occasionally and this has the effect of reducing their mean speed. Observed motor velocities are reported, on the other hand, on the basis of tracking individual motors during one of their processive trajectories.

Membrane Bound Proteins Transport Molecules From One Side of a

Membrane to the Other

Another way in which cells manipulate transport rates is by selectively and transiently altering the permeability of cell membranes through protein channels and pumps. Many ionic species are effectively unable to permeate (at least on short time scales) biological membranes. What this means is that concentration gradients can be maintained across these membranes until or if these protein channels open which then permits a flow of ions down their concentration gradient. In fancier cases such as indicated schematically in fig. 15.12, ions can be pumped up a concentration gradient through mechanisms involving ATP hydrolysis. In fig. 3.2(RP) we showed the process of ion transport across a membrane with a characteristic time of microseconds.

- **Estimate: Ion Transport Rates in Ion Channels.** An ion channel embedded in the cell membrane can be thought of as a tube with a diameter of approximately $d = 0.5\text{nm}$ (size of hydrated ion) and a length $l = 5\text{nm}$ (width of the lipid bilayer). With these numbers in hand, and a typical value of the diffusion constant for small ions (eg. sodium), $D \approx 2000\mu\text{m}^2/\text{s}$, we can estimate the flux of ions through the channel, assuming that their motion is purely diffusive.

  To make this estimate, we invoke an approximate version of Fick's law (to be described in detail in chap. 8) which says that the flux (number of molecules crossing unit area per unit time) is proportional to the difference in concentration and inversely proportional to the distance between the two "reservoirs". Mathematically, this can be written as

  $$J_{\text{ion}} \approx D\frac{\Delta c}{l}, \tag{3.13}$$

  where Δc is the difference in ion concentration across the cell membrane. For typical mammalian cells the concentration difference for sodium or potassium is $\Delta c \approx 100\text{mM}$, and

  $$J_{\text{ion}} \approx 2000\mu\text{m}^2/\text{s} \times \frac{100 \times 6 \times 10^{20}\text{dm}^{-3}}{5\text{nm}} \approx 2 \times 10^7\text{nm}^{-2}\text{s}^{-1}. \tag{3.14}$$

  Given the cross-sectional area of a typical channel $A_{\text{channel}} = d^2\pi/4 \approx 0.2\text{nm}^2$, the number of ions traversing the membrane per second is estimated to be

  $$\frac{dN_{\text{ion}}}{dt} = J_{\text{ion}}A_{\text{channel}} \approx 10^4\text{nm}^{-2}\text{s}^{-1} \times 0.2\text{nm}^2 = 4 \times 10^6\text{s}^{-1}, \tag{3.15}$$

  or alternatively, we can say that a single ion makes it across in roughly $1/N_{\text{ion}} = 1/2$ a millisecond. This estimate does remarkably well at giving a sense of the time scales associated with ion transport across ion channels.

  RP: not satisfied with this estimate. Can use the known conductances in pS and then see what number crossing is and it is a lot higher.

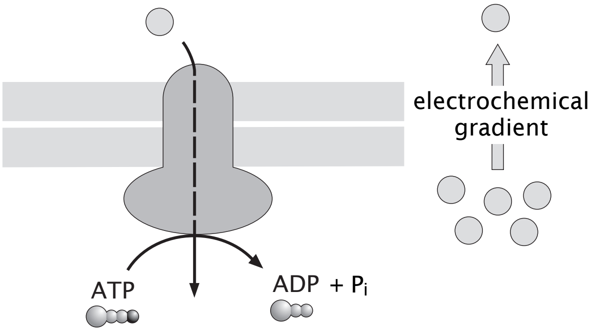


Figure 3.26: Active transport across membranes. Molecular pumps consume energy in the form of ATP hydrolysis and use the liberated free energy to pump molecules against their concentration gradient.

Enzymes, molecular motors and ion channels (and pumps) are all ways in which the cell uses proteins to circumvent the intrinsic rates of different physical or chemical processes.

3.4.3 Beating the Replication Limit

The most fundamental process of cellular life is to form two new cells. A minimal requirement for this to take place is that an individual cell must duplicate its genetic information. Replication of the genetic material proceeds through the action of DNA polymerase, an enyzme which copies the DNA sequence information from one DNA strand into a complementary strand. Like all biochemical reactions, this requires a certain amount of time which we estimated earlier in the chapter. Why should any cell accept this speed limit on its primary directive of replication? When we calculated the replication time for the *E. coli* chromosome, we concluded that two replication forks operating at top speed would be sufficient to replicate the chromosome in approximately the 3000 second division time that we stipulated for a bacterium growing in a minimally defined medium with glucose as the sole carbon source under a continuous supply of oxygen. While this set of conditions is in many ways convenient for the human experimentalist, it is by no means ideal for the bacterium. If instead of only supplying glucose we add a rich soup of amino acids, *E. coli* will grow much faster with a doubling time of order twenty minutes (1200 seconds).

How can the bacterium double more quickly than its chromosome can replicate? For *E. coli* and other fast growing bacteria, the answer is a simple and elegant manipulation of the procedural time limit imposed by the DNA replication apparatus. The bacterium begins replicating its chromosome a second time before the first replication is complete. In a rapidly growing *E. coli* there may be between four and eight copies of the chromosomal DNA close to the replication origin, even though there may be only one copy of the chromosome close to the replication terminus. In other words, the bacterium has started replicating its daughter's, grandaughter's or even great grandaughter's chromosome before its own replication is complete. The newborn *E. coli* cell is thus essentially already pregnant with partially replicated chromosomes preparing for the next one or two generations.

As we have noted above, the genome size for bacteria tends to be substantially smaller than the genome size for eukaryotic cells. Nonetheless, eukaryotic cells are still capable of replicating at a remarkably fast rate. For example, early embryonic cells of the South African clawed frog *Xenopus laevis* can divide every 30 minutes. Despite the fact that its genome (3100MB) is roughly six-fold larger than the genome of *E. coli*, two mechanical changes enable rapid replication of the *Xenopus* genome (and the genome of other eukaryotes). First, the genome is subdivided into multiple linear chromosomes rather than a single circular chromosome. Second, and more importantly, replication is initiated simultaneously from many different origins sprinkled throughout the chromosome as opposed to the single origin of bacterial chromosomes. This parallel processing for the copying of genomic information enables the task to be completed more rapidly than would be dictated by the procedural time limit dictated by a single molecule of DNA polymerase.

3.4.4 Eggs and Spores: Planning for the Next Generation

We have been considering the processes of cell growth and division as though they are tightly coupled with one another. In some cases, however, organisms may separate the processes of growth and division so that they occur over different spans of time. The most dramatic example of this is in the growth of giant egg cells followed by extremely rapid division of early embryos after fertilization. For example, in the frog *Xenopus laevis*, each individual egg cell is enormous - up to 1mm across - and grows gradually within the body of the female over a period of three months. Following fertilization, cell division occurs without growth so that a tadpole hatches after 36 hours that has the same mass as the egg from which it is derived.

Even for organisms where cell growth normally is coupled to cell division, there are several mechanisms whereby cells may choose to postpone either growth or division if conditions are unfavorable. For example, many cells ranging from bacteria through fungi to protozoans such as *Dictyostelium* are capable of creating spores. Spores are nearly metabolically inert and serve as a storage form for the genomic information of the species that can survive periods of drought or low nutrient availability. This is a mechanism by which an organism can effectively exist in suspended animation waiting for however much time is needed to pass until conditions become favorable again. When fortune finally favors the spore, it can germinate releasing a rapidly growing cell. The maximum survival time of spores is unknown. However, there are *Bacillus* spores that were put into storage by Louis Pasteur in the late 1800s that appear to be fully viable today and it is generally accepted that some spores may be able to survive for thousands of years. Some controversial reports even suggest that viable bacterial spores can be recovered from bodies trapped in amber over at least a few million years. The seeds of flowering plants perform a similar role, though they are typically not as hearty as spores.

Although animals do not form true spores, several do have forms that permit long term survival under starvation conditions. The most familiar examples are hibernation of large animals such as bears which can survive an entire winter season without eating. Smaller animals perform similar tricks. These include dauer form larvae of several worms. The most impressive example of "suspended animation" among animals is presented by the tardigrade or water bear, a particularly adorable segmented metazoan that rarely grows larger than 1mm. The tardigrade normally lives in water, but when it is dried out it, it slows its metabolism and alters its body shape, extruding almost all of its water, to form a dried out form called a tun. Tardigrade tuns can be scattered by wind and can survive extreme highs and lows of temperature and pressure. When the tun falls into a favorable environment like a pond, the animal will reanimate. Each of these examples shows how organisms have evolved mechanisms that are completely indifferent to the absolute passage of time.

3.5 Summary and Conclusions

Because life processes are associated with constant change, it is important to understand how long these processes take. In the case of the diurnal clock, organisms are able to measure the passage of time with great regularity to determine their daily behaviors. For most other kinds of biological processes, times are not absolute. In this chapter, we explored several different views of time in biological systems starting with the straightforward assignment of procedural time as measured by the amount of time it takes to complete some process. In complex biological systems, processes that occur at intrinsically different rates may be linked together such that one must be completed before another can begin. Examples of this kind of measurement of relative time are found in the regulation of the cell cycle, the assembly of complex structures such as the bacterial flagellum, etc. Finally, we briefly explored some of the ways that organisms manipulate biological processes to proceed faster or slower than the normal intrinsic rates. Armed with these varying views of time, we will use time as a dimension in our estimates and modeling throughout the remainder of the book. Time will particularly take center stage in Part III, Life in Motion, where dynamic processes will be revealed in all their glory.

3.6 Further Reading

Bier E., **The Coiled Spring: How Life Begins**, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000. This is the best introduction to developmental biology that we are aware of. This book is a wonderful example of the seductive powers of development.

Carroll S. B., **Endless Forms Most Beautiful**, W.W. Norton and Company, New York, New York, 2005. One of us (RP) read this book twice in the first few weeks after it hit the shelves. From the perspective of the present chapter, this book illustrates the connection between developmental and evolutionary time scales.

F. Neidhardt, "Bacterial Growth: Constant Obsession with dN/dt", J. Bacteriol., **181**, 7405 (1999). Bacterial growth curves are one of the simplest and most enlightening tools for peering into the inner workings of cells. Neidhardt's ode to growth curves is both entertaining and educational.

D. Dressler and H. Potter, **Discovering Enzymes**, W. H. Freeman and Company, New York, New York, 1991. This book is full of fascinating insights into enzymes.

John Gerhart and Marc Kirschner, **Cells, Embryos, And Evolution**, Blackwell Science, 1997 and **The Plausibiity of Life**, Yale University Press, 2006.

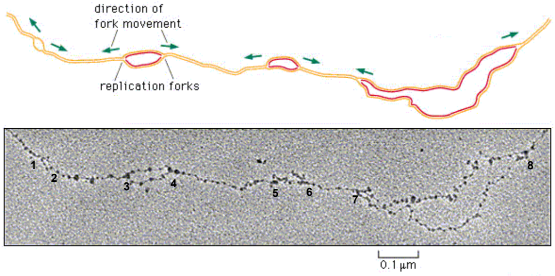


Figure 3.27: Replication forks in Drosophila (from ECB).

3.7 Problems

1. Numerics of the cell cycle

2. ***E. coli* cell cycle** Improve the estimates for the synthesis rates *E. coli* during a cell cycle from section 3.1.3 by including the effect of degradation.

3. DNA replication rates Look at fig. 3.27 and assuming that this is a representative sample of the replication process, estimate the number of DNA polymerase molecules in a eukaryotic cell like this one from the fly. Note that the fly DNA is about 1.8×10^8 nucleotide pairs in size. Estimate the fraction of the total fly DNA shown in the micrograph. There are eight forks in the micrograph, numbered 1-8. Estimate the lengths of the DNA strands between replication forks 4 and 5 where we count the forks from left to right. If a replication fork moves at a speed of 100 nucleotides/s, how long will it take for forks 4 and 5 to collide. Also, given the mean spacing of the bubbles, estimate how long it will take to replicate the entire fly genome.

4. Metabolic rates Bacterial cells have a much higher rate of metabolism than animal cells. For example, most bacterial cells under optimal conditions divide on a $20 \sim 30$ minutes time scale, while animal cells can take a day or more. To gain some understanding of the large difference in metabolic rates consider the fact that for energy production to occur in an aerobic cell, oxygen must be transported through the cell membrane and distributed across the cellular interior.

(a) Argue that the maximum metabolic rate is larger for cells with a larger surface to volume ratio.
(b) Compare the surface to volume ratio of *E. coli* to that of a globular amoeba which can be thought of as a sphere of radius 150μm.
(c) In this problem we examine the limitations imposed on the size of a bacterium by metabolism. Lets take that the bacterium burns oxygen at a rate of 0.02 mole/kg$s$; this is the amount of oxygen spent per unit time per unit mass of the bacterium, which we assume is a sphere of radius $R = 1\mu m$. This oxygen gets into the bacterium by diffusion through its surface at a rate given by $\Phi = 4\pi DRc_0$. $D = 2\mu m^2/ms$ is the diffusion constant for oxygen in water, and $c_0 = 0.2$mole$/m^3$ is the oxygen concentration.

Chapter 8

Random Walks and the Structure of Macromolecules

"The journey of a thousand miles begins with a single step." - Chinese proverb

Chapter Overview: In Which We Think of Macromolecules as Random Walks

A useful alternative to the deterministic description of structure in terms of well defined atomic coordinates is the use of statistical descriptions of structure. For example, the arrangement of a large DNA molecule within the cell is often best characterized statistically in terms of average quantities such as the mean size and position. The goal of this chapter is to examine one of the most powerful ideas in all of science, namely, the random walk, and to show its utility in characterizing biological macromolecules such as DNA. We will show how these ideas culminate in a probability distribution for the end-to-end distance of polymers and how this distribution can be used to compute the "structure" of DNA in cells as well as to understand recent single-molecule experiments in which molecules of DNA (or proteins) are pulled on and the subsequent deformation is monitored as a function of the applied force. In addition, we will show how these same ideas may be tailored to thinking about proteins.

8.1 What is a Structure: PDB or R_G?

The study of structure is often a prerequisite to tackling the more interesting question of the functional dynamics of a particular macromolecule or macromolecular assembly. Indeed, this notion of the relation between structure and function has been elevated to the status of the true central dogma of molecular biology, namely, "sequence determines structure determines function" (Petsko and Ringe, 2004), which calls for uncovering the relation between sequence and consequence. The idea of structure is hierarchical and subtle, with the relevant detail that is needed to uncover function often living at totally disparate spatial scales. For example, in thinking about phosphorylation-induced conformational changes, an atom-by-atom description is required, whereas in thinking about cell division, a much coarser description of DNA is likely more useful. The key message of the present chapter is that there is much to be gained in some circumstances by abandoning the deterministic, pdb mentality described in earlier chapters for a *statistical* description in which we attempt only to characterize certain average properties of the structure. We will argue that this type of thinking permits immediate and potent contact with a range of experiments.

8.1.1 Deterministic vs Statistical Descriptions of Structure

PDB Files Reflect a Deterministic Description of Macromolecular Structure

The notion of structure is complex and ambiguous. In the context of crystals, we can think of structure at the level of the monotonous regular packing of the atoms into the unit cells of which the crystal is built. This thinking applies even to crystals of nucleic acids, proteins or complexes such as ribosomes, viruses and RNA polymerase. Indeed, it is precisely this regularity that makes it possible to deposit huge pdb files containing atomic-coordinates on databases such as the Protein Data Bank and VIPER. In this world view, a structure is the set $(\mathbf{r}_1, \mathbf{r}_2, \cdots \mathbf{r}_N)$, where $\mathbf{r}_i$ is the vector postion $\mathbf{r}_i = (x_i, y_i, z_i)$ of the i^{th} atom in this N-atom molecule. However, the structural descriptions that emerge from x-ray crystallography provide a deceptively static picture which can only be viewed as a starting point for thinking about the functional dynamics of macromolecules and their complexes in the crowded innards of a cell.

Statistical Descriptions of Structure Emphasize Average Size and Shape Rather Than Atomic Coordinates

As noted above, in the context of polymeric systems, the notion of structure is subtle and brings us immediately to the question of the relative importance of universality (for example, how size scales with the number of monomers) and specificity in macromolecules. In particular, there are certain things that we might wish to say about the structure of polymeric systems that are indifferent to the precise chemical details of these systems. For example, when a DNA molecule is ejected from a bacteriophage into a bacterial cell, all that we may really care to say about the disposition of that molecule is how much space it takes up and where within the cell it does so. Similarly, in describing the geometric character of a bacterial genome, it may suffice to provide a description of structure only at the level of characterizing a blob of a given size and shape. Indeed, these considerations bring us immediately to the examination of

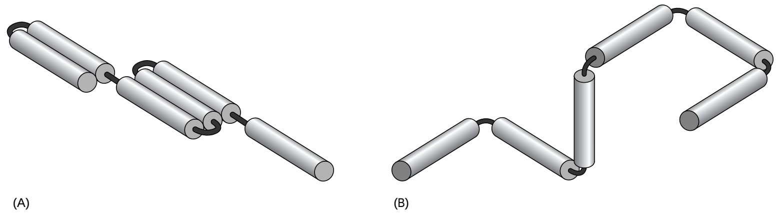


Figure 8.1: Random walk model of polymer. Schematic representation of a (A) one-dimensional random walk and a (B) three-dimensional random walk as an arrangement of linked segments of length a.

statistical measures of structure. As hinted at in the title to this section, one such statistical measure of structure is provided by the radius of gyration, R_G, which, roughly speaking, gives a measure of the size of a polymer blob. It is the business of the remainder of the chapter to show the calculable consequences of adopting such a description of structure.

