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"

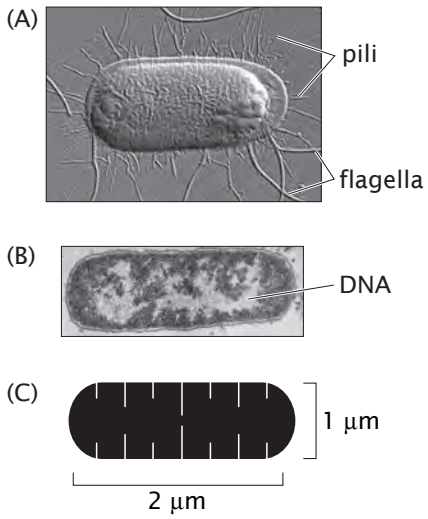


Figure 2.1: *E. coli* as a standard ruler for characterizing spatial scales. (A) Atomic force microscopy image of an *E. coli* cell (courtesy of C. T. Lim), (B) Electron micrograph of *E. coli* bacterium, (C) the *E. coli* ruler.

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 $V_{E.coli} \approx 1\ \mu\text{m}^3 = 1$ 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

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

  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

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

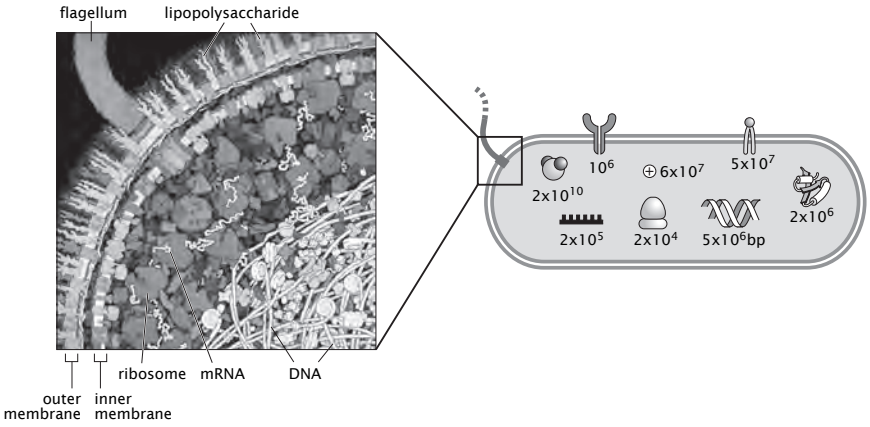


Figure 2.2: Molecular contents of the bacterium *E. coli*. The cartoon on the left shows the crowded cytoplasm of the bacterial cell. The cartoon on the right shows an order-of-magnitude molecular census of the *E. coli* bacterium with the approximate number of different molecules in *E. coli*.

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

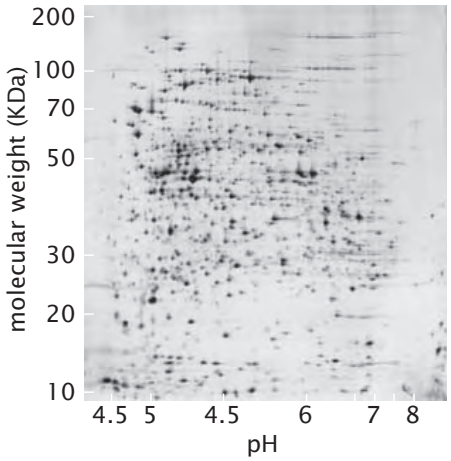


Figure 2.3: Experimental census of the cell. Measurement of protein census using two-dimensional polyacrylamide gel electrophoresis. Figure adapted from the 2DPage database.

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.*.

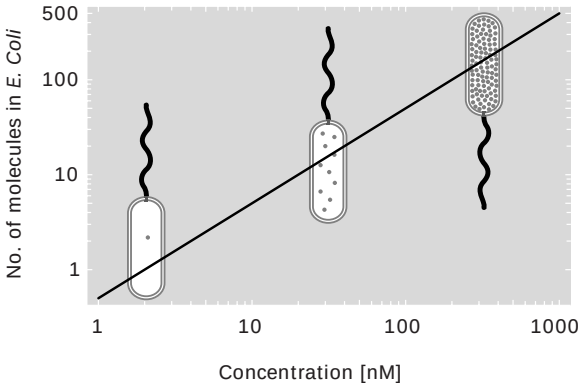


Figure 2.4: Concentration in *E. coli* units. Number of copies of a given molecule in a volume the size of an *E. coli* cell as a function of the concentration.

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

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

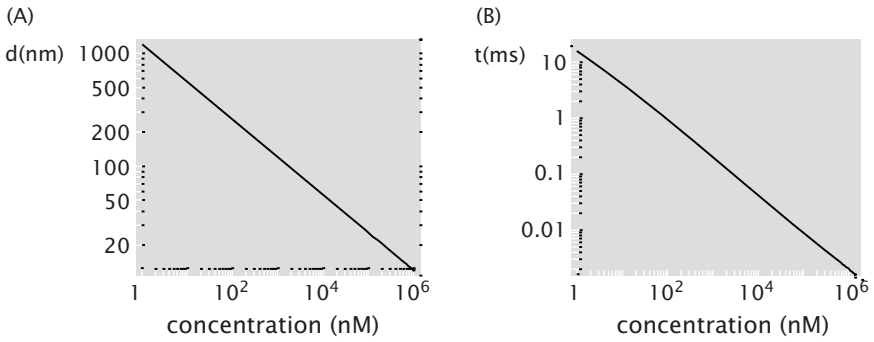


Figure 2.5: Different representations of concentration. (A) Concentration expressed in units of typical distance (d) between neighboring molecules measured in nanometers. (B) Diffusion time over the distance between neighboring molecules as a function of concentration. The diffusion constant $D = 100\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

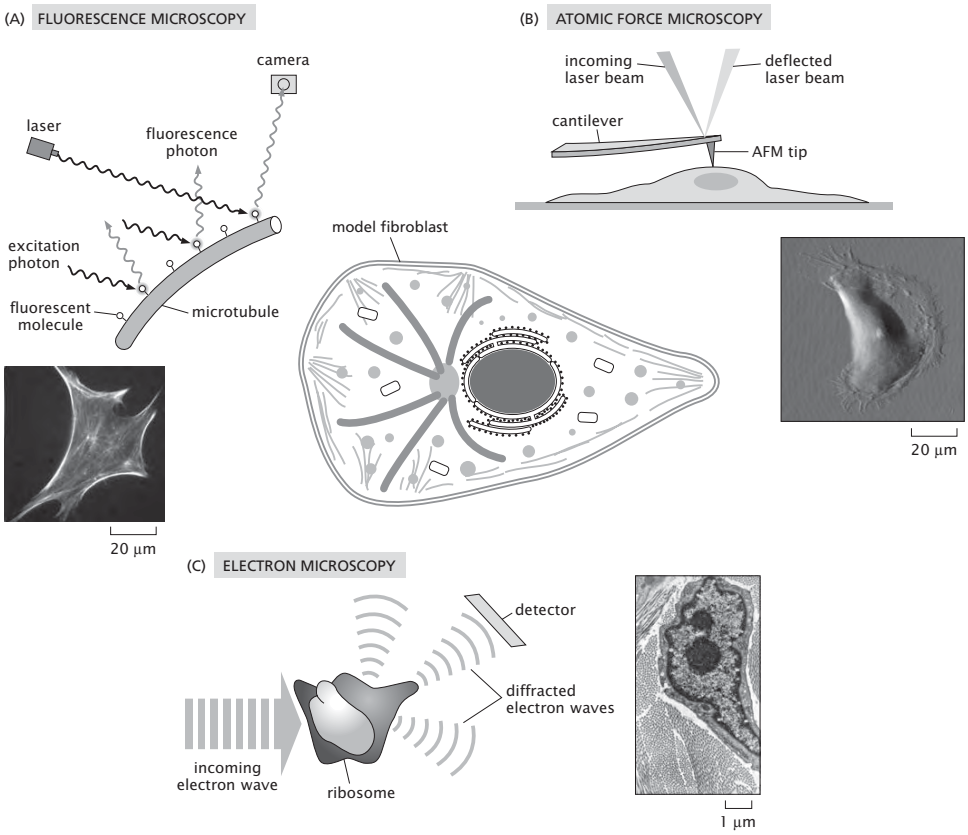


Figure 2.6: Experimental techniques which have revealed the structure of both cells and their organelles. (A) Fluorescence microscopy and associated image of fibroblast with labeled actin, (B) Atomic force microscopy schematic and associated image of surface topography of fibroblast. (C) Electron microscopy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