8.2 Macromolecules as Random Walks

Random Walk Models of Macromolecules View Them as Rigid Segments Connected by Hinges

One way to characterize the geometric disposition of a macromolecule such as DNA is through the *deterministic* function $\mathbf{r}(s)$. This function tells us the position ($\mathbf{r}$) of that part of the polymer which is a distance s along its contour. An alternative we will explore here is to discretize the polymer into a series of segments, each of length a, and to treat each such segment as though it is rigid. The various segments that make up the macromolecular chain are then imagined to be connected by flexible links that permit the adjacent segments to point in various directions. Both one- and three-dimensional versions of this idea are shown in fig. 8.1. Note that in the figure, we illustrate the case in which the links are restricted to 90 degree angles, though there are many instances in which we will consider links that can rotate in arbitrary directions (the so-called freely jointed chain model).

Fig. 8.2 shows an example of the correspondence between the real structures of these molecules and their idealization in terms of the lattice model of the random walk. In particular, fig. 8.2 shows a conformation of DNA on a surface. Using the discretization advocated above, we show how this same structure can be approximated using a series of rigid rods (the Kuhn segments) connected by flexible hinges. We will argue that this level of description can be useful in settings ranging from estimating the entropic cost to confine DNA to the response of DNA when subjected to mechanical forces.

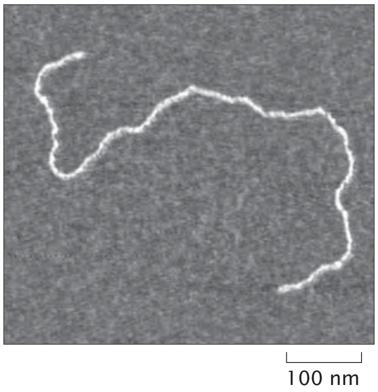


Figure 8.2: Structure of DNA on a surface as seen experimentally using atomic-force microscopy.

8.2.1 A Mathematical Stupor

Every Macromolecular Configuration Is Equally Probable When the Polymer Is Viewed as a Random Walk

In this section we work our way up by degrees to some of the full beauty and depth of the random walk model. The aim of the analysis is to obtain a probability distribution for each and every macromolecular configuration and to use these probabilities to compute properties of the macromolecule that can be observed experimentally, such as the mean size of the macromolecule and the free energy required to deform that molecule. Our starting point will be an analysis of the random walk in one-dimension, with our discussion being guided by the ways in which we will later generalize these ideas and apply them in what might at first be considered unexpected settings.

We begin by imagining a single random walker confined to a one-dimensional lattice with lattice parameter a as already shown in fig. 8.1(A). The life history of this walker is built up as a sequence of left and right steps, with each step constituting a single segment in the polymer. In addition, for now we postulate that the probabilities of left and right steps are given as $p_r = p_l = 1/2$. The trajectory of the walker is built up by assuming that at each step the walker starts anew with no concern for the orientation of the previous segment. We note that for a chain with N segments, this implies that there are a total of 2^N different permissible macromolecular configurations, each with probability $1/2^N$.

The Mean Size of a Random Walk Macromolecule Scales as the Square Root of the Number of Segments, $\sqrt{N}$

Given the spectrum of possible configurations and their corresponding probabilities, one of the most immediate questions we can pose concerns the mean distance of the walker from its point of departure as a function of the number of segments in the chain. In the context of biology, this question is tied to problems such as the cyclization of DNA, the likelihood that a tethered ligand and receptor will find each other and to the gross structure of plasmids and chromosomal DNA in cells. To find the end-to-end distance for the molecule of interest we can use both simple arguments as well as brute force calculation, and we will take up both of these options in turn. The simple argument notes that the expected value of the walkers distance from the origin, R, after N steps can be obtained as

$$\langle R \rangle = \langle \sum_{i=1}^{N} x_i \rangle, \tag{8.1}$$

where $x_i = \pm a$ is the excursion suffered by the walker during the i^{th} step and where we have introduced the bracket notation $\langle \cdots \rangle$ to signify an average. Recall that to obtain such an average we sum over all possible configurations with each configuration weighted by its probability (in this case they are all equal). This result may be simplified by noting that the averaging operation represented by the brackets $\langle \cdots \rangle$ on the righthand side of the equation can be passed within the summation symbol (i.e. the average of a sum is the sum of the averages) and through the recognition that $\langle x_i \rangle = 0$. Indeed, this leaves us with the conclusion that the mean excursion undertaken by the walker is identically zero.

A more useful measure of the walker's departure from the origin is to examine

$$\langle R^2 \rangle = \langle \sum_{i=1}^{N} \sum_{j=1}^{N} x_i x_j \rangle \ . \tag{8.2}$$

This is the variance of the probability distribution of R, while $\sqrt{\langle R^2 \rangle}$ is the standard deviation. Its significance is that the probability of finding our random walker within one standard deviation of the mean is close to 70%. In other words the standard deviation is the measure of the typical excursion of the random walker after N steps, and therefore serves as a good surrogate for the typical size of the related polymer.

In order to make progress on eqn. 8.2 we break up the sum into two parts as

$$\langle R^2 \rangle = \sum_{i=1}^{N} \langle x_i^2 \rangle + \sum_{i \neq j=1}^{N} \langle x_i x_j \rangle. \tag{8.3}$$

Note that each and every step is independent of all steps that precede and follow it. This implies that the second term on the righthand side is zero. In addition, we note that $\langle x_i^2 \rangle = a^2$, with the result that

$$\langle R^2 \rangle = N a^2. \tag{8.4}$$

Thus, we have learned that the walker's departure from the origin is characterized statistically by the assertion that $\sqrt{\langle R^2 \rangle} = a\sqrt{N}$, meaning that the distance from the origin grows as the square root of the number of segments in the chain.

The Probablity of a Given Macromolecular Configuration Depends Upon it's Microscopic Degeneracy

In addition to the simple argument spelled out above, it is also possible to carry out a brute force analysis of this problem using the conventional machinery of probability theory. We consider this an important alternative to the analysis given above since it highlights the fact that there are many microscopic configurations that correspond to a given macroscopic configuration. In particular, in the case in which the walker makes a total of N steps, we pose the question, what is the probability that n_r of those steps will be to the right (and hence $n_l = N - n_r$ to the left)? Since the probability of each right or left step is given by $p_r = p_l = 1/2$, the probability of a *particular* sequence of N left and right steps is given by $(1/2)^N$. On the other hand, we must remember that there are many ways of realizing n_r right steps and n_l left steps out of a total of N steps. In particular, there are

$$W(n_r; N) = \frac{N!}{n_r!(N - n_r)!}, \tag{8.5}$$

distinct ways of achieving this outcome. A particular example of this thinking to the case $N = 3$ is shown in fig. 8.3 where we see that there is one configuration where all three segments are right pointing, one configuration in which all three segments are left pointing and three configurations each for the cases in which $n_r = 2, n_l = 1$ and $n_r = 1, n_l = 2$.

We have now enumerated the microscopic degeneracies of each macroscopic configuration (characterized by a given end-to-end distance). As a result, we are poised to write down the probability of an overall departure n_r from the origin which is given by

$$p(n_r; N) = \frac{N!}{n_r!(N - n_r)!}(\frac{1}{2})^N. \tag{8.6}$$

With this probability distribution in hand, we can now evaluate any average characterizing the geometric disposition of the chain by summing over all of the configurations.

To develop facility in the use of this probability distribution, we begin by confirming that it is normalized. To do so, we ask for the outcome of the sum

$$\sum_{n_r=0}^{N} p(n_r; N) = \sum_{n_r=0}^{N} \frac{N!}{n_r!(N - n_r)!}(\frac{1}{2})^N. \tag{8.7}$$

To evaluate this sum, we recall the binomial theorem that tells us

$$(x + y)^N = \sum_{n_r=0}^{N} \frac{N!}{n_r!(N - n_r)!} x^{n_r} y^{N-n_r}. \tag{8.8}$$

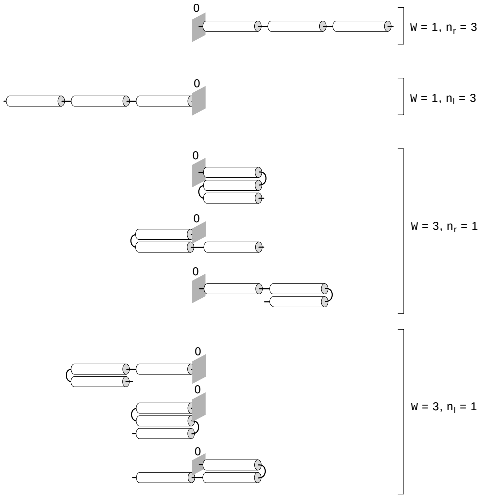


Figure 8.3: Random walk configurations. The schematic shows all of the allowed conformations of a polymer made up of three segments ($2^3 = 8$ conformations) and their corresponding degeneracies.

For the case in which $x = y = 1$, we see that this implies

$$\sum_{n_r=0}^{N} \frac{N!}{n_r!(N-n_r)!} = 2^N. \tag{8.9}$$

Plugging this result back into eqn. 8.7 demonstrates that the probability distribution is indeed normalized (i.e. $\sum_{n_r=0}^{N} p(n_r; N) = 1$).

Entropy Determines the Elastic Properties of Polymer Chains

The probability distribution for n_r can be used to deduce a more telling quantity, the probability distribution for the end to end distance, $R = (n_r - n_l)a$. If we use the condition $n_r + n_l = N$ to solve for n_l and substitute this into $R = (n_r - n_l)a$, it follows that $n_r = (N + R/a)/2$ and eqn. 8.6 can be rewritten as

$$p(R; N) = \frac{N!}{\left(\frac{N}{2} + \frac{R}{2a}\right)!\left(\frac{N}{2} - \frac{R}{2a}\right)!}\left(\frac{1}{2}\right)^N, \tag{8.10}$$

to give the probability distribution of the end-to-end distance. This distribution is plotted in fig. 8.4. For large N this probability distribution is sharply peaked at $R = 0$. Next we show that it takes on the form of a Gaussian distribution for $R \ll Na$. This calculation involves two math methods we have discussed previously, the Stirling approximation (pg. 256), $\ln n! \approx n \ln n - n + \frac{1}{2}\ln(2\pi n)$ for $n \gg 1$, and the Taylor expansion (pg. 249), $\ln(1 + x) \approx x - x^2/2$ for $x \ll 1$. Note that here we take the first three terms in the Stirling approximation, and

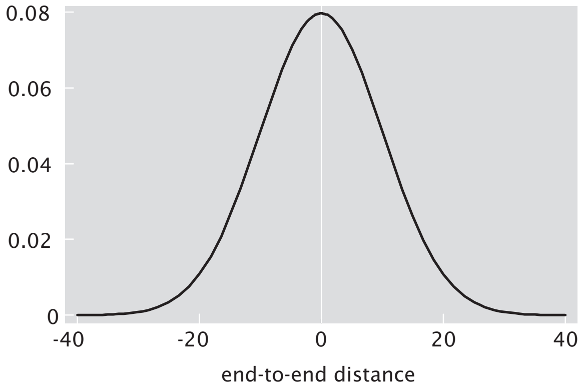


Figure 8.4: End-to-end probability distribution for a one-dimensional "macromolecule" with 100 segments. RP: Fix the figure so that it shows a comparison of the Binomial distribution and the approximate Gaussian for different values of N.

keep terms up to x^2 in the Taylor expansion, in anticipation that the leading term of $\ln p(R;N)$ is of order R^2.

We begin by taking the logarithm of the probability distribution for R shown in eqn. 8.10 and then we apply the Strirling approximation to each of the three factorials resulting in,

$$\begin{aligned}\ln p(R;N) &= \underbrace{N\ln N - N + \frac{1}{2}\ln(2\pi N)}_{\ln N!} \\ &- \underbrace{\left[\left(\frac{N}{2}+\frac{R}{2a}\right)\ln\left(\frac{N}{2}+\frac{R}{2a}\right) - \left(\frac{N}{2}+\frac{R}{2a}\right) + \frac{1}{2}\ln\left(2\pi\left(\frac{N}{2}+\frac{R}{2a}\right)\right)\right]}_{\ln(N/2+R/2a)!} \\ &- \underbrace{\left[\left(\frac{N}{2}-\frac{R}{2a}\right)\ln\left(\frac{N}{2}-\frac{R}{2a}\right) - \left(\frac{N}{2}-\frac{R}{2a}\right) + \frac{1}{2}\ln\left(2\pi\left(\frac{N}{2}-\frac{R}{2a}\right)\right)\right]}_{\ln(N/2-R/2a)!} \\ &- N\ln 2\ . \end{aligned} \tag{8.11}$$

In the next step we rewrite the logarithms,

$$\ln\left(\frac{N}{2}\pm\frac{R}{2a}\right) = \ln\left[\frac{N}{2}\left(1\pm\frac{R}{Na}\right)\right] = \ln\frac{N}{2} + \ln\left(1\pm\frac{R}{Na}\right) \tag{8.12}$$

where we have used the rule about logarithms that $\ln[AB] = \ln(A)+\ln(B)$. We can now make use of the Taylor expansion,

$$\ln\left(1\pm\frac{R}{Na}\right) \approx \pm\frac{R}{Na} - \frac{1}{2}\left(\pm\frac{R}{Na}\right)^2 \tag{8.13}$$

which we substitute repeatedly in eqn. 8.11. After some annoying algebra (which is left as an exercise for the reader) we arrive at the formula

$$\ln p(R;N) = \ln 2 - \frac{1}{2}\ln(2\pi N) - \frac{R^2}{2Na^2}. \tag{8.14}$$

If we now exponentiate both sides of this equation, we find the coveted Gaussian distribution,

$$p(R;N) = \frac{2}{\sqrt{2\pi N}} e^{-\frac{R^2}{2Na^2}} \ . \tag{8.15}$$

Note that the derived approximate formula is a probability for values of R which come in multiples of $2a$. To turn this into a probability distribution function, $P(R;N)$, such that $P(R;N)dR$ is the probability that R falls within an interval of length dR, all that remains is to divide out the result in eqn. 8.15 by the density of integer R values per unit length, which is $1/2a$. This yields the result for the probability distribution function for the end to end distance of a freely jointed chain,

$$P(R;N) = \frac{1}{\sqrt{2\pi N a^2}} e^{-\frac{R^2}{2Na^2}} \ , \tag{8.16}$$

which we will make use of repeatedly throughout the book.

The result derived above is a special case of the so-called central-limit theorem which is arguably the most important result of probability theory. In a nutshell, it states that the probability distribution of $x_1 + x_2 + \cdots + x_N$, which is a sum of identical, independently distributed random variables, is Gaussian in the limit of large N, as long as the mean and variance of each individual x_i is finite. Since the individual displacements of the random walker satisfy this condition, it immediately follows that for large number of steps N, the total displacement R will be Gaussian distributed, with mean $\langle \mathbf{R} \rangle = 0$ and variance $\langle \mathbf{R}^2 \rangle = Na^2$. Note that this will hold regardless of whether the walk is executed in 1, 2 or 3 dimensions.

We leave it as a homework problem to show that the Gaussian distribution of R for a 1-dimensional walk given in eqn. 8.16 indeed has the required mean and variance. Here we make use of this result to derive the large-N distribution for the end-to-end distance of a 3-dimensional random walk. Since the mean is zero the distribution is of the form

$$P(\mathbf{R};N) = \mathcal{N} e^{-\kappa R^2} \tag{8.17}$$

where the parameters $\mathcal{N}$ and κ are to be determined from two conditions that the distribution must satisfy

$$\begin{aligned} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} P(\mathbf{R},N)d^3R &= 1 \text{ (Normalization)} \\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} R^2 P(\mathbf{R},N)d^3R &= Na^2 \text{ (Variance)} \ . \end{aligned} \tag{8.18}$$

Since both integrands are functions of R^2 we can transform the volume integral in both cases to an integral over spherical shells of radius R to obtain,

$$\begin{aligned}\int_0^{+\infty} P(\mathbf{R}, N) 4\pi R^2 dR &= 1 \text{ (Normalization)} \\ \int_0^{+\infty} R^2 P(\mathbf{R}, N) 4\pi R^2 dR &= Na^2 \text{ (Variance)} . \end{aligned} \tag{8.19}$$

To compute the integrals in the above equations we make use of the Gaussian integral formulas

$$\begin{aligned}\int_0^{+\infty} 4\pi \mathcal{N} R^2 e^{-\kappa R^2} dR &= 4\pi \mathcal{N} \frac{1}{4}\sqrt{\frac{\pi}{\kappa^3}} = 1 \\ \int_0^{+\infty} 4\pi \mathcal{N} R^4 e^{-\kappa R^2} dR &= 4\pi \mathcal{N} \frac{3}{8}\sqrt{\frac{\pi}{\kappa^5}} = Na^2. \end{aligned} \tag{8.20}$$

To compute κ we can divide the second equation by the first to give

$$\kappa = \frac{3}{2Na^2} \ . \tag{8.21}$$

Substituting this result into the first of the two integrals above gives us

$$\mathcal{N} = \left(\frac{\kappa}{\pi}\right)^{\frac{3}{2}} = \left(\frac{3}{2\pi Na^2}\right)^{\frac{3}{2}} \ , \tag{8.22}$$

the normalization constant. Putting this all together we obtain the end-to-end distribution for a 3-dimensional random walk with N Kuhn segments of length a,

$$P(\mathbf{R}; N) = \left(\frac{3}{2\pi Na^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Na^2}} \ . \tag{8.23}$$

- **Estimate: End-End Probability for the *E. coli* genome.** One interesting application of these ideas that will be explored more throughout the chapter is to the structure of chromosomal DNA. The DNA associated with an *E. coli* cell is roughly 5 million nucleotides long, and can be modeled as a random walk of roughly $N = 15000$ steps since the Kuhn length for bare DNA is roughly 300 bp in length. The probability that the end-to-end distance is zero for a one-dimensional walk of this many steps is 7×10^{-3}. The probability that $R = 500a$ is 2×10^{-6} while for $R = 1000a$ the probability drops all the way down to 2×10^{-17}. This overwhelming probability that R is close to zero is responsible for the elastic properties of polymer chains. Namely, if you imagine stretching a polymer (say, the *E. coli* DNA) so that R is non-zero, then upon release it will quickly find itself in the $R \approx 0$ state solely by virtue of this being a much more likely state. Note that this is not the result of any real physical force, such as, for example, the electric force which is ultimately responsible for the elastic properties of crystals, but purely a result of statistics. As such it is, like the case of pressure of the ideal gas, another example of an entropic force.