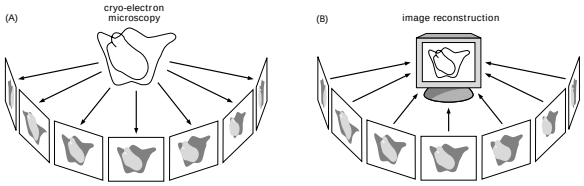


Figure 2.7: Schematic of tomographic reconstruction. (A) The sample is rotated so that radiation is scattered from a series of different orientations, (B) three-dimensional reconstruction of the structure giving rise to the pattern of scattering.

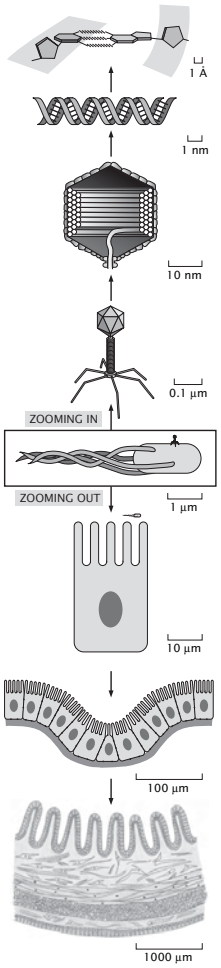


Figure 2.8: Powers of ten representation of biological length scales. The hierarchy of scales is built around the *E. coli* standard ruler. Starting with *E. coli* the first part of the chapter will consider a succession of tenfold increases in resolution as are shown in the figure. The second part of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

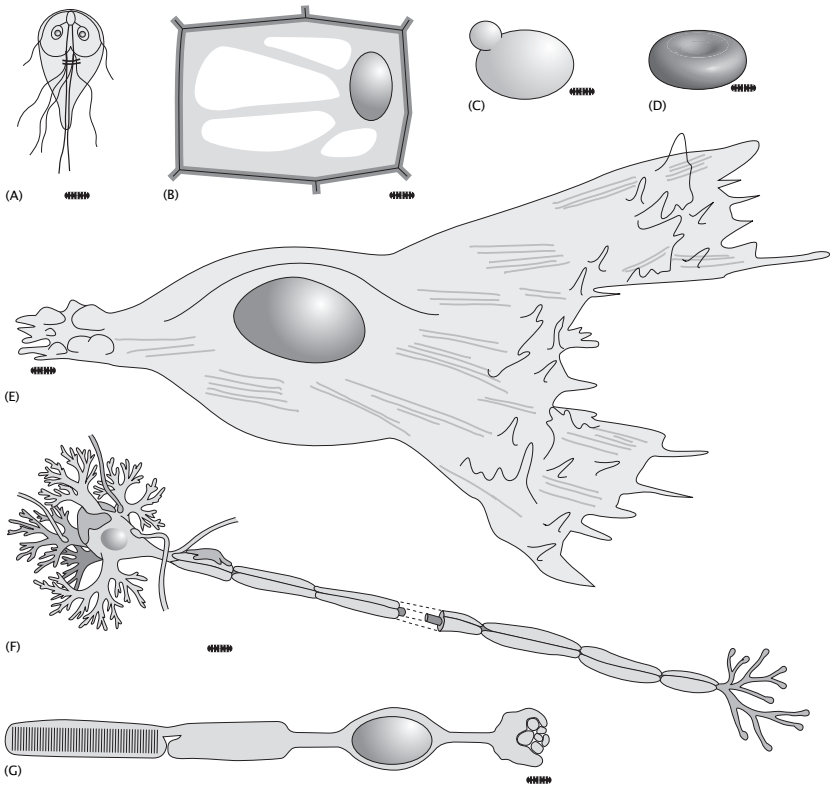


Figure 2.9: Cartoons of several different types of cells all referenced to the standard *E. coli* ruler. (A) the protist *Giardia lamblia*, (B) plant cell, (C) *Saccharomyces cerevisiae*, yeast cell (D) red blood cell, (E) fibroblast cell, (F) eukaryotic nerve cell and (G) rod cell.

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 $V_{E.\ coli} \approx 1.0\mu\text{m}^3$ 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

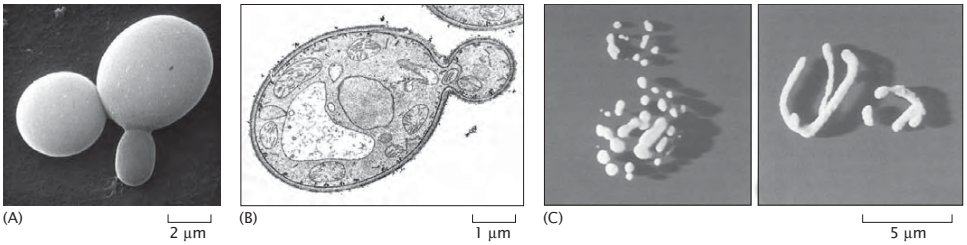


Figure 2.10: Microscopy images of a yeast cell. (A) Scanning electron micrograph of the yeast *Saccharomyces cerevisiae* revealing the overall size scale of these budding yeast. (B) Electron microscopy image of a budding yeast cell. (C) Confocal microscopy images of the mitochondria of *Saccharomyces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

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

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

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

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

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

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

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

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

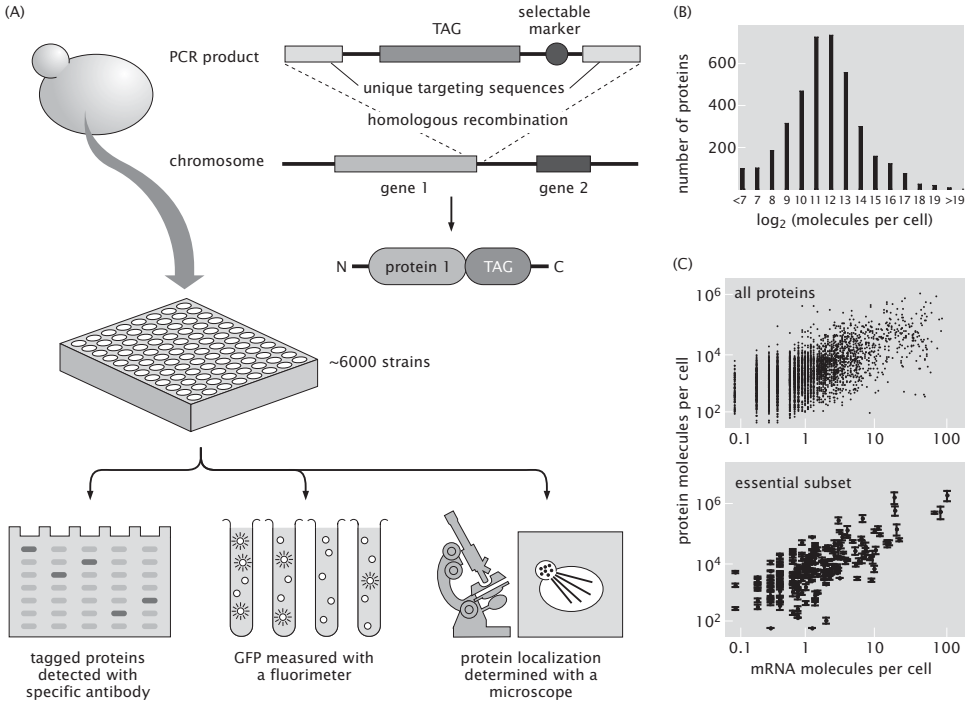


Figure 2.11: Protein copy numbers in yeast. (A) Result of antibody detection of various proteins in yeast showing the number of proteins that have a given copy number. The number of copies of the protein is expressed in powers of 2 as 2^N. (B) The mean number of proteins associated with various processes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