The Persistence Length Is a Measure of the Length Scale Over Which a Polymer Remains Roughly Straight

With the random walk model in hand we can describe the structure of long polymers, whose contour length L is much larger than the persistence length ξ_p, which is the length over which the polymer is essentially straight. In particular, the persistence length is the scale over which the tangent-tangent correlation function decays along the chain. To see this idea more clearly, we imagine a polymer as a curve in three dimensional space. At each point along that curve, we can draw a tangent vector which points along the polymer at that point. As a result of thermal fluctuations, the polymer meanders in space and the persistence length is the length scale over which "memory" of the tangent vector is lost. From a mathematical perspective, we can write the tangent-tangent correlation function as $\langle \mathbf{t}(s) \cdot \mathbf{t}(s') \rangle$, where $\mathbf{t}(s)$ is the tangent vector evaluated at the point a distance s along the polymer and the notation $\langle \cdots \rangle$ is an instruction to average over all the configurations. The persistence length determines the scale over which correlations in tangent vectors decay through the equation

$$\langle \mathbf{t}(s) \cdot \mathbf{t}(s') \rangle = e^{-\frac{|s-s'|}{\xi_p}} \ . \tag{8.24}$$

A good example of a long flexible polymer is provided by genomic DNA of viruses such as λ-phage with a contour length of 16.6μm. This should be compared to the persistence length $\xi_p \approx 50$nm of DNA at room temperature and solvent conditions typical of the cellular environment. Since the persistence length is the length over which the tangent vectors to the polymer backbone become uncorrelated, we can think of the polymer as consisting of $N \sim L/\xi_p$ connected links which take random orientations with respect to each other. This is the logic which gives rise to the *freely jointed chain* model (essentially the random walk picture undertaken in the previous section).

As already described, in the freely-jointed-chain model, polymer conformations are random walks of N steps. The length of the step is the *Kuhn length* which is roughly equal to the persistence length. As promised in the earlier discussion, we now establish the relation between the persistence length and the Kuhn length invoked in the random walk model. To make a more precise determination of the Kuhn length we calculate the mean-squared end-to-end distance of an elastic beam undergoing thermal fluctuations, and compare it to the same quantity obtained for the freely jointed chain. The end-to-end vector $\mathbf{R}$ of a beam can be expressed in terms of the tangent vector $\mathbf{t}(s)$,

$$\mathbf{R} = \int_0^L ds \mathbf{t}(s) \tag{8.25}$$

Therefore

$$\langle \mathbf{R}^2 \rangle = \langle \int_0^L ds \mathbf{t}(s) \int_0^L du \mathbf{t}(u) \rangle \tag{8.26}$$

where $\langle \cdots \rangle$ is the thermal average. Using the tangent-tangent correlation function, eqn. 8.24, we find

$$\langle \mathbf{R}^2 \rangle = 2 \int_0^L ds \int_s^L du e^{-(u-s)/\xi_p}. \tag{8.27}$$

The above integral is obtained by splitting up the integration over the $L \times L$ box in s-u space to integrals over the two triangles, one with $s < u$ and the other with $s > u$, which give equal contributions (thus the factor of two). In the limit $L \gg \xi_p$ we are considering here, we have

$$\langle \mathbf{R}^2 \rangle \approx 2 \int_0^L ds \int_0^\infty dx e^{-\frac{x}{\xi_p}} = 2L\xi_p \ . \tag{8.28}$$

Comparing this to the result that follows from the random walk model, $\langle \mathbf{R}^2 \rangle = aL$, we see that Kuhn length a is twice the persistence length. We are now prepared to make estimates of the physical size of genomes in solution.

8.2.2 How Big is a Genome?

In previous sections we have demonstrated how the size of a polymer, when viewed as a random walk, can be written in terms of key parameters such as the persistence length ξ_p and the number of Kuhn lengths making up the entire contour. In particular, we deduced the size of the polymer in solution may be written as

$$\sqrt{\langle R^2 \rangle} = 2\xi_p \sqrt{N}. \tag{8.29}$$

This equation may be rewritten in terms of the polymer length once we recall that the number of "monomers" (more correctly, the number of Kuhn lengths) in the chain is given by $N = L/2\xi_p$. In light of this result, we then have

$$\sqrt{\langle R^2 \rangle} = \sqrt{2L\xi_p}. \tag{8.30}$$

The radius of gyration is perhaps a more intuitive measure of the size of a polymer in solution and is defined through the expression

$$\langle R_G^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle (\mathbf{R}_i - \mathbf{R}_{CM})^2 \rangle. \tag{8.31}$$

The center of mass can be defined as

$$\mathbf{R}_{CM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{R}_i. \tag{8.32}$$

With this definition of the radius of gyration in hand, a simple relation between radius of gyration, contour length (L) and persistence length (ξ_p) can be written as (proven by the reader in the problems at the end of the chapter)

$$\sqrt{\langle R_G^2 \rangle} = \sqrt{\frac{L\xi_p}{3}}. \tag{8.33}$$

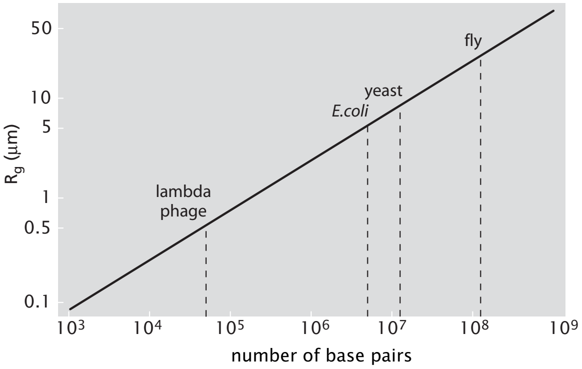


Figure 8.5: Plot of the average size of a DNA molecule in solution as a function of the number of base pairs using the random walk model.

We may write this result in an alternative form in terms of the number of base pairs in the genome of interest by noting that $L \approx .34 N_{bp}$ nm, and hence,

$$\sqrt{\langle R_G^2 \rangle} \approx .3\sqrt{N_{bp}\xi_p} nm. \qquad (8.34)$$

This relation between the radius of gyration of DNA in solution and the number of base pairs is plotted in fig. 8.5.

- **Estimate: The Size of Viral and Bacterial Genomes.** One application of ideas like those described above in the setting of biological electron microscopy is to images of viruses and cells that have ruptured and are thus surrounded by the DNA debris from their genome. We already mentioned in conjunction with fig. 1.12 (pg. 41) that the appearance of DNA in electron microscopy images can be used as the basis of an estimate of genome length. A second example is shown in fig. 8.6 where it is seen that the DNA adopts a configuration in solution which is much larger than the configuration it has when packed inside of the virus or bacterium. To develop intuition for what is seen in such images, we exploit eqn. 8.33 to formulate an estimate of the size of the DNA. Consider fig. 1.12 which shows bacteriophage T2. As seen in the figure, the viral genome has leaked from what is apparently a ruptured capsid and we will assume that this DNA in solution has adopted an equilibrium configuration. The genomes of $T2$ and $T4$ are very similar with a genome length of roughly 150 kB. For a genome of length $L = N_{bp} 3.4 \mathring{A} \approx 510,000 \mathring{A}$ and recalling that the persistence length is $\xi_p \approx 500\mathring{A}$, eqn. 8.33 tells us that the mean size of the DNA seen in fig. 1.12 is $\sqrt{\langle R_G^2 \rangle} = \sqrt{2 \times 500 \times 510 \times 10^3} \mathring{A} \approx 2\mu m$. This result is comparable to though larger than the length scale of the exploded DNA seen in fig. 1.12. Given the crudeness of the model and probably more importantly, the fact that the DNA seems to be constrained via links

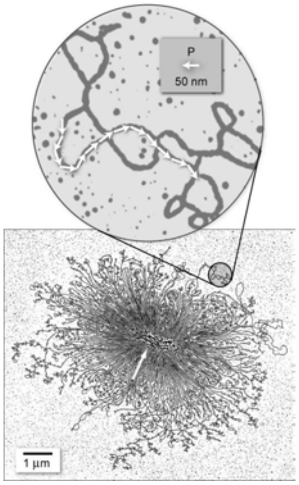


Figure 8.6: Illustration of the spatial extent of a bacterial genome which has escaped the bacterial cell. The expanded region in the figure shows a small segment of the DNA and has a series of arrows on the DNA, each of which have a length equal to the persistence length in order to give a sense of the scale over which the DNA is stiff.

to the capsid itself, this analysis provides a satisfactory first approximation to the structures seen in electron microscopy.

These same arguments can be invoked again to coach our intuition concerning the size of the DNA cloud surrounding a bacterium that has lost its DNA as well. In this case, the genome length is substantially larger than that of the T2 phage, namely, $L \approx 4.6 \times 10^6 \times 3.4\ \AA \approx 1.5 \times 10^7\ \AA \approx 1600\ \mu m$. Once again invoking eqn. 8.33 tells us that the mean size of the DNA seen in fig. 8.6 is $\sqrt{\langle R_G^2 \rangle} \approx 12\ \mu m$. As with the phage calculation, the random walk calculation should be seen as an overestimate since the DNA is clearly forced to return to the bacterium repeatedly, inhibiting the structure from adopting a fully expanded configuration.

8.2.3 The Geography of Chromosomes

Genetic Maps and Physical Maps of Chromosomes Describe Different Aspects of Chromosome Structure.

In our discussion of DNA so far, we have described it as a featureless, self-similar polymer chain. However, of course, DNA is much better known and appreciated as the carrier of genetic information. Classical genetics focused on identification and characterization of genes as abstract entities, ignoring the importance of their physical location on chromosomes and overlooking the consequences of the physical nature of the carrier DNA molecule. The ground breaking work of Thomas Hunt Morgan and his gene hunters which we described in chap. 4 was an early and vivid illustration of the fact that the abstract informational entities known as genes exist with concrete physical relationships to one another. As we have learned more about the regulation and activity of genes, it has become more and more clear that the physical location and dynamic properties of the DNA molecule that carries them are critical components of their biological activity. For example, Morgan's mapping strategy relied on measuring the frequency of recombination between two or more genes. The physical process of recombination requires that two homologous DNA molecules be mobile within a nucleus such that they can physically encounter one another with a measurable frequency. Recombinations do not seem to occur in all nuclei. In the fruit fly, chromosomes are able to recombine in meiosis during oogenesis in the female germline, but not during spermatogenesis in the male germline. Why is it that sometimes DNA segments are able to physically encounter one another and sometimes they are not? What determines the probability of such encounters? These issues in polymer conformations set physical limits on genetic events ranging from transformation and transduction in bacterial cells to the generation of diverse antibodies in the immune system of mammals.

Different Structural Models of Chromatin Are Characterized by the Linear Packing Density of DNA.

One of the themes that we will keep revisiting is the question of DNA packing. In eukaryotic cells DNA is condensed into chromatin fibers. The basic unit of chromatin is the nucleosome. How nucleosomes are packaged into chromatin depends on whether the cell is dividing or not. In the interphase the cell is actively transcribing genes, and the chromosomes are not as condensed as during mitosis when the two copies of the complete genome need to be equally divided among the two daughter cells.

One measure of the degree of DNA packaging into chromosomes is the liner density of chromatin ν, which specifies the number of base pairs of DNA in a nanometer of chromatin fiber. For the 30nm-fiber, shown in fig. 8.7(A), $\nu \approx$ 100bp/nm, while for the 10nm-fiber the packing density is about an order of magnitude smaller. A simple estimate of ν can be made based on the micrograph in fig. 8.7(B) which shows individual nucleosomes along the 10nm-fiber. We see that there are on average 2 nucleosomes for every 50nm of fiber. In yeast cells, for example, there is 200bp per nucleosome (150bp wound around the histones plus 50bp of linker DNA) therefore $\nu \approx 2 \times 200\text{bp}/50\text{nm} = 8\text{bp/nm}$. For comparison, for metaphase chromosomes $\nu \approx 30,000\text{bp/nm}$.

Spatial Organization of Chromosomes Shows Both Elements of Randomness and Order.

Until recently it was believed that interphase chromosomes were randomly distributed within the cell nucleus resembling a bowl of spaghetti. Contrary to this view there is mounting evidence from experiments with fluorescently tagged

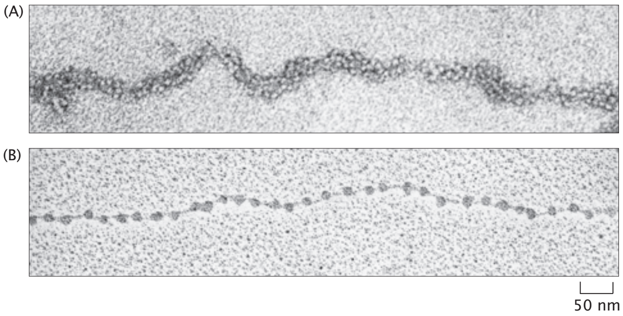


Figure 8.7: Chromatin under the electron microscope. (A) Chromatin extracted from an interphase nucleus appears as a 30nm thick fiber. (B) The 10nm fiber structure shows individual nucleosomes.

chromosomes that the spatial organization of genes in the cell is ordered, as depicted in fig. 8.8. These experiments have put forward the notion of chromosome territories whereby individual chromosomes and particular genetic loci are always found in the same region of the nucleus. The existence of chromosome territories raises a number of questions about how gene expression and pairing interactions of genes (such as during recombination) are orchestrated in space and time.

The observation that interphase chromosomes are segregated would not be surprising if we were dealing with a polymer system which is very dilute. In a dense situation free polymers in solution will interpenetrate each other. Simple estimates can be made for the density of chromatin within the nucleus, and they typically lead to the conclusion that the expected, equilibrium state of chromosomes should be that of a dense polymer system. The fact that segregation is not observed points to the existence of mechanisms beyond polymer chain entropy and confinement, that affect the spatial distribution of chromosomes. We will examine chromosome tethering as one such mechanism. Possible tethering scenarios are shown in fig. 8.9.

- **Estimate: Chromosome Packing in the Yeast Nucleus.** To examine the question of whether the separate chromosomes in yeast are expected to behave as independent blobs or an interpenetrating mess, we pursue the discussion given above in quantitative detail. The yeast cell has 16 chromosomes in its nucleus. The diameter of the interphase nucleus is about 2μm. The chromosome size varies between 230kb to 1500kb, with a total genome size of 12Mb. This gives a density of $c = 12\ \mathrm{Mb}/(4\pi/3 \times 1\mu\mathrm{m}^3) \approx 3\mathrm{Mb}/\mu\mathrm{m}^3$. Lets compare this density with the density of a typical yeast chromosome released from the confines of the cell nucleus. If we adopt the random walk model of a polymer to describe chromatin free in solution, this density can be estimated as $c^* = N_G/(4\pi/3R_{\mathrm{g}}^3)$

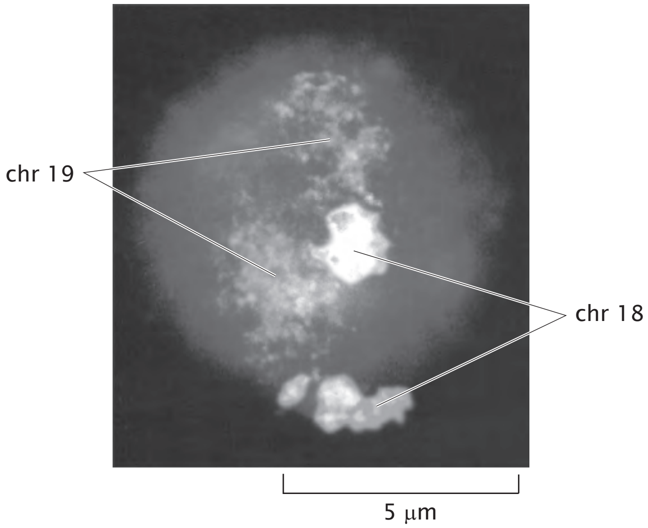


Figure 8.8: Fluorescently stained chromosomes 18 and 19 in a human cell. The chromosomes assume separate territories within the nucleus.

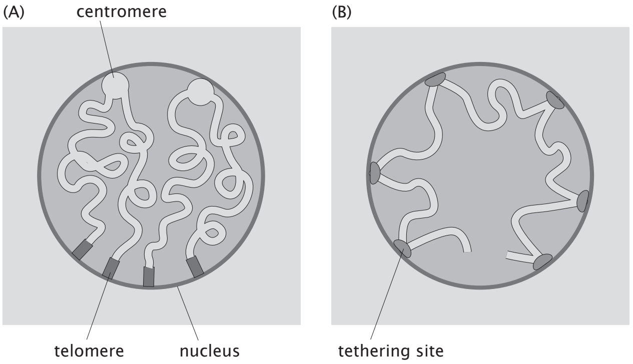


Figure 8.9: Cartoon representation of possible tethering scenarios of interphase chromosomes. The left panel shows tethering at the centromere and the two telomeres at the nuclear periphery. The right panel shows tethering at intermediate locations.

where N_G is the chromosome size in base pairs, and R_g is the radius of gyration of the polymer. If we take an average size of a yeast chromosome to be 12 Mb/16 = 750 kb and a packing density of 8bp/nm the length of this polymer is 750kb/(8bp/nm) = 94 μm. Using the *in vitro* measured value of the persistence length for a 10nm-fiber, $\xi_\mathrm{P} = 30$nm, the estimate for the radius of gyration is, $R_\mathrm{g} = 0.97\mu$m. This then leads to a density for an "free" chromosome of $c^* = 750\mathrm{kb}/(4\pi/3 \times (0.97\ \mu\mathrm{m})^3) \approx 200\ \mathrm{kb}/\mu\mathrm{m}^3$ which is about 10 times smaller than the density of chromosomes in the nucleus. The same qualitative conclusion is reached assuming a 30nm-fiber model for the chromosomes. Using a packing density of 100 bp/nm and the reported persistence length of 200nm an average chromosome has a density of $c^* \approx 500\ \mathrm{kb}/\mu\mathrm{m}^3$. This indicates that the chromosomes in the yeast nucleus should typically be found in an entangled melt-like configuration. The fact that yeast chromosomes are segregated with each chromosome taking up a well defined region of the nucleus indicates the need for a specific mechanism for segregation, such as tethering to the nuclear periphery, as shown in fig. 8.9.