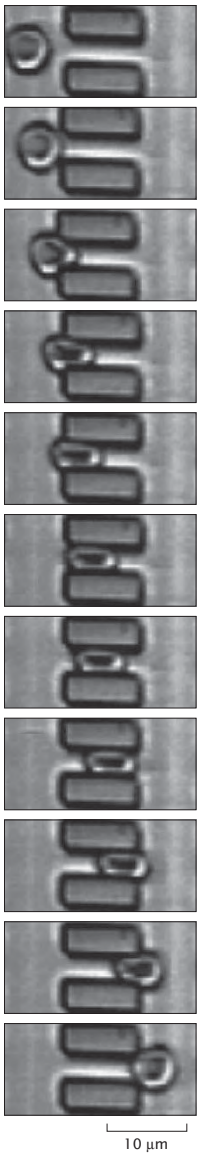


Figure 2.12: Deformability of red blood cells. To measure the deformability of human red blood cells, an array of blocks was fabricated in silicon, each block was $4 \times 4 \times 12$ microns. The blocks were spaced by 4 microns in one direction and 13 microns in the other. A glass coverslip covered 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)

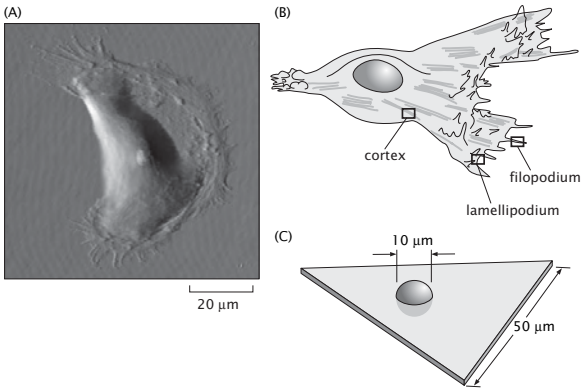


Figure 2.13: Structure of a fibroblast. (A) Atomic-force microscopy image of a fibroblast (courtesy of Manfred Radmacher). (B) cartoon of the external morphology of a fibroblast, (C) characteristic dimensions of the morphology of a fibroblast.

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

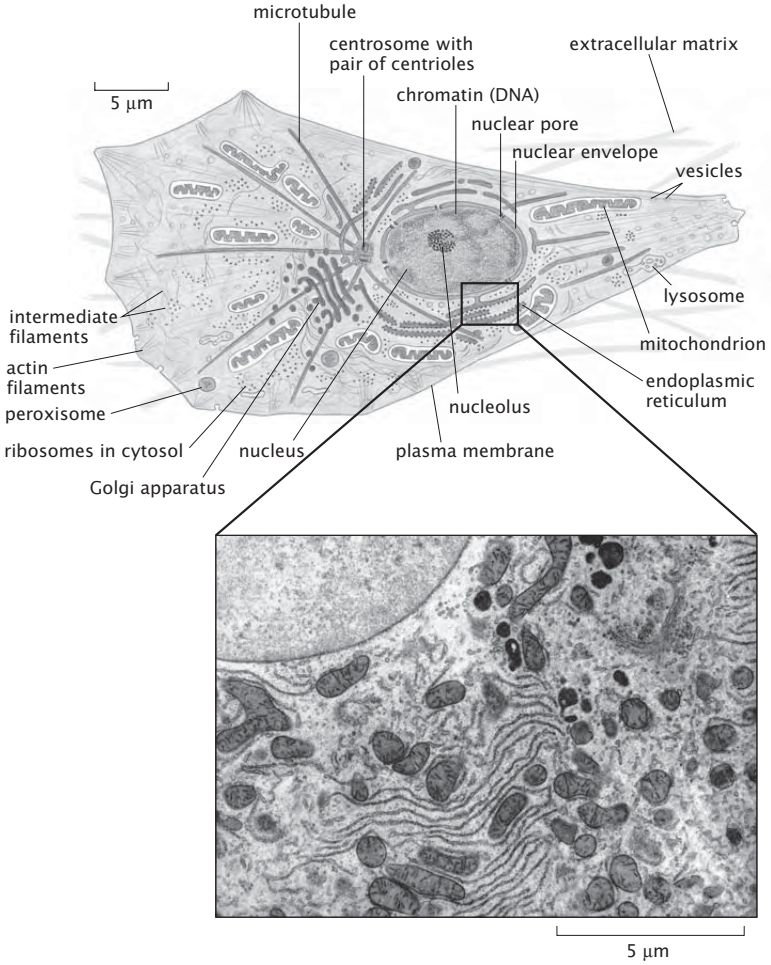


Figure 2.14: Eukaryotic cell and its organelles. The schematic shows a eukaryotic cell and a variety of membrane bound organelles. A thin-section electron microscopy image shows a portion of a rat liver cell approximately equivalent to the boxed area on the schematic. A portion of the nucleus can be 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)

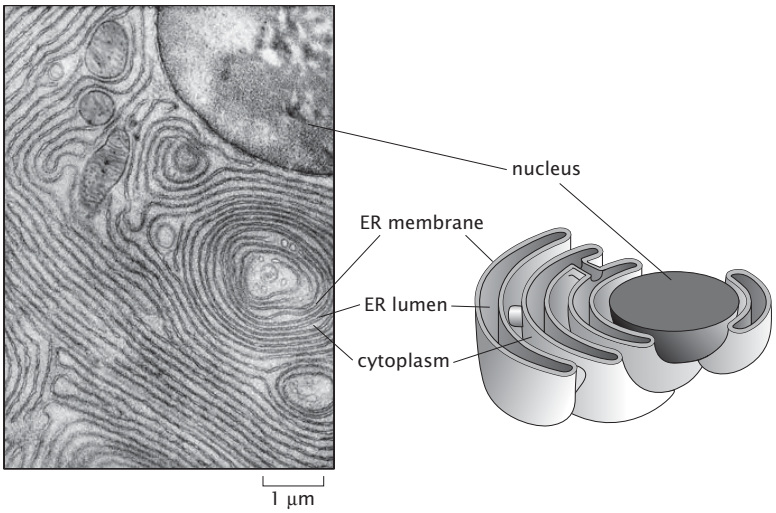


Figure 2.15: Electron micrograph and associated schematic of the endoplasmic reticulum. The left panel shows a thin-section electron micrograph of an acinar cell from the pancreas of a bat. The nucleus is visible at the upper right and the dense and elaborate ER structure is strikingly evident. The right 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

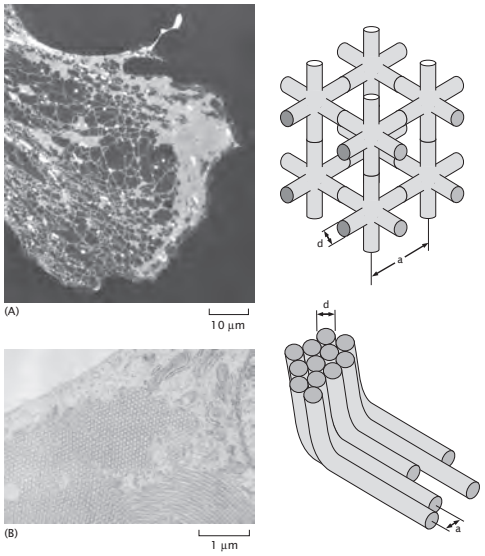


Figure 2.16: Variable morphology of the ER. (A) In most cultured cells, the ER is a combination of a web-like reticular network of tubules and larger flattened cisternae. In this image, a cultured fibroblast was stained with a fluorescent dye called DiOC6 that specifically labels ER membrane. On the 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

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

where A_i is the area of the $i^{\rm th}$ concentric sphere and d is the distance between adjacent cisternae. Since two membranes bound each cisterna the total area of the ER membrane is $A_{\rm ER} = 2 \times \sum_i A_i$. 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:

$$V_{\rm ER} = \frac{4\pi}{3} R^3_{out} - \frac{4\pi}{3} R^3_{nucleus} \ . \tag{2.13}$$