Chromosomes Are Tethered at Different Locations.

One of the recent experimental tricks that has made it possible to examine chromosome geography is the use of repeated DNA binding sites that are the target of particular fluorescently labeled proteins. Conceptually, the experiment can be designed by having two distinct sets of DNA binding sites that are separated by a known *genomic* distance. Then, by measuring the *physical* distance between these binding sites in space as revealed by where the colored spots appear in a fluorescence image, it is possible to map out the spatial distribution of different sites on the genome. Experiments that utilize fluorescence in-situ hybridization, or *lacO* arrays inserted into the chromosomes and labeled with GFP fused Lac repressors, can yield detailed information about the distribution of distances between chromosomal loci. In the absence of tethering a random walk model of chromatin leads to a Gaussian distribution of distances between two tagged loci,

$$P(\mathbf{r}) = \left(\frac{3}{2\pi N a^2}\right)^{3/2} \exp\left(\frac{-3\mathbf{r}^2}{2Na^2}\right) \ , \qquad (8.35)$$

while the presence of a tether at position $\mathbf{R}$ would simply lead to a displaced Gaussian,

$$P(\mathbf{r}) = \left(\frac{3}{2\pi N a^2}\right)^{3/2} \exp\left(\frac{-3(\mathbf{r}-\mathbf{R})^2}{2Na^2}\right) \ . \qquad (8.36)$$

In these formulas $a = 2\xi_p$ is the Kuhn or segment length of the polymer, while N is the total number of segments; Na is the polymer contour length. Using the linear packing density of DNA in chromatin ν, the contour length can be written in terms of the genomic distance as N_G/ν. For example, two genomic loci N_G = 100kb apart would be separated by a 30-nm fiber which is 100kb/100bp/nm =

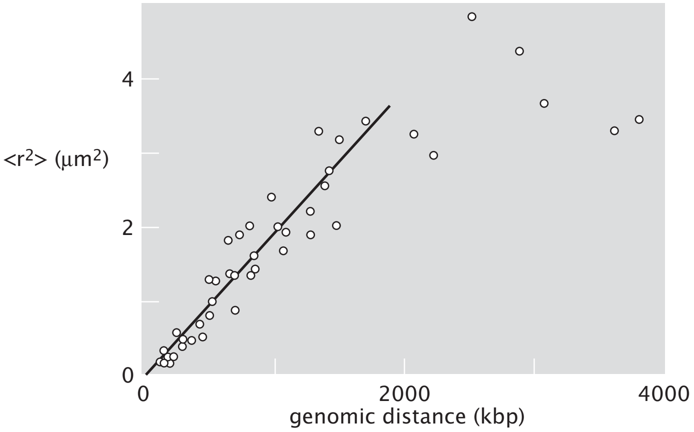


Figure 8.10: Physical distance between two fluorescently labeled loci on human chromosome four as a function of the genomic distance. The physical distance is measured in terms of the average squared distance between the two labels.

1μm in contour length. Assuming that the chromatin structure is that of a 10nm fiber the contour distance along the fiber between the loci would be ten times as large given the ten times smaller packing density.

The end-to-end distribution function for a random walk polymer is determined by a single parameter Na^2, the mean end-to-end distance squared. Since the contour length $Na = N_G/\nu$, the mean end-to-end distance squared can also be written as $\langle R^2 \rangle = N_G a/\nu$. Therefore the material parameter that characterizes the random-walk model of chromosomes is the ratio of the Kuhn length and the packing density. This parameter can be determined from measurements of the average distance squared between two regions of the chromosome as a function of their genomic distance. The results of such a measurement on human chromosome four are shown in fig. 8.10, where the fit to the data yields an estimate of $a/\nu = 2\text{nm}^2/\text{bp}$, which is nothing but the initial slope of the linear portion of the data. The fact that the plot levels off at large genomic distance can be contributed to the effect of chromosome confinement within the cell nucleus. Below we analyze this confining effect using a random walk model in the context of the chromosomes of the bacterium *V. cholerae*.

It is interesting to use the measured value of a/ν to estimate the Kuhn length for the 30-nm and the 10-nm chromatin fiber. Since $\nu_{30-\text{nm}} \approx 100\text{bp/nm}$ and $\nu_{10-\text{nm}} \approx 10\text{bp/nm}$, the corresponding persistence lengths are 100nm and 10nm. Even more interestingly the measured a/ν makes a prediction for the probability distribution of distances between fluorescently tagged loci on the chromosome, which we take up next.

Typically, due to random orientations of cells in the microscope, experiments with tagged chromosomes only yield information about the magnitude r of the

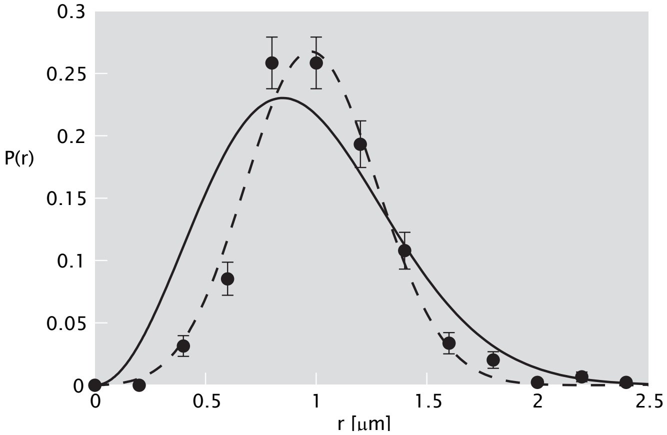


Figure 8.11: Statistics of yeast chromosome III. Distribution of distances between two fluorescent tags placed in proximity of the cetromere and the HML region on yeast chromosome III. These two regions are separated by approximately 100kb in genomic distance.

distance vector **r** between the two marked spots on the chromosome. These can be obtained from eqn. 8.35 and eqn. 8.36 by integrating out the angular variables θ and ϕ associated with the vector **r**. This procedure yields

$$P(r) = \left(\frac{3}{2\pi N a^2}\right)^{3/2} 4\pi r^2 \exp\left(\frac{-3r^2}{2Na^2}\right) \,, \tag{8.37}$$

for the untethered case and

$$P(r) = \left(\frac{3}{4\pi N a^2}\right)^{1/2} \frac{r}{R}\left[\exp\left(\frac{-3(r-R)^2}{2Na^2}\right) - \exp\left(\frac{-3(r+R)^2}{2Na^2}\right)\right]. \tag{8.38}$$

when the polymer is tethered. The parameter characterizing the mechanical properties of the DNA is $Na^2 = N_G a/\nu$. Note that that tethering gives a different functional form for the distribution of distances.

Measurement of the distribution of distances between tagged regions on yeast chromosome III demonstrates that this difference in distributions can be observed *in vivo*. Namely, in fig. 8.11 we show the distance distribution measured between two florescent tags, one placed near the HML region of chromosome III of budding yeast and the other on the spindle pole body, which essentially marks the location of the centromere. The measured distribution is poorly fitted by the free-polymer formula, eqn. 8.37, while the tethered polymer formula, eqn. 8.38 does the job nicely.

The fit to the tethered-polymer distribution yields two quantities that characterize the model, the mean squared distance, $Nb^2 = 0.5\mu\text{m}^2$, and $R \approx 0.9\mu\text{m}$,

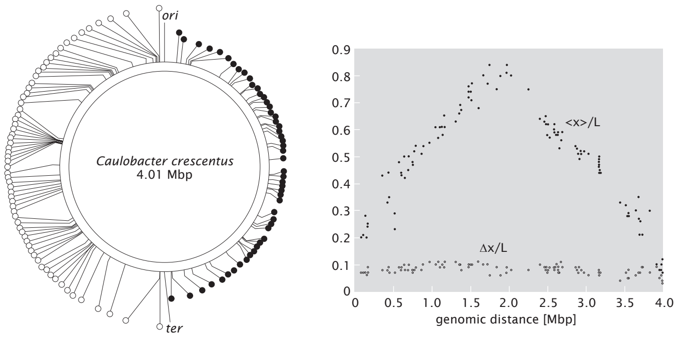


Figure 8.12: Chromosome geography in *Caulobacter crescentus*. Average positions (x/L) and the standard deviation $(\Delta x/L)$ of the position along the long axis of the cell, for 112 different fluorescently tagged locations along the chromosome of *C.crescentus*. The locations of the fluorescent tags are shown on the diagram.

the distance to the tethering point. Note that in order to compute the genomic location of the putative tethering point we need the parameter b/ν which characterizes chromatin structure. For yeast chromosomes measurements of the physical distance as a function of the genomic distance yield $anu \approx 3\text{nm}^2/\text{bp}$ which in turn predicts a genomic distance of $N_G = Na^2/(a/\nu) = 160\text{kb}$. More importantly the tether model makes quantitative predictions for the distance distribution if the marker at HML is moved to a new genomic location.

Chromosome Territories Have Been Observed in Bacterial Cells.

Bacterial chromosomes were until recently thought of as unstructured and random. This view has been seriously challenged by experiments that utilize fluorescent markers placed at different genomic locations, as shown in fig. 8.12. In this experiment 112 different mutants of *C.cresentus* were created with fluorescent tags placed at 112 different locations covering the length of its circular chromosome. Measurements of the average position of the markers along the length of the cell revealed a linear relationship between the genomic distance from the origin of replication and the physical distance away from the pole of the bacterium. This is not too be expected assuming a simple model of the 4Mbp circular chromosome as a polymer loop confined to the cell.

- **Estimate: Chromosome organization in *C. crescentus*.** Another measure of the organization of chromosome in *C.cresentus* is provided by the width of the distribution of positions of the marked regions. As shown in fig. 8.12 the standard deviation of the position is independent of genomic distance from the origin of replication, and is approximately $0.2\mu\text{m}$ (cell length $L \approx 2\mu\text{m}$). We can rationalize this measurement within a simple model where the chromosome is partitioned into loops. This can be affected by proteins that make contact between different locations on the chromosome (H-NS is a possible candidate). To estimate the size of a loop we assume that the observed dispersion of the position is due to the random walk nature of the loop. Since the mean of the square of the three-dimensional end-to-end distance is Na^2 the mean of x^2 is three times less, or $Na^2/3$. Using the relation between genomic distance and the mean distance squared, $Na^2 = N_G a/\nu$, and assuming that the chromosome has the same Kuhn length ($a = 100$nm) and packing density ($\nu = 3$bp/nm) as naked DNA, we arrive at an estimate $(0.2\mu\mathrm{m})^2 = Na^2/3 = N_G/3(100/3)\mathrm{nm}^2/\mathrm{bp}$, $N_G \approx 4$kb, which means that the loop should be 8kb or less. (A more careful analysis would take into account the closed nature of a loop yielding an estimate which is higher by a factor of two.) This correlates nicely with other measurements of topological domains in bacterial chromosomes which find them to be roughly 10kb in size.

Chromosome Territories in *V. cholera* Can Be Explained by Models of Polymer Confinement and Tethering

Another experiment placed a fluorescent markers close to each of the two origins of replication on the two chromosomes of the bacterium *V.cholerae*. This bacterium has two chromosomes, 3Mb and 1Mb in size. In this case the position along the length of the cell (x) and perpendicular to it (y) were both measured. The distribution of x and y are shown in fig. 8.13 for the origin of replication for the larger of the two chromosomes. For comparison, the length of the cell is about 3.2μm, while its diameter is roughly 0.8μm.

The width of the distribution of x positions is roughly half a micron, which is considerably less than the length of the cell. The distribution is centered around $x_0 = 0.6\mu$m, consistent with a tether located at this position in the cell, and is well described by a Gaussian, as expected for a random walk polymer that is unaffected by the presence of cell walls. By fitting the Gaussian distribution for the end-to-end distance of a simple one-dimensional random walk polymer,

$$P(x) = \sqrt{\frac{1}{2\pi Na^2}} e^{-(x-x_0)^2/Na^2} \tag{8.39}$$

we extract the parameter $Na^2 = 0.16\mu\mathrm{m}^2$. Assuming once again the Kuhn length of bare DNA, $a = 0.1\mu$m, we conclude that the number of Kuhn segments between the fluorescent marker and the tethering point at $x_0 = 0.6\mu$m, is $N = 16$. Taking $\nu = 3bp/nm$ this gives a genomic distance of $16 \times 0.1\mu\mathrm{m} \times 3\mathrm{bp/nm} = 4.8$kb to the tether. Therefore the simple one-dimensional model of the chromosome predicts a tether at genomic position roughly 5kb away from the location of the fluorescent marker.

The distribution of positions along the y-direction is spread over the width of the cell and is centered at zero. The latter is a consequence of the experimental procedure whereby distance data was collected from cells whose orientation

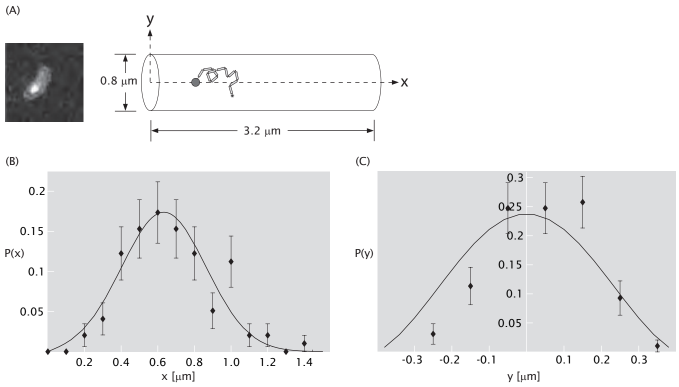


Figure 8.13: Chromosome position distributions *in vivo.* (A) The position of the fluorescently tagged origin of replication on the larger of the two *V.cholerae* chromosomes, is measured along the long axis of the cell (x-direction) and perpendicular to it (y-direction). The cell can be modeled as a cylinder, while the distribution of x and y positions can be explained with a model of a chromosome as a confined and tethered random walk polymer. (B-C) Measured distance distribution functions and comparison to theory.

along the azimuthal direction was random. Furthermore, the distribution is not Gaussian, indicative of confinement by the cell walls.

To develop quantitative intuition about confinement we develop a model of a one-dimensional polymer made up of N segments, each of length a, tethered at position x_0 and confined to a cell of size L; see fig. 8.14. We would like to calculate the distribution of the end-to-end distance $P(x;N)$.

To compute $P(x;N)$ we once again make use of the mapping to the random walk model whereby polymer configurations are identified with trajectories of a random walker that has taken N steps starting at position x_0. As we are only interested in those random walks that stay within the box, we impose absorbing boundary conditions at the boundaries. This guarantees that any walk that crosses the boundary of the box is excluded from the ensemble of allowed walks. The fraction of random walks that start at $x = x_0$ and end up at x without leaving the box is then $G(x;N)$. This quantity satisfies the diffusion equation,

$$\frac{\partial G(x;N)}{\partial N} = \frac{a^2}{2}\,\frac{\partial^2 G(x;N)}{\partial x^2}. \tag{8.40}$$

The probability that a walk which stays in the box also ends up at position x, is then

$$P(x;N) = \frac{G(x;N)}{\int_0^L G(x;N)dx}\ . \tag{8.41}$$

Therefore to obtain the probability distribution $P(x;N)$ we must first solve

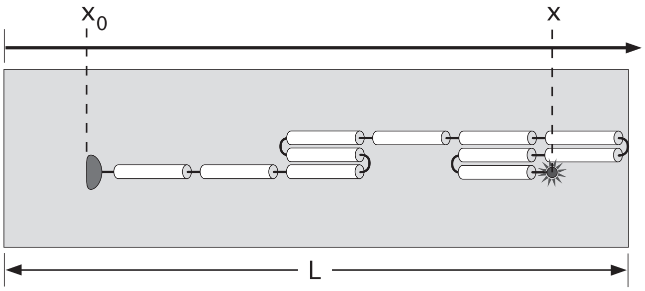


Figure 8.14: Simplified one-dimensional model of a chromosome confined to a cell of size L and tethered at position x_0. The model makes a prediction for the distribution of distances to the fluorescent marker $P(x)$.

eqn. 8.40 with boundary conditions $G(0; N) = G(L; N) = 0$ and the initial condition $G(x; 0) = \delta(x - x_0)$.

To solve eqn. 8.40 we expand the function $G(x; N)$ into a Fourier series,

$$G(x;N) = \sum_{n=1}^{\infty} A_n(N) \sin\left(\frac{n\pi}{L}x\right) ; \tag{8.42}$$

Note that every term in the sum satisfies the absorbing boundary condition. We still need to satisfy the initial condition and the differential equation itself.