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,

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

Using the values $R_{\rm nucleus} = 5\mu$m, $R_{\rm out} = 10\mu$m and $d = 0.05\mu$m, we get at an estimate $A_{\rm ER} = 15 \times 10^4 \mu{\rm m}^2$. 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 $10^4 \mu{\rm m}^2$ 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

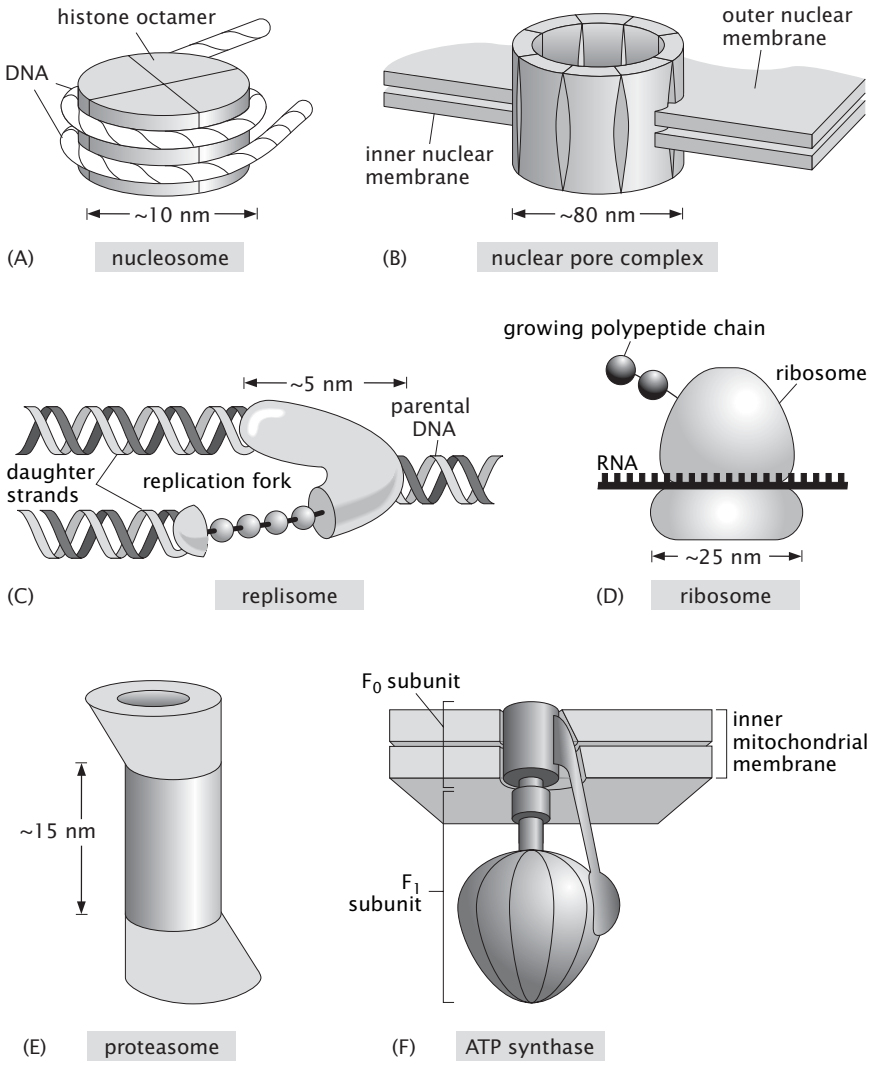


Figure 2.17: The macromolecular assemblies of the cell.

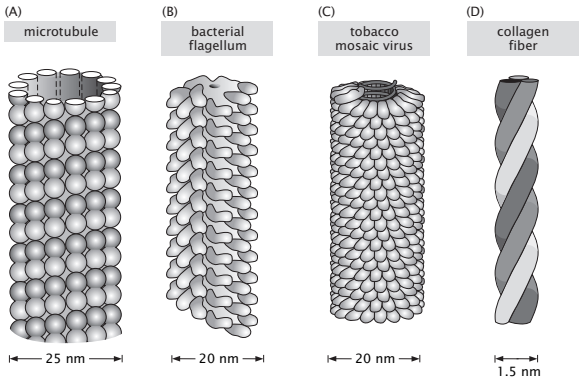


Figure 2.18: Helical assemblies of the cell. Cells have a variety of different helical assemblies, some formed from individual monomeric units (such as (A)-(C)) and others resulting from coils of proteins.