The initial condition states

$$\delta(x - x_0) = \sum_{n=1}^{\infty} A_n(0) \sin\left(\frac{n\pi}{L}x\right) \tag{8.43}$$

and it needs to be solved for the constants $A_n(0)$. To do this we multiply both sides with $\sin(m\pi x/L)$ and integrate the equation from 0 to L. The left hand side gives $\sin(m\pi x_0/L)$ while the right hand side is

$$\sum_{n=1}^{\infty} A_n(0) \int_0^L \sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{m\pi}{L}x\right) dx = A_m(0)\frac{L}{2} \tag{8.44}$$

where we have used the orthogonality property of sine functions:

$$\int_0^L \sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{m\pi}{L}x\right) dx = \delta_{n,m}\frac{L}{2} \ . \tag{8.45}$$

Putting the results of integration of the left and right hand side of eqn.8.43 together, we find

$$A_m(0) = \frac{2}{L} \sin\left(\frac{m\pi}{L}x_0\right) \ . \tag{8.46}$$

Now we turn to the differential equation itself. The question at hand is what should we choose for the coefficients $A_n(N)$ so that the diffusion equation, eqn. 8.40, is satisfied. To figure this out we simply substitute the Fourier expansion of $G(x;N)$ into the differential equation. This yields:

$$\sum_{n=1}^{\infty} \frac{\partial A_n(N)}{\partial N} \sin\left(\frac{n\pi}{L}x\right) = -\frac{a^2}{2} \sum_{n=1}^{\infty} A_n(N) \left(\frac{n\pi}{L}\right)^2 \sin\left(\frac{n\pi}{L}x\right) \ . \tag{8.47}$$

Now we once again use the trick of multiplying both sides of this equation with $\sin(m\pi x/L)$ and integrating from 0 to L. Employing the orthogonality property this time yields a differential equation for the coefficient $A_m(N)$:

$$\frac{\partial A_m(N)}{\partial N} = -\frac{a^2}{2} \left(\frac{m\pi}{L}\right)^2 A_m(N) \ . \tag{8.48}$$

The solution to this equation is an exponential function,

$$A_m(N) = A_m(0) \exp\left(-\left(\frac{m\pi}{L}\right)^2 \frac{a^2}{2} N\right) \ , \tag{8.49}$$

where the coefficient $A_m(0)$ was determined above (eqn.8.46) from the initial condition.

Finally, the solution to eqn.8.40 that satisfies the initial condition that all walkers start at x_0 and the absorbing boundary conditions at the box boundaries, is

$$G(x;N) = \sum_{n=1}^{\infty} \frac{2}{L} \sin\left(\frac{n\pi}{L}x_0\right) \sin\left(\frac{n\pi}{L}x\right) \exp\left(-\left(\frac{n\pi}{L}\right)^2 \frac{a^2}{2} N\right) \ . \tag{8.50}$$

To turn this quantity into the sought out probability distribution for the end-to-end distance of a polymer confined in a box, we make use of eqn.8.41, to yield

$$P(x;N) = \frac{1}{L} \ \frac{\sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{L}x_0\right) \sin\left(\frac{n\pi}{L}x\right) \exp\left(-\left(\frac{n\pi}{L}\right)^2 \frac{a^2}{2} N\right)}{\sum_{n=1}^{\infty} \sin\left(\frac{n\pi}{L}x_0\right) \frac{1}{n\pi}(1-\cos(n\pi)) \exp\left(-\left(\frac{n\pi}{L}\right)^2 \frac{a^2}{2} N\right)} \ . \tag{8.51}$$

This probability distribution is plotted in fig. 8.15a for DNA ($a = 100$nm) confined to a box 2μm in length, for DNA lengths ranging from 0.5μm to 10μm. Note that for the shortest chain the confining box has no effect and the end-to-end distance distribution is a simple Gaussian function, eqn.8.39. For the intermediate chain length, $Na = 2\mu$m, the effect of the box is to skew the distribution owing to the fact that the tethering point, $x_0 = 0.75\mu$m, was chosen closer to the left box boundary. Finally, for very long DNA lengths the distribution is once again symmetric, with all memory of the tethering point lost. This provides us with the quantitative intuition that allows us to conclude that the observed distribution of average positions of markers along the *C.crescentus*

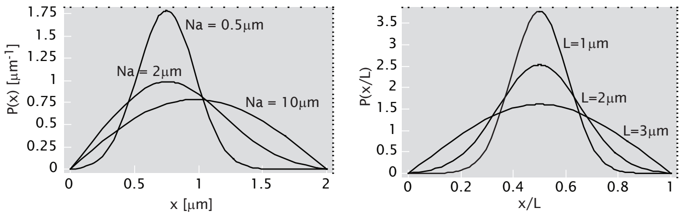


Figure 8.15: A. The distribution of distances to the fluorescent marker for the one-dimensional chromosome model for different contour lengths of the chromatin fiber between the tethering point (at $x_0 = 0.75\mu$m) and the fluorescent marker. The cell size is $L = 2\mu$m, and the packing density and Kuhn length are that of bare DNA. (B) Same as in A, for a 1μm long chromatin fiber confined to cells of different size and tethered in the middle of the cell.

chromosome is inconsistent with a model of a polymer confined to the cell interior which is only tethered at the pole of the bacterium. In other words, further constraints need to be imposed on the chromosome to establish the observed chromosome geography.

In fig. 8.15b we plot once again the end-to-end distance distribution using eqn.8.51, but this time for a $Na = 1\mu$m long DNA molecule ($a = 100$nm) tethered at the center of the confining box, for box sizes ranging from 1μm to 3μm. We note that the effect of confinement sets in rather rapidly: there is little evidence for it in the largest box size, while for the smallest one the distribution is practically that of a very long polymer confined to a small box. This provides an explanation of the difference in the observed distance distributions in the x and y direction for the fluorescent markers placed on the *V.cholerae* chromosome. We can check this assertion quantitatively by fitting the measured x-distribution to the derived formula. This gives two parameters, the position of the assumed tether x_0 and the size of the chain characterized by the quantity Na^2. With the quantity Na^2 in hand and assuming the y position of the tether to be at $y = 0$ (turns out this has little effect given the strong confinement in the y-direction, which, as remarked above, erases the effect of the tether position) we can simply plot the expected y-distribution and ask whether it matches the data. This comparison is shown in fig.8.13. A better match to the data can be achieved by taking the cell to be a cylinder and further taking into account the fact that the y measurement is the projection of the radial distance onto the plane of the cover-slip on which the cells rest.

JK: replace the data fits for Vibrio with the 1d result so that it matches what we do in the chapter. Make the cylinder case a homework

- **The Math Behind the Models: Expanding in Sines and Cosines.** Throughout the book we are often invited to consider functions that are defined on the interval between 0 and L. A useful property of such functions that we employ over and over again is that they can be expanded into a Fourier series:

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n}{L}\, x\right) + b_n \sin\left(\frac{2\pi n}{L}\, x\right) \ . \tag{8.52}$$

Here a_n and b_n are Fourier coefficients, numbers that need to be computed for a given function f. The above equality is true for all points on the interval with the possible exception for $x = 0$ and $x = L$. Namely, since all the functions appearing in the sum on the right hand side take on the same value at 0 and L, we would have to conclude that $f(0) = f(L)$ is also true. If this if not the case, it can be shown that the Fourier series representation of $f(x)$ takes on the value $(f(0)+f(L))/2$ at the boundaries of the interval.

Computing the Fourier coefficients relies on the orthogonality property of sine and cosine functions. Namely, the integral of the product of two such functions is non-zero only in the case when both functions are sines, or both are cosines, and they have the same period; the period of $\sin\left(\frac{2\pi n}{L}\right)$ is L/n. Mathematically stated

$$\begin{aligned}
\int_0^L \sin\left(\frac{2\pi n}{L}\, x\right) \cos\left(\frac{2\pi m}{L}\, x\right) dx &= 0 \\
\int_0^L \sin\left(\frac{2\pi n}{L}\, x\right) \sin\left(\frac{2\pi m}{L}\, x\right) dx &= \delta_{n,m}\, \frac{L}{2} \\
\int_0^L \cos\left(\frac{2\pi n}{L}\, x\right) \cos\left(\frac{2\pi m}{L}\, x\right) dx &= \delta_{n,m}\, \frac{L}{2}
\end{aligned} \tag{8.53}$$

where the Kronecker symbol, $\delta_{n,m}$, is one for $n = m$ and zero otherwise. With these identities in hand, we can compute the Fourier coefficients of the function $f(x)$ by multiplying it with sines and cosines with different periods, and integrating over the interval between 0 and L. Looking at the right hand side of eqn. 8.52 and taking into account the orthogonality identities above, we see that the only surviving term on the right hand side will be the sine or cosine term with the same period. Therefore, we have the following identities

$$\begin{aligned}
\int_0^L f(x) dx = \frac{a_0}{2} \\
\int_0^L f(x) \cos\left(\frac{2\pi n}{L}\, x\right) dx &= a_n \frac{L}{2} \\
\int_0^L f(x) \sin\left(\frac{2\pi n}{L}\, x\right) dx &= b_n \frac{L}{2}
\end{aligned} \tag{8.54}$$

from which we can compute the Fourier coefficients

$$\begin{aligned} a_0 &= \frac{2}{L}\int_0^L f(x)dx \\ a_n &= \frac{2}{L}\int_0^L f(x)\cos\left(\frac{2\pi n}{L}\;x\right)dx \\ b_n &= \frac{2}{L}\int_0^L f(x)\sin\left(\frac{2\pi n}{L}\;x\right)dx\;. \end{aligned} \tag{8.55}$$

Its important to note that the Fourier series representation of the function $f(x)$ on the interval zero to L obtained in this way is not unique. The representation developed above corresponds to a function $F(x)$ that is periodic on the whole x axis, with period L. and is obtained from $f(x)$ by simply repeating it on intervals $(L, 2L)$, $(2L, 3L)$, etc. and $(-L, 0)$, $(-2L, L)$, and so on. Of course, this is not the only way of obtaining a periodic function in x from a function $f(x)$ defined on $(0, L)$. One can for instance take $-f(-x)$ on the interval $(-L, 0)$ and then repeat this new function, now defined on the interval $(-L, L)$, over all interval of length $2L$ that cover the x axis. Unlike the previous procedure such a function would be $2L$ periodic, but would still give a faithful representation of $f(x)$ on the interval of interest, $(0, L)$. Which representation one ends up using is often a matter of convenience.

To illustrate the procedure of expanding a function into a Fourier series, lets consider the simple example given by the function $f(x)$, which is equal to 1 for $0 < x < L/2$ and equal to zero for $L/2 < x < L$. Extending this function to the whole x axis gives a square wave. Fourier coefficients are computed using eqn. 8.55, and we find $a_0 = 2/L$, $a_n = 0$, $b_n = 0$ for n even and $b_n = 2/(\pi n)$ for n odd. How the function $f(x)$ emerges from the Fourier series as more and more terms are kept in the sum is shown in fig. 8.16.

8.2.4 DNA Looping: From Chromosomes to Gene Regulation

The organization of genomes occurs at many different scales. A shorter scale phenomenon of widespread significance is the formation of loops of various kinds in both genomic DNA and RNAs as well. These looping events can be fruitfully examined from the random-walk perspective. Fig. 8.17 shows how nucleic acids form “loops” in a wide variety of different settings. For example, as shown in the previous chapter and illustrated in fig. 8.17(A), melting of DNA results in bubbles of single stranded fragments and the meandering of the single-stranded fragments can be evalulated as a problem in random walks. Similar ideas are relevant in evaluating the propensity of RNA to form hairpin loops. Another favorite example involves the formation of DNA loops by transcription factors as part of the process of gene regulation. Yet another example shown in fig. 8.17(D)

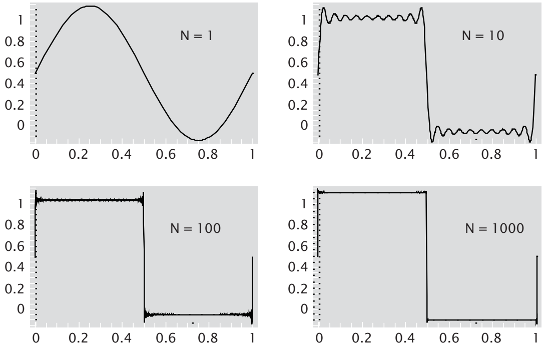


Figure 8.16: Fourier series representation of a square wave. Different graphs correspond to the Fourier series representation of the square wave function where the first N terms have been retained in the sum on the right hand side of eqn. 8.52.

involves genetic recombination in which distant parts of chromosomal DNA find one another as a precursor to the recombination event itself. These events are important in situations ranging from mating type switching in yeast to V(D)J recombination in B cells to the stochastic decision making that attends olfactory receptor selection.

The Lac Repressor Molecule Acts Mechanistically By Forming a Sequestered Loop in DNA

In fig. 4.13 (pg. 182) and section 4.4.3 (pg. 184), we introduced the *lac* operon as a particularly notable example of gene regulation. One part of the *lac* operon story is how the genes of this operon are repressed by the Lac repressor molecule as shown in fig. 8.18. Thus far, our description of Lac repressor has been largely schematic without particular reference to the mechanical actions responsible for repression. The actual story of the action of Lac repressor is more complicated than that illustrated in fig. 4.16 (pg. 187). In fact, there are several other operator sites (O_2 and O_3) in addition to the primary operator site (O_1) described there where the repressor can bind resulting in a DNA loop like that shown in fig. 8.18. The effectiveness of repression is highest when the Lac repressor tetramer (built up from four copies of the *lacI* gene) binds to two operators simultaneously.

Looping of Large DNA Fragments Is Dictated by the Difficulty of the Distant Ends to Find Each Other

In order for a protein molecule such as the Lac repressor to spontaneously form a loop in the DNA, the DNA and protein must together suffer a fluctuation that brings all of the pieces into physical proximity. As will be shown in chap. 10,

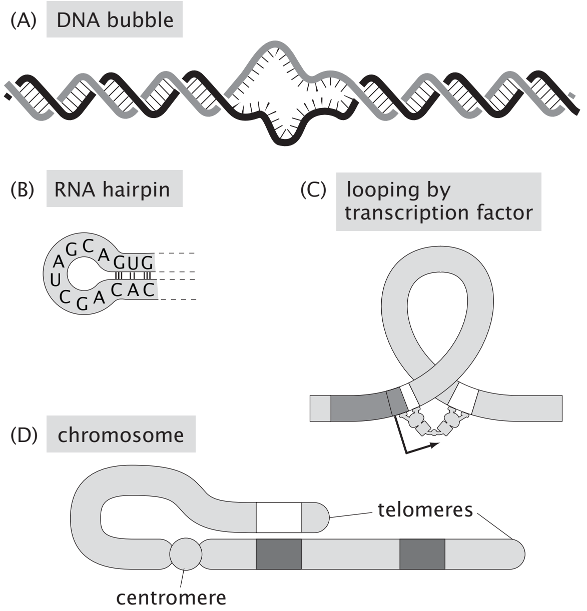


Figure 8.17: Examples of looping. (A) bubble formation in a double-stranded DNA helix, (B) Hairpin loop in RNA secondary structure, (C) DNA looping due to a transcription factor, (D) long distance DNA looping of chromosomal DNA.

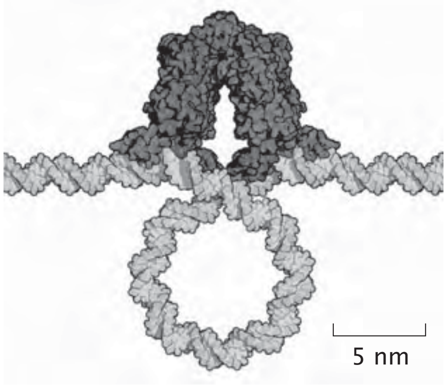


Figure 8.18: Model for DNA loop formation by the Lac repressor. The interface between the protein and the DNA was determined by x-ray crystallography, but the overall position and shape of the DNA in the loop is an artist's rendition.

for the DNA to bend in this way costs elastic energy. However, there is also a contribution to the free energy of looping from entropy since when the DNA is looped, there are fewer conformations available to the system and hence a reduction in the entropy. As a warm-up exercise to evaluate the entropic cost of loop formation we consider a one-dimensional model and examine the fraction of conformations which close on themselves. The probability, $p_\circ$, of loop formation is the probability that the one-dimensional random walker returns to the origin. Using eqn. 8.10, we conclude

$$p_\circ = \frac{\text{number of looped configs.}}{\text{total number of configs.}} = \frac{\frac{N!}{(\frac{N}{2})!(\frac{N}{2})!}}{2^N} \tag{8.56}$$

where N is the number of Kuhn segments. Here we are interested in the long chain limit, which corresponds to $N \gg 1$. This is also the limit in which the random walk model can be applied to DNA conformations, as discussed previously. To further simplify eqn. 8.56 we make use of our trusty Stirling formula, $N! \approx (N/e)^N \sqrt{2\pi N}$, which for $N \gg 1$ implies

$$p_\circ \approx \sqrt{\frac{2}{\pi N}}. \tag{8.57}$$

The interesting prediction of the model is that the cyclization probability of long DNA strands will decay with polymer length to the power $-1/2$.

This result for the probability that the two ends will be within some small distance of each other can also be obtained using the Gaussian approximation to the end-to-end distribution derived earlier in the chapter. To use the continuous distribution, we need the probability that the two ends of the chain are within some critical distance of one another, namely, $\delta \ll \sqrt{Na^2}$. In this case the end-to-end distribution of eqn. 8.16 can be approximated by

$$P(R;N) \approx \frac{1}{\sqrt{2\pi N a^2}} \tag{8.58}$$

where we have made the substitution $\exp(-R^2/2Na^2) \approx 1$, valid for $-\delta < R < \delta$. The cyclization probability is obtained by integrating over all the distances of near contact in the form

$$p_\circ = \int_{-\delta}^{\delta} \frac{1}{\sqrt{2\pi N a^2}}\, dR = \sqrt{\frac{2}{\pi N}} \frac{\delta}{a} \tag{8.59}$$

which is identical to eqn. 8.57 for $\delta = a$.

Unlike the scaling of the polymer size with its length which we found to be independent of the dimensionality of space, the effect of dimensionality on cyclization is quite significant. In particular, the cyclization probability has a different form depending upon whether we evaluate this quantity for one-, two- or three-dimensional random walks. To see this, consider the 3-dimensional random walker of N steps. The probability of returning to the origin can be written as the ratio of the number of walks that return to the origin to the total number of walks in much the same way as we did above (the precise details of this calculation in the discrete language is left to the problems at the end of the chapter). However, a more immediate route to the result can be obtained by exploiting the continuous distribution.

To see this, consider the end-to-end distribution of a three-dimensional random walk. In particular, the probability that the two ends of the chain are at distance δ or smaller, is given by the integral

$$p_\circ = \int_0^\delta 4\pi R^2 P(R;N)dR = \int_0^\delta 4\pi R^2 \left(\frac{3}{2\pi N a^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Na^2}} dR \ . \tag{8.60}$$

Since we are interested in cyclization we can assume that the distance δ is much smaller than the polymer size, $N^{1/2}b$. In this case the exponential function in the integrand can be approximated by one, and the resulting integral is

$$p_\circ = \int_0^\delta 4\pi R^2 \left(\frac{3}{2\pi N a^2}\right)^{\frac{3}{2}} dR = \left(\frac{6}{\pi N^3}\right)^{\frac{1}{2}} \left(\frac{\delta}{a}\right)^3 \ . \tag{8.61}$$

The main conclusion that follows from this calculation is that the cyclization probability decays as the number of Kuhn segments of the chain to the power $-3/2$. In section 10.3 (pg. 463), we will finish these arguments by showing how to link the entropic and energetic description of DNA looping. These ideas will then be applied to compute the probability of gene expression in section 19.2.5 (pg. 873).

8.2.5 PCR, DNA Melting and DNA Bubbles

So far we examined biological processes associated with DNA loops where the double stranded molecule stays intact. During DNA processing by various polymerases loops of single stranded DNA are formed by local melting of the double helix. This melting process is also at the heart of the polymerase chain reaction, which is one of the key tools of modern molecular biology. Here we use random walk models of DNA to consider how complementary base pairing competes with the melted state in which the bases are no longer linked in pairs.

DNA Melting Is the Result of Competition Between the Energy Cost and the Entropy Gain of Separating the Two Complementary Strands

DNA melting is the process by which two strands of the double stranded helix come apart. This is one of the main steps in the polymerase chain reaction (PCR) and it plays a crucial role in transcription and replication since dsDNA needs to be “melted” locally so as to allow RNA or DNA polymerase to initiate transcription or replication. The melting process is a competition between entropy which favors the melted state and energy which is minimized when all the bases are paired up and hydrogen bonds are formed between them. As a result, melting can be induced by an increase in temperature, which changes the relative weights of entropy and energy in the DNA free energy, or, for example, by changing salt concentrations which change the energetics of hydrogen bonding. When a cell needs to melt its DNA helix, it doesn't change the temperature or salt concentration, but rather uses an energy-consuming enzyme called a helicase to pay the energetic penalty of separating the two DNA strands.

The polymerase chain reaction (PCR) has been a revolution within the revolution of molecular biology. PCR permits the amplification of DNA fragments so that these fragments can be used for processes such as cloning genes for expressing insulin in bacteria, finding rare mutations in a population, identifying the origin of a blood sample at a crime scene and comparing the sequence of human vs neanderthal. The basic idea is shown schematically in fig. 8.19. The goal of the PCR reaction is to take some fragment of a DNA molecule and make a huge number of copies of it. In fig. 8.19, it is seen that the reaction consists of the template DNA (the piece to be copied), "primers" which are small ($\approx$ 20bp) DNA fragments that are complementary to sites on the DNA adjacent to the region of interest, DNA polymerase which is the molecular xerox machine that makes the copies and a host of nucleotides (the As, Gs, Ts and Cs) that are the raw material for constructing new DNA molecules.

The way that a typical PCR reaction goes is based on a series of cycles in which the temperature is alternately raised and lowered. The point of raising the temperature is to melt the DNA. Once the DNA has been melted into single strands, there is an annealing step during which the primers bind to their target sites. After this, there is an elongation stage where the polymerase molecules add the appropriate nucleotides to the nascent DNA double helix. Once this cycle is finished, the whole thing is repeated, but now there are more template molecules to use to build new DNA molecules. As a result, the overall concentration of reaction product increases exponentially. Our aim in this section is to perform a simple estimate of one part of the overall PCR reaction, namely, DNA melting. The goal of this estimate is to illustrate some important ideas rather than to shed any deep light on DNA melting or PCR themselves.

DNA Melting Temperatures Can Be Estimated Using a Random Walk Model

A simple model of DNA melting is based on a two-state internal-variable model, like the ones introduced in chapter 7. In this model the base pairs are either in the double-helical state or the melted state. A number of consecutive base-pairs in the melted state are said to form a "bubble". A bubble costs an energy due to the breaking of the favorable hydrogen bonds but is favored by entropy since the single stranded DNA that makes up the bubble is considerably more flexible than its double stranded counterpart and can therefore assume many more configurations. The melting transition is therefore the result of the contest between the energy and the entropy of bubble formation.

To examine this competition quantitatively we consider a simplified version of the so-called Poland-Scheraga model where we allow the formation of only one bubble as shown in fig. 8.20. This is a reasonable assumption for a DNA strand of moderate length ($100 - 1000$bp) as the energy penalty for initiating a bubble

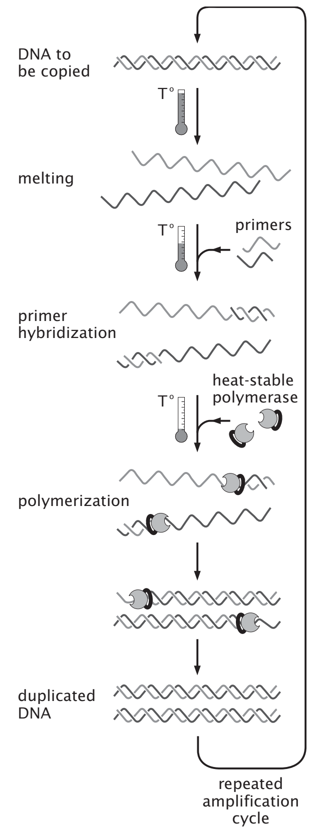


Figure 8.19: Schematic of the polymerase chain reaction (PCR) and its dependence upon DNA melting.

is considerably larger than for elongating the bubble. For short strands the entropy gained by having more than one bubble will not be enough to overcome this energy penalty for bubble initiation.

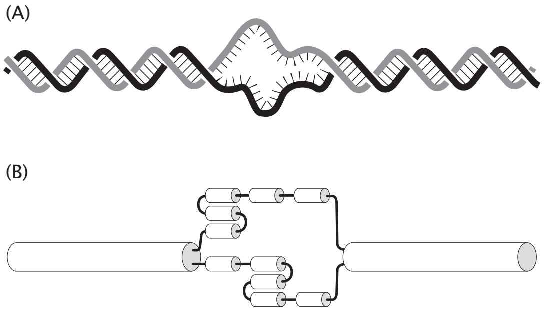


Figure 8.20: One-bubble Poland-Scheraga model. The possible states of a DNA strand of length N base pairs are labelled by the length of the single bubble, $1 \leq n \leq N$. (A) Schematic of a single bubble in the DNA. (B) One-dimensional random walk picture of the DNA with a bubble. The significance of the lengths of the cylinders is to characterize the difference in persistence length between the dsDNA (stiff) and the much more flexible ssDNA. .

The quantity of interest for the one-bubble model is the equilibrium probability that the bubble is of length n base pairs. Statistical mechanics tells us that this probability is given by

$$p_1(n) = \frac{e^{-\Delta G_1(n)/k_B T}}{Z} \tag{8.62}$$

where $\Delta G_1(n)$ is the free energy of formation for a bubble of length n and

$$Z = \sum_{n=1}^{N} e^{-\Delta G_1(n)/k_B T} \tag{8.63}$$

is the partition function of the one-bubble model. The free energy of formation can be written as

$$\Delta G_1(n) = E_{\text{in}} + nE_{\text{el}} - k_B T \ln\left(\Omega_\circ(n)(N-n)\right) \tag{8.64}$$

where E_{in} and E_{el} are the energies for initiating and for elongating a bubble by one base pair, respectively, while $\Omega_\circ(n)$ is the number of ways of making a bubble of two strands of ssDNA each n nucleotides long. The factor $N-n$ accounts for the number of ways of choosing the position along the DNA chain at which the bubble is located. The precise form of the bubble entropy will depend on the polymer model one adopts for the ssDNA. Here, in the name of simplicity, we adopt the one-dimensional random walk model of a polymer. In this case we can write the number of configurations of the part of the DNA that is single stranded

$$\Omega_\circ(n) = 2^{2n} p_\circ(2n) \tag{8.65}$$

which is nothing but the number of random walks of total $2n$ steps that return to the origin, introduced in eqn. 8.56. This reduces to

$$\Omega_\circ(n) = \frac{2^{2n}}{\sqrt{\pi n}} \ . \tag{8.66}$$

for $n \gg 1$, where we have made use of eqn. 8.57 for the cyclization probability $p_\circ(2n)$.

The (reduced) free energy of our one-bubble model of DNA melting is therefore

$$\frac{\Delta G_1(n)}{k_B T} = (\epsilon_{\rm el} - 2\ln 2)\, n + \frac{1}{2}\ln n - \ln(N-n) \ , \tag{8.67}$$

where the energy parameter is given by $\epsilon_{\rm el} \equiv E_{\rm el}/k_B T$, and we have dropped the initiation energy which is the same for all one-bubble states, and another unimportant, n-independent constant.

In order to tease out quantitative intuition provided by this model, we examine how the bubble length n^* at which the free energy is minimum (which is also the most likely bubble length in thermal equilibrium) depends on the temperature, or equivalently, the dimensionless elongation energy $\epsilon_{\rm el}$. Setting the first derivative of the free energy with respect to n to zero, leads to the equation

$$(\epsilon_{\rm el} - 2\ln 2) + \frac{1}{2n} + \frac{1}{N-n} = 0 \tag{8.68}$$

whose solutions are

$$n^*_\pm = N\frac{1 + \Delta\epsilon \pm \sqrt{1 + 6\Delta\epsilon + \Delta\epsilon^2}}{\Delta\epsilon} \tag{8.69}$$

where we have introduced a new variable

$$\Delta\epsilon \equiv 2(\epsilon_{\rm el} - 2\ln 2) \ . \tag{8.70}$$

Consider first the situation when $\Delta\epsilon > 0$. In this case both solutions, $n^*_\pm$ are not of interest as they do not correspond to bubbles whole length is positive and smaller than N. This means, that on the interval $0 < n \le N$ the free energy is monotonically increasing and therefore we expect that the state with no bubble wins out as one with the lowest free energy (this is not 100% guaranteed because the Stirling approximation gets worse as n becomes smaller). Going back to the original parameters in the models this means that for temperatures low enough, so that $E_{\rm el}/k_B T > 2\ln 2$, the no-bubble state wins out. At higher temperatures, when $\Delta\epsilon > 0$, the situation is very different. In this case both solutions $n^*_\pm$ are of interest as they are both positive and less than N. One of the solutions is typically small compared to N and is a local maximum while the other is close in value to N and is a local minimum. In fig. 8.21 we show plots of the reduced

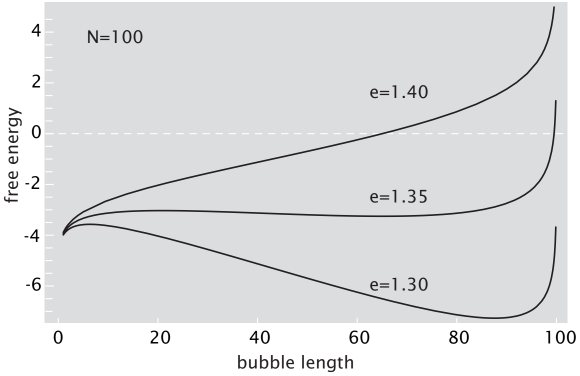


Figure 8.21: Free energy of the one-bubble model as a function of the bubble size. As the temperature is increased ϵ becomes smaller and smaller. At high temperatures the most likely bubble size is close to the DNA length (ie. the chain is completely melted), while for small temperatures it is zero. At intermediate temperatures, the model predicts strong fluctuations of the bubble size.

free energy as given in eqn. 8.67 (ie.without the Stirling approximation) for values of ϵ_{el} close to $2\ln 2 \approx 1.39$ which explicitly demonstrate this behavior. It is interesting to note that for $E_{\text{el}}/k_B T < 2\ln 2$ even though the no-bubble configuration has the lowest free energy one should observe fluctuations into the one-bubble states with a typical bubble size that will depend on temperature. Also, close to the critical value of temperature the free energy as a function of bubble size becomes relatively flat so one should observe bubbles of varying sizes appear simply due to thermal fluctuations.

8.3 The New World of Single Molecule Mechanics

Models such as the random walk model described here have extraordinary reach. Yet another interesting application of these ideas is to the recent development of single-molecule techniques for measuring the response of macromolecules to external forcing.

Single Molecule Measurement Techniques Lead to Force Spectroscopy

There are a number of different ways of applying forces to individual macromolecules. Several of these techniques are represented in schematic form in fig. 8.22. One such technique shown in fig. 8.22(A) involves the use of micron-sized cantilevers which are attached to a macromolecule which is, in turn, tethered to a surface. Through control of the height of the surface to which the molecule is tethered, for example, the cantilever will suffer a deflection which can be measured using reflected laser light. A second example shown in fig. 8.22(B) is optical tweezers which permit the application of forces of order 1-50 pN on macromolecules of interest. In this case, the key idea is that by attaching a macromolecule to a micron-sized bead, it is possible to pull on the bead (and hence the molecule) by shining laser light on the bead and using the resulting radiation pressure from the laser light to manipulate the bead. The same concept is similarly played out in the context of the magnetic tweezers shown in fig. 8.22(C) where the bead is manipulated by magnetic fields rather than laser light. One of the interesting variations on the forcing scheme provided by the magnetic tweezer is the opportunity to apply torsional forces which examine the response of molecules to twist. The final example shown in fig. 8.22(D) is the use of a pipette-controlled force apparatus in which the strengths of ligand receptor interactions as well as the mechanical response of lipid bilayer vesicles can be examined. Our main point in this discussion is to alert the reader to the emergence of single-molecule techniques that complement the tools of traditional solution biochemistry and permit the measurement of not only the average properties of the various macromolecules of biological interest, but also the fluctuations about this average response.

8.3.1 Force-Extension Curves: A New Spectroscopy

Different Macromolecules Have Different Force Signatures When Subjected to Loading

The techniques introduced above permit the explicit measurement of the force-extension characteristics of a range of different molecules. Fig. 8.23 shows the force-extension properties of several characteristic examples ranging from DNA to proteins. In particular, fig. 8.23(A) shows the force-extension characteristics of a single DNA molecule subjected to loading. Note that the same characteristic force-extension signature will be found for a given DNA molecule regardless of which of the various techniques is used to measure it, and further, that this curve provides a unique fingerprint which serves as the basis of *force spectroscopy* of macromolecules. Fig. 8.23(B) shows a plot of the force-extension properties of a particular RNA molecule. Note that the character of the secondary structure associated with a given RNA molecule is translated, in turn, into the character of the force-extension curve, illustrating the idea that the force-extension curve provides a spectroscopic fingerprint of different macromolecules. Fig. 8.23(C) shows yet a third example of the intriguing diversity of force-extension curves associated with different macromolecules, this time revealing how the multidomain protein titin unfolds in the presence of force. One immediate statement that can be made in this example is that the number of load drops in the curve corresponds to the number of domains in the protein. We emphasize that these three examples are but a tiny representation of the broad class of measurements that have been made on polysaccharides, lipids, proteins and nucleic acids as well as their assemblies.

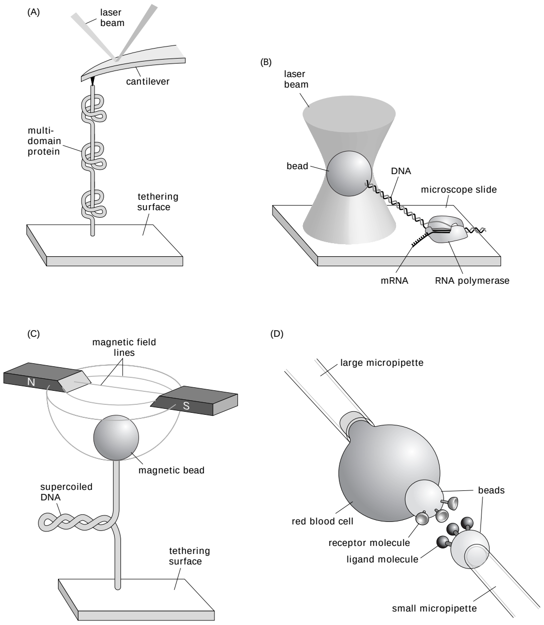


Figure 8.22: Schematic showing a variety of single molecule techniques. (A) single molecule atomic-force microscopy being used to stretch a multi-domain protein, (B) optical tweezers being used to measure the rate of transcription, (C) magnetic tweezers being used to measure the torsional properties of DNA and (D) pipette-based force apparatus being used to measure ligand-receptor adhesion forces.

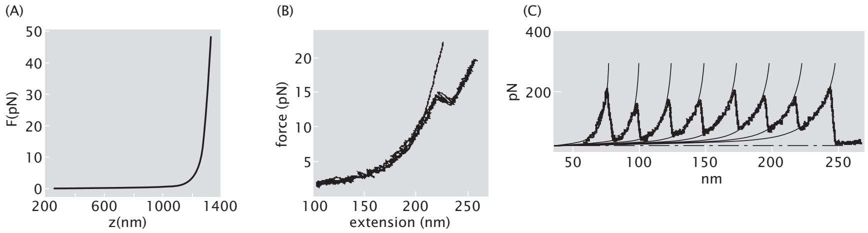


Figure 8.23: Force-displacement curve for a variety of different molecules illustrating the sense in which single molecule experiments serve as the basis of force spectroscopy.

8.3.2 Random Walk Models for Force-Extension Curves

Given that different macromolecules exhibit different force-extension signatures, it is of interest to see if we can compute some characteristics of these curves using what we know about random walks. Indeed, the calculation of these force-extension curves gives us the opportunity to further explore entropic forces.

The Low-Force Regime in Force-Extension Curves Can Be Understood Using the Random Walk Model

One of the simplest models that can be written to capture the relation between force and extension in polymers is based on a strictly entropic interpretation of the free energy. In particular, by remembering that as the chain molecule is stretched to lengths approaching its overall contour length, the overall number of configurations available to the molecule goes down, and with it so too does the entropy. This reduction in entropy corresponds to an increase in the free energy. To the extent that the pulling experiment is done sufficiently slowly, we can think of the force as being given by

$$\text{force} = -\frac{\partial G}{\partial L}, \tag{8.71}$$

where G is the free energy and L is the length.

We begin with a one-dimensional rendition of the freely-jointed chain model. We imagine a polymer of overall length $L_{\text{tot}} = Na$, where N is the number of monomers and a is the length of each monomeric segment. The basic thrust of our argument will be to construct the free energy $G(L)$ as a function of the length $L = (n_r - n_l)a$ from which the force necessary to arrive at that extension is given by eqn. 8.71. As before, we use the notation n_r and n_l to signify how many of the total links are right pointing (n_r) and how many are left pointing (n_l). In order to proceed, we need an explicit formula for the free energy. As noted above, in this simplest of models we ignore any enthalpic contributions to the free energy, with the entirety of the free energy of the molecule taking the form,

$$G(L) = -k_B T \ln W(L; L_{\text{tot}}), \tag{8.72}$$

where $W(L; L_{\text{tot}})$ is the number of configurations of the molecule which have length L given that the total contour length of the molecule is L_{tot}.

As shown in fig. 8.24, we are interested in the equilibrium of our random walk representation of the polymer when it is subjected to external forcing such as can be provided by an optical tweezers setup. A particularly transparent way to imagine this problem is to think of weights being dangled from the ends of the polymer as shown in fig. 8.25 (this idea of representing the energy of the loading device via weights was introduced in fig. 5.12 (pg. 240)). In this case, the free energy of eqn. 8.72 must be supplemented with a term of the form $U_{weights} = -2mgL$. What this term says physically is that the more the molecule is stretched, the lower the weights will dangle with the result that their potential energy is decreased.

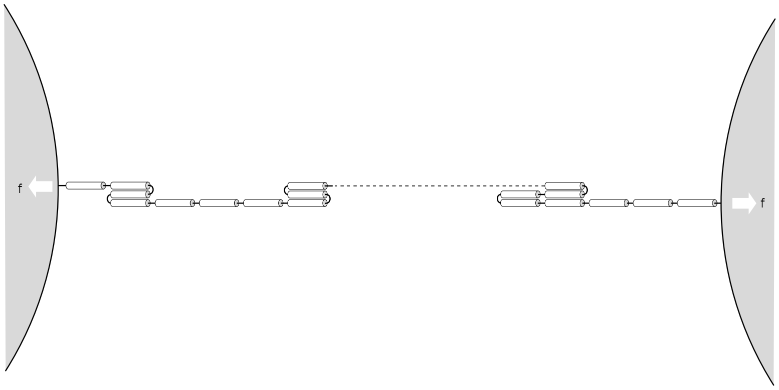


Figure 8.24: Schematic of a model one-dimensional polymer subjected to external forcing. The cartoon is meant to suggest that our one-dimensional random walk polymer is attached to two optical beads at its extremities and that forcing is applied using optical tweezers.

Putting together this term with the contribution from eqn. 8.72, we have for the total free energy of the system

$$G(L) = \underbrace{-2mgL}_{\text{contribution from weights}} - \underbrace{k_B T \ln W(L; L_{\text{tot}})}_{\text{entropic contribution of polymer conformations}}. \quad (8.73)$$

To make further progress with this result, and in particular, to obtain the free energy minimizing length as a function of the applied force, we must first find a concrete expression for $W(L; L_{\text{tot}})$. To that end, we note that this reduces to nothing more than the combinatoric question of how many different ways there are of arranging N arrows, n_R of which are right pointing and $n_L = N - n_R$ of which are left pointing. The result is

$$W(n_R; N) = \frac{N!}{n_R!(N-n_R)!}, \quad (8.74)$$

where we have found it convenient to replace our reference to L and L_{tot} with reference to the number of right pointing arrows and the total number of such arrows with the recognition that they are related by $L = (n_R - n_L)a$ and $L_{\text{tot}} = Na$.

Given the free energy, our task now is to minimize it with respect to length (or n_R). To that end, we first invoke the Stirling approximation (pg. 255), which we remind the reader allows us to replace $\ln N!$ by $N \ln N - N$. In light of this approximation, the overall free energy may be written as

$$G(n_R) = -2Mgan_R + k_B T(n_R \ln n_R + (N - n_R)\ln(N - n_R)). \quad (8.75)$$

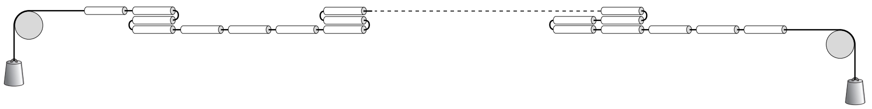


Figure 8.25: Schematic of a model one-dimensional polymer subjected to external forcing through the attachment of weights on the end. This scenario is a pedagogical device to illustrate how to include the forcing in the overall free energy budget.

Note that we have neglected all constant terms since they will not contribute during the minimization. Differentiation of this expression with respect to n_R results in

$$\frac{\partial G}{\partial n_R} = -2Mga + k_b T \mathrm{ln} n_R - k_B T \mathrm{ln}(N - n_R) = 0 \tag{8.76}$$

which may be rewritten in a more transparent fashion as

$$z = \frac{\langle L \rangle}{L_{\text{tot}}} = \tanh \frac{mga}{k_B T}. \tag{8.77}$$

The construct of using weights to load the molecule was a convenient pedagogical device to provide a concrete mechanism for seeing how the energy of the loading device can be included in the free energy budget. More generally, the two ends are subjected to a force f with the result that $z = \tanh(fa/k_B T)$. This force-extension relation is shown in fig. 8.26. To gain further insight into the quantitative aspects of the model we consider the limiting case of a small force, i.e. $fa \ll k_B T$. For a dsDNA molecule in physiological conditions ($a \approx 100$nm) this corresponds to $f \ll 40$ fN while for the much more flexible ssDNA ($a \approx 1.5$nm) the small force regime is obtained for $f \ll 3$ pN. In the small force limit the force-extension curve is linear (as shown in the problems at the end of the chapter),

$$\langle L \rangle = \frac{L_{\text{tot}} a}{k_B T} f \ , \tag{8.78}$$

ie. in this regime the polymer behaves like an ideal Hookean spring with a stiffness constant $k = k_B T / L_{\text{tot}} b$. The fact that the stiffness of this spring is linearly proportional to the temperature reveals its true entropic nature. For λ-phage dsDNA whose contour length is $L_{\text{tot}} = 16.6$ μm the effective spring constant is $k \approx 2.3$ fN/μm while for the same length ssDNA the stiffness is given by $k \approx 160$ fN/μm. Note that the larger flexibility of ssDNA, as evidenced by its smaller persistence length, leads to a larger value for the effective spring stiffness.

Thus far, our model of the macromolecule has been highly idealized in that we have imagined that each monomer can only point in one of two directions. Though that model is instructive, clearly it is of interest to expand our horizons to the more physically realistic three-dimensional case. The generalization of our freely-jointed chain analysis to three dimensions holds no particular surprises.

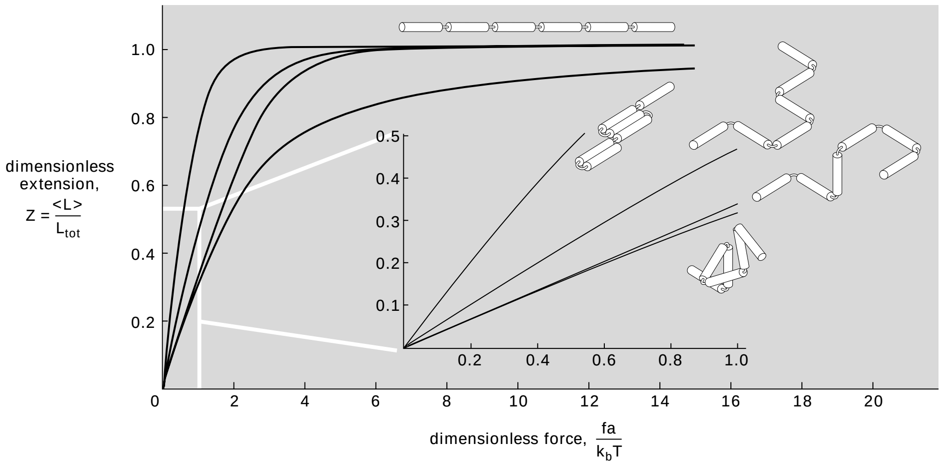


Figure 8.26: Relation between force and extension as obtained using the freely jointed chain model. Results for one-, two- and three-dimensions are shown and the three-dimensional case is shown for both the version in which the monomers can only point in the Cartesian directions and for the case in which they can point in any direction. The curves are related to their corresponding model by the cartoon showing the random-walk chain.

The fundamental idea is that now instead of constraining the monomers that make up the molecule of interest to point only right or left, we give them full three-dimensional motion. The simplest variant of this model is to permit each monomer to point in one of six directions (i.e. $\mathbf{e}_1$, $-\mathbf{e}_1$, $\mathbf{e}_2$, $-\mathbf{e}_2$, $\mathbf{e}_3$ and $-\mathbf{e}_3$). We quote the result for this model, namely,

$$z = \frac{\langle L \rangle}{L_{\rm tot}} = \frac{2\text{sinh}\beta f a}{4 + 2\text{cosh}\beta f a}, \tag{8.79}$$

and leave the details as an exercise for the reader.

The more interesting case which we work out in greater detail is that in which each monomer can point in *any* direction. In this case, rather than writing out the free energy explicitly, we compute the partition function and use it to deduce the relevant averages, such as the average length at a given applied force. As each link in the chain is independently fluctuating the partition function for $N = L_{\rm tot}/a$ links is $Z_N = Z_1^N$ with

$$Z_1 = \int_0^{2\pi} d\phi \int_0^{\pi} e^{fa\cos\theta/k_BT} \sin\theta d\theta. \tag{8.80}$$

This equation instructs us to compute the Boltzmann factor for every orientation of the monomer (characterized by the angles ϕ and θ. The integral over the unit sphere can be evaluated with the change of variables $x = \cos\theta$, to give

$$Z_1 = 4\pi \frac{k_B T}{fa} \sinh \frac{fa}{k_B T}. \tag{8.81}$$

Now the free energy $G(f) = -k_B T \ln Z_N$ is a function of the applied force f and we differentiate it with respect to f to obtain an expression for its thermodynamic conjugate, the average polymer length,

$$\langle L \rangle = -\frac{\partial G}{\partial f} = Na \left(\coth \frac{fa}{k_B T} - \frac{k_B T}{fa} \right). \tag{8.82}$$

The small force limit, $fa/k_B T \ll 1$ in this case gives the same Hookean expression, $f = k\langle L \rangle$ as the one-dimensional freely jointed chain, except the effective spring constant is three times as large, $k = 3k_B T/L_{\text{tot}} a$. The same result follows from eqn. 8.79. Not surprisingly, the two-dimensional version of the model, whether it be defined on a lattice or not, gives $k = 2k_B T/L_{\text{tot}} a$.

8.4 Proteins as Random Walks

So far, we have shown how the random walk model can be applied to nucleic acids. Similar ideas have proven useful for thinking about proteins as well. Globular proteins in their native state form compact structures. One of the key ideas driving research in structural biology, which seeks to describe protein structure in atomic detail, is that protein function follows from its structure. Proteins are polymers comprised of amino acids. Therefore, a natural question to ask is what, if any, aspects of protein structure can be understood from simple coarse-grained models of polymers, such as the various random walks introduced in this chapter.

In this section we examine a lattice model of proteins, the compact polymer model. Usually when representing the polymer by a random walk on a lattice, the sites not occupied by the monomers are thought of as representing the solvent. Random walks described in the previous sections are open structures with the monomer sites typically surrounded by solvent sites. This is inadequate for describing protein conformations which are compact with solvent typically making contact only with amino-acids at the surface of the protein. To mimic this property of proteins we invoke compact random walks (also referred to as Hamitonian walks) which are self-avoiding random walks that visit every site of the lattice, usually taken to be cubic; see fig. 8.27. By virtue of covering all the lattice sites by monomers, all the solvent sites are pushed to the surface. These compact random walks, are a very coarse grained model of proteins and, as with all coarse-grained models, one is limited in scope and precision of the questions that the model is equipped to address. The rewards on the other hand come in the form of simplicity and generality of the answers obtained. Furthermore, as any good model does, compact random walks also reveal new questions and sharpen old ones, about the structure of naturally occurring proteins.

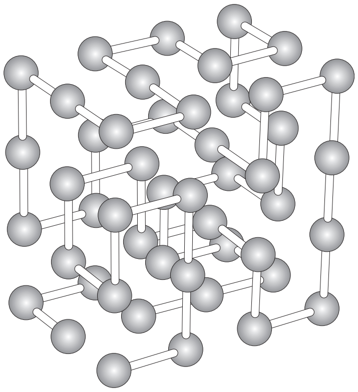


Figure 8.27: Compact polymer configuration on a 4x4x3 cubic lattice. Taken from Dill's Protein Science review

8.4.1 Compact Random Walks and the Size of Proteins

Random Walk Models Permit an Estimate of the Size of Proteins

Possibly the simplest property of a globular protein is its size, as measured by its linear dimensions, or more precisely, its radius of gyration. The Protein Data Bank reveals a systematic dependence of the protein size on its mass. Namely, for globular proteins, the radius of gyration scales roughly with the cube root of the mass. The relation between the physical size of proteins and their sequence size is shown in fig. 8.28. This is a property of compact polymers as witnessed by the configuration shown in fig. 8.27. As a compact polymer completely fills the lattice, its linear size will scale with the linear dimension of the lattice or with the cube root of the number of lattice sites, given that we have in mind a three-dimensional lattice. If we attach a single residue with each site, and take these to be of roughly equal mass, we arrive at the scaling law observed for real proteins. Compactness implies that all the space occupied by proteins is filled, with no holes present. Therefore, the volume occupied by the protein, which necessarily scales as the cube root of its linear dimension, is proportional to the mass. For proteins in the unfolded state the structures are better described as random walks. The size of a random walk polymer, unlike compact polymers, scales as the 1/2 power of the mass. If one were to examine random self-avoiding walks, an argument due to Flory predicts scaling of the linear size with mass to the 3/5 power, indicating a structure which is even more expanded that that of a simple random walk.

Compact Polymers Exhibit Secondary Structure Motifs Similar to Proteins

When examining the structure of globular proteins one of their most striking features is the preponderance of symmetric motifs such as helices and sheets,

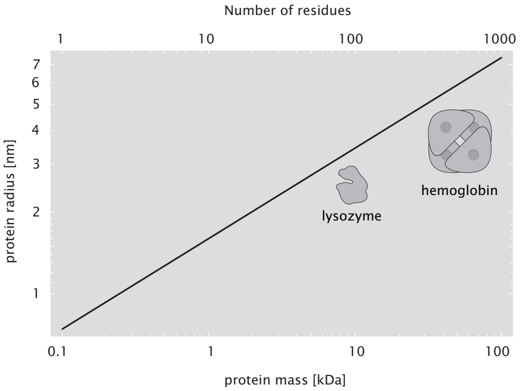


Figure 8.28: Scaling of protein size as a function of the number of amino acid residues.

which are referred to as the secondary structure. These are precisely the features of protein structure accentuated by ribbon diagrams. What gives rise to secondary structure? One idea is that secondary structure motifs are a consequence of the compact state of the protein which in turn is affected by the hydrophobicity of amino-acid residues. If indeed compactness alone drives secondary structure formation then compact polymers should also exhibit a rather large tendency towards secondary structure motifs. This hypothesis is readily testable in the lattice model.

First we need to define secondary structure motifs on the lattice. There is a certain amount of arbitrariness to this and we must be careful in interpreting the results. One possible definition, for the case of two dimensional compact polymers is given in fig. 8.29(A). These results are calculated by taking the ensemble of all possible compact random walks, and for each such structure the percentage of residues taking part in secondary structure motifs, such as helices, sheets and turns, is computed. For small structures this combinatorial problem can be done by hand, but on larger lattices a computer needs to be employed. The distribution of the percent of residues participating in secondary structure over the ensemble of compact polymers, for different polymer sizes, obtained in this way is shown in fig. 8.29(B).

From fig. 8.29(B) we see that the percentage of monomers participating in secondary structure approaches 70% as the size of the compact polymer increases. Of course this number will vary depending on the precise definition of secondary structures on the lattice. Nonetheless, the lattice model predicts that compactness alone can lead to secondary structure. These observations have lead to more detailed computer studies using compact polymer models that are no longer restricted to lattice sites. These have shown that compactness can aid in the formation of secondary structures but that specific interactions between residues in close proximity of each other are also required to produce

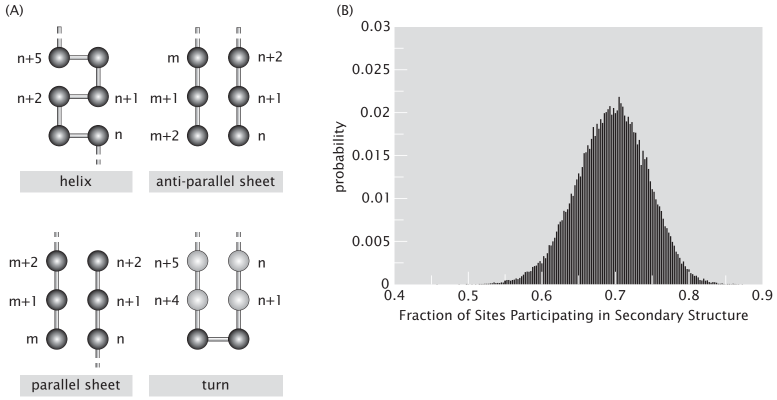


Figure 8.29: "Secondary structure" in lattice models of proteins. (A) Monomers shaded dark participate in secondary structure motifs: helices, parallel and anti-parallel sheets, and turns. (B) Histogram of the fraction of monomers participating in secondary structure motifs for compact polymers on a 20 × 20 square lattice (total 400 monomers).

the observed secondary structure motifs.

One of the challenges brought in on the heels of the successes of the great genomic sequencing initiatives is that of figuring out the structural and functional implications of these vast libraries of genes. One step in unraveling the meaning of all of this genomic data is to figure out how to go from a particular protein sequence to the corresponding structure. The problem is that when confronted with some new genome sequence, one would like to be able to state what proteins are implied by the various sequences and what structures these proteins have. Like for the analysis of protein-ligand binding in chap. 7, here too we will find that the use of internal-state variables to characterize the amino acid identity of a given residue is extremely powerful.

The process by which a chain of amino acids assumes the specific three-dimensional native structure of a protein is often not understood in enough detail to allow for a prediction of the structure based on the known sequence. The complexity of the problem is illustrated in part by the observation that the number of possible three-dimensional conformations of a protein is so large that a random search in structure space would never uncover the native state. Though nature is clever enough to wiggle its way out of this problem, sometimes we are not. Even if we are to model structures using a highly simplified and contracted scheme in which a given structure is viewed as random walks on a cubic lattice as introduced above, the number of structures for a 100-monomer chain is 6^{100} or 6.5×10^{77}. The way we obtain this estimate is based on the idea that the link connecting every successive set of residues can point in one of the 6 directions along the three Cartesian axes. If we imagine doing a random search among these structures at a (very optimistic) rate of one structure per femtosecond (10^{-15} seconds), it would take roughly 2×10^{55} years to complete the search. This is about 10^{45} times the age of the Universe!

8.4.2 Hydrophobic and Polar Residues: The HP Model

The above estimate tells us that the folding of a protein into its native structure is most certainly not a random process. The hydrophobic interaction between amino-acid residues and the water molecules that surround them leads to a collapse of the chain as was illustrated in fig. 5.8 (pg. 236). As a result the hydrophobic residues are sequestered to the interior of the protein, while the surface is populated by polar residues. Thus hydrophobicity is one force that can steer the protein to a folded state avoiding a random search of configuration space. Indeed, the spirit of the class of models introduced here is that collapse induced by hydrophobic effects drives the formation of secondary structure as opposed to an alternative view in which the formation of the hydrogen bonds that define secondary structure lead to collapse.

The HP Model Divides Amino Acids Into Two Classes: Hydrophobic and Polar

The idea that the hydrophobic force plays a prominent role in protein folding has led to coarse-grained models of proteins where the 20 naturally occurring amino acids are replaced with a two-letter alphabet that identifies each amino acid as being hydrophobic (H) or polar (P). This leads to a drastic reduction of the complexity of the sequence space as the number of possible sequences for a 100-mer goes down from $20^{100} \approx 10^{130}$ to $2^{100} \approx 10^{30}$. To implement such a model, we need to decide how to partition the 20 amino acids into the two categories H and P. An example of such a partitioning is shown in fig. 8.30. Indeed, as shown in fig. 8.31, there is a hierarchy of possible classifications of the amino acids based on various properties for grouping them.

In the remainder of the book, we will use the HP model introduced here as the basis of a variety of different discussions. Our reasoning is that classifying amino acids according to just these two broad categories allows us to take otherwise analytically intractable problems and to render them tractable. For example, in section 18.4.1 (pg. 843), we will consider an HP model of translation and kinetic proofreading featuring only two species of tRNA. This simplification will allow us to carry out the analysis completely. Similarly, the entirety of chap. 18 on bioinformatics will be based on sequence alignments using only the HP alphabet. Though we compromise on biological realism, our sense is that the pedagogical payoff is worth it.

8.4.3 HP Models of Protein Folding

The protein folding problem of finding the native structure given the amino acid sequence of a protein is one of a class of problems concerning the relationship between the sequence space and the space of three-dimensional structures. Just

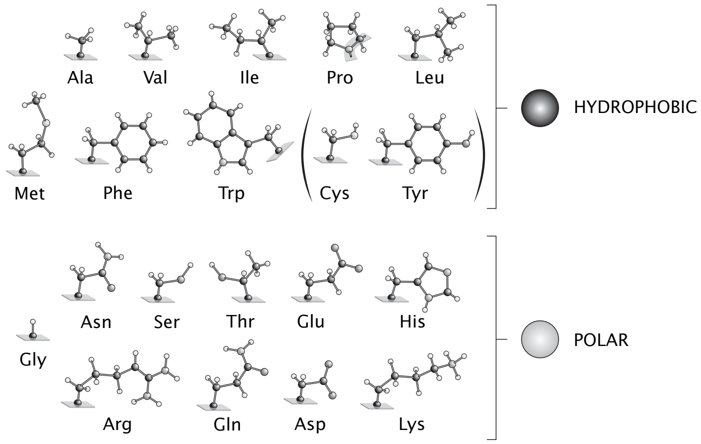


Figure 8.30: Mapping of the amino acids onto an HP alphabet. The 20 amino acids are coarsely separated into two categories, hydrophobic (H) or polar (P).

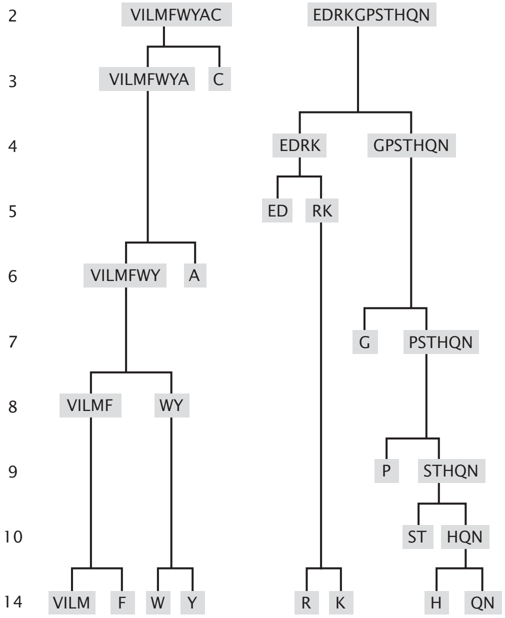


Figure 8.31: Hierarchy of amino-acid classifications. Groupings of amino acids into "classes" with similar properties. At the top of the figure, the amino acids are grouped into two categories, hydrophobic (H) on the left and polar (P) on the right. At each level the number of distinct classes is shown by the integer on the left.

as introducing a two letter alphabet greatly reduces the sequence space, constraining the space of structures to compact random walks on a lattice makes the exploration of structure space more tractable. In particular the number of compact polymer structures on a $3 \times 3 \times 3$ lattice, often used in numerical studies, is 103,346, while the number of possible sequences is $2^{27} = 134, 217, 728$.

To gain intuition about lattice HP models we investigate the toy model that consists of 6 monomers on a 2×3 lattice. The number of possible *sequences* is $2^6 = 64$ while the number of compact structures that are unrelated by lattice rotations, translations or reflections is only 3. These are shown in fig.8.32(A). The final ingredient of the model is the hydrophobic energy which measures the extent to which the H-monomers make energetically unfavorable contacts with the solvent. A simple model of this interaction is to assign a free energy penalty ϵ for every H monomer in contact with either a solvent molecule or a P monomer. (A more refined model might distinguish the interaction energy associated with an H-solvent and an H-P contact.) Solvent molecules are represented as lattice sites not occupied by the monomers, while a contact is a bond between two nearest neighbor sites not occupied by the polymer chain.

The protein folding problem within this toy model can be formulated in the following way: Given an HP sequence which of the possible structures minimizes the hydrophobic interaction energy? We examine two sequences in light of this question: HPHPHP and PHPPHP. The energies for each of these two sequences in each of the 3 possible compact configurations are given in fig.8.32(B). We see that the first sequence has the same energy regardless of the compact conformation the 6-mer assumes. This implies that independent of temperature the probability of finding the polymer in any of the three compact conformations is 1/3. Such a sequence is not protein-like in the sense that it does not have a unique low energy, native state.

On the other hand the sequence PHPPHP has a unique native state, the Π shaped conformation shown in fig.8.32(B). The probability of finding the chain in the native conformation is proportional to the Boltzmann factor associated with its energy,

$$p_{\text{fold}} = \frac{e^{-2\beta\epsilon}}{e^{-2\beta\epsilon} + 2e^{-4\beta\epsilon}} \; ; \tag{8.83}$$

the denominator is nothing but the partition function for the three possible conformations. The probability of this toy protein to be in the folded state as a function of temperature is shown in fig.8.33. Note the sigmoidal character of the plot which is characteristic of many real proteins.

Another interesting question we can pose in the context of this toy model of folding is: What sequences are protein like? Such questions are practically impossible to address in more realistic models of proteins given the astronomically large (literally!) number of sequences and conformations. The hope is that by asking these types of questions in simple lattice models one might uncover patterns that are also present in real proteins.

In the context of our toy model we can address this question systematically if we notice that a necessary condition for a sequence to have a unique native

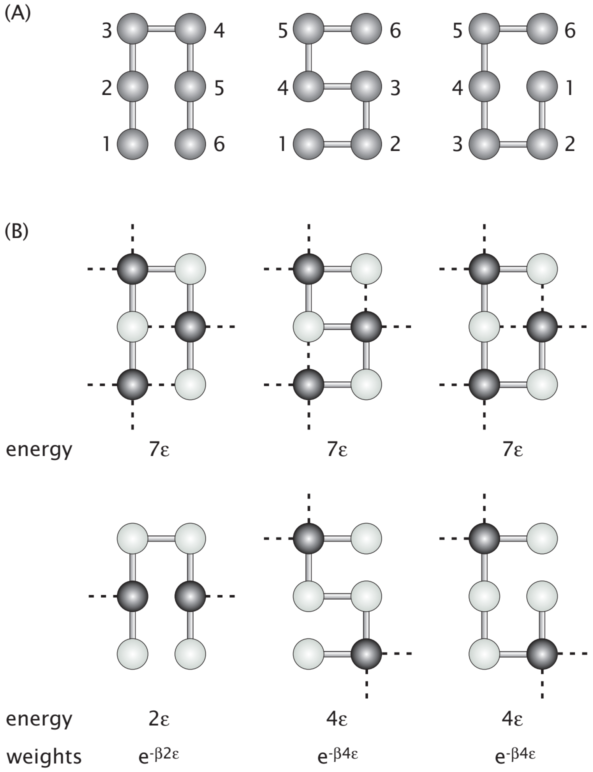


Figure 8.32: Lattice models of protein folding in the HP model. (A) Possible compact conformations of an HP 6-mer in a toy model of protein folding. (B) The hydrophobic energy of an HP polymer in a particular compact conformation depends on its sequence. Sequences which have a unique lowest energy state are protein-like.

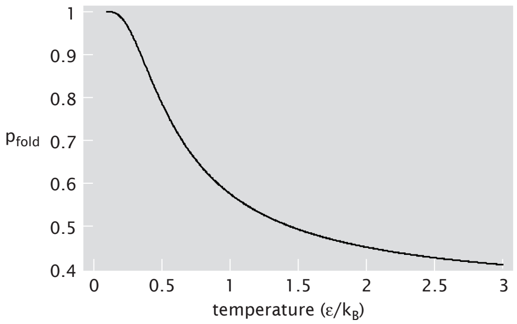


Figure 8.33: Probability of finding the PHPPHP polymer in its native state.

conformation is for there to be at least one HH contact, like the one between the two H monomers in the native state of the PHPPHP sequence in fig.8.32(B). Then we can construct for each of the 3 possible compact structures all the sequences that have that particular structure as its unique native state. One strategy is to begin by choosing two residues that are in contact in the chosen conformation and not in any other; for example this is the case for residue 2 and 5 in the Π structure. We make both these residues an H and then we assign an H or a P to all the others so that no contacts are made in any of the other compact conformations. The outcome of implementing this algorithm is shown in fig.8.34.

An interesting feature of this model is that it predicts the Π structure to be the most designable one. Namely, this structure has 9 sequences of total 64 which fold into it. The least designable structure has only 3 sequences that fold into it. This observation suggests a question whether observed protein structures in Nature are highly designable or not.

The HP model of proteins suggests an interesting strategy for protein design. The idea is to use the degeneracy of the genetic code to create a library of amino-acid sequences which are identical when translated into HP language. For any particular sequence the amino acids are chosen randomly from the pool of H or P residues. For example, a four-helix bundle has been designed by following the pattern: HPPHHPPHPPHHPPH... which ensures that there is a hydrophobic residue every three or four amino acids in the sequence; see fig. 8.35. This is consistent with the structural repeat of 3.6 amino-acids per turn of an alpha-helix. It has been shown experimentally that these sequences not only properly fold into helices but also have enzymatic activity. Identical design principles have been used to make β-sheets which can aggregate into structures akin to amyloid fibers.

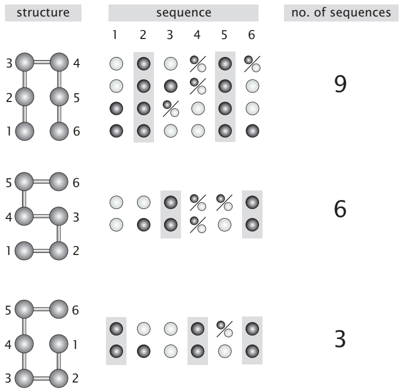


Figure 8.34: Protein-like sequence fold into a unique compact conformation. The number of protein-like sequences varies from compact structure to compact structure. The structures with a particularly large number of protein-like sequences associated with them are highly *designable*.

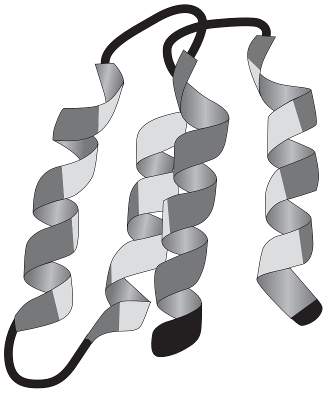


Figure 8.35: The four-helix bundle designed by using an HP sequence strategy.

8.5 Summary and Conclusions

The random-walk model is useful in many different scientific settings. One powerful application of these ideas is to the structure and properties of polymers, including many of the “giant molecules” of life. In this chapter, we have shown how simple ideas from the physics of random walks can be used to explore the size and distribution of DNA, the force-extension properties of polymers and the emergence of entropic elasticity and as a toy model that captures some of the features of protein folding.

8.6 Further Reading

Grosberg A. Y. and Khokhlov A. R., **Giant Molecules**, Academic Press, San Diego, California, 1997. This book is a thoughtful discussion of polymer physics that is pleasing to novices and professionals alike.

Benedek G. B. and Villars F. M. H., **Physics With Illustrative Examples from Medicine and Biology: Statistical Physics**, Springer-Verlag, Inc., New York: New York, 2000.

Berg H., **Random Walks in Biology**, Princeton University Press, Princeton: New Jersey, 1993.

Doi M., **Introduction to Polymer Physics**, Oxford University Press, Oxford: England, 1995.

Doi M. and Edwards S. F., **The Theory of Polymer Dynamics**, Oxford University Press, Oxford: England, 1986.

Chandrasekhar article
de Gennes P.-G., **Scaling Concepts in Polymer Physics**, Cornell University Press, Ithaca: New York, 1979.

K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas and H. S. Chan, "Principles of protein folding - A perspective from simple exact models", Protein Sci., **4**, 561 (1995) and K. Dill, "Polymer principles and protein folding", Protein Sci., **8**, 1166 (1999). These articles give many interesting insights into the use of lattice models and reduced alphabet amino acid repertoires to examine protein folding.

M. H. Hecht, A. Das, A. Go, L. H. Bradley and Y. Wei, "De novo proteins from designed combinatorial libraries", Protein Sci., **13**, 1711 (2004). This very interesting review describes the use of the HP model in carrying out protein design.

8.7 Problems

How Big is a Genome?

(a) In the text, we claimed that the radius of gyration of a polymer can be written in the form

$$\sqrt{\langle R_G^2 \rangle} = \sqrt{\frac{L\xi_p}{3}}. \qquad (8.84)$$

In this part of the problem, deduce this relation.

(a) Compute the entropic size of the DNA associated with a human chromosome if it is not associated with any proteins.

Entropic Cost of DNA Packing. Work out the free energy cost associated with packing the *E. coli* genome inside of the bacterium assuming that the entirety of this free energy cost is entropic. (RP: be careful about the Flory-Huggins version of this story that Ken likes to talk about). RP: also do the problem for the virus.

30 nm Fiber and Packing

Use the numbers for packing density (ν) and persistence length of 30 nm fiber to estimate the R_G of each chromosome and compare to the 10 nm case.

End-to-end distribution

(a) Complete the algebra leading up to eqn. 8.15, for the probability distribution of the end-to-end distance of a one-dimensional freely jointed chain.

(b) Compute the average end-to-end distance $\langle R\rangle$, and $\langle R^2\rangle$ using the same continuous Gaussian distribution. Compare your results to those obtained using the binomial distribution.

Cyclization in 3D

(a) Do the discrete calculation as a ratio to get the cyclization probability.

Force-Extension in the Freely Jointed Chain.

Work out the force extension properties of the three dimensional freely jointed chain, both for discrete and continuous allowed orientations. Key point: show the linearized version leads to Hooke's law.