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.

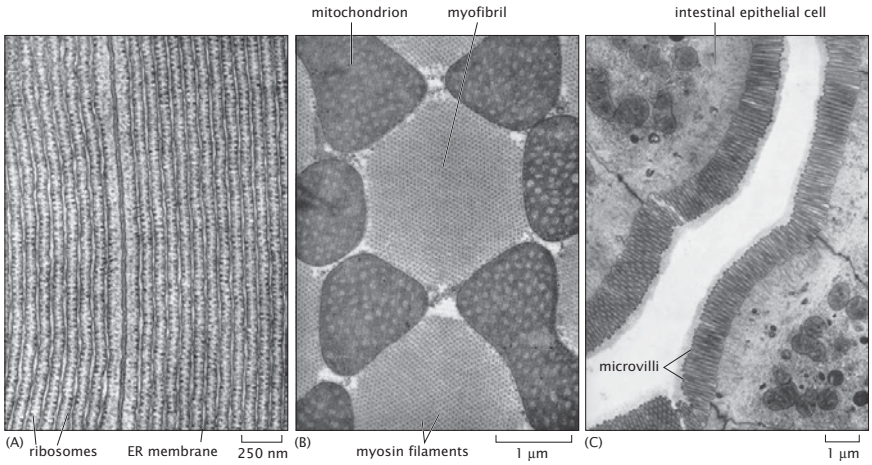


Figure 2.19: Ordered macromolecular assemblies. Collage of examples of macromolecules organized into superstructures. (A) ribosomes on the endoplasmic reticulum, (B) myofibrils in the flight muscle, (C) microvilli at the epithelial surface.

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

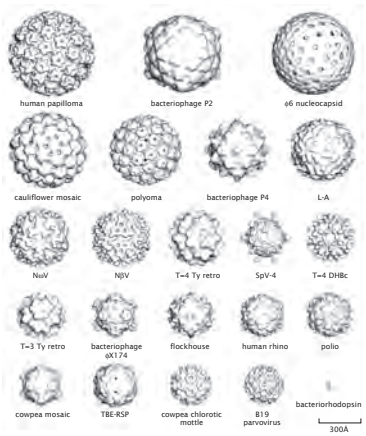


Figure 2.20: Structures of viral capsids. The regularity of the structure of viruses has enabled detailed, atomic-level analysis of their construction patterns. This gallery shows a variety of the different geometries explored by the class of nearly spherical viruses. For size comparison, a large protein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

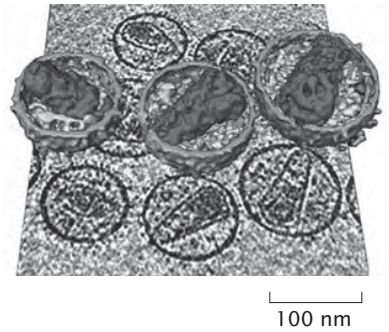


Figure 2.21: Structure of HIV viruses. The planar image shows a single frame from an electron microscopy tilt series. The three-dimensional images show reconstructions of the mature viruses featuring the ice-cream cone shaped capsid on the interior. figure from Briggs, Structure 2006

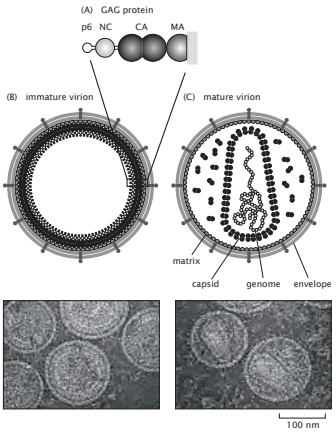


Figure 2.22: HIV architecture. (A) Schematic of the Gag polyprotein, a 41,000 Da architectural building block. (B) Immature virions showing the lipid bilayer coat and the uncut Gag shell on the interior, (C) mature virions in which the Gag protein has been cut by proteases and the separate components 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

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

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.

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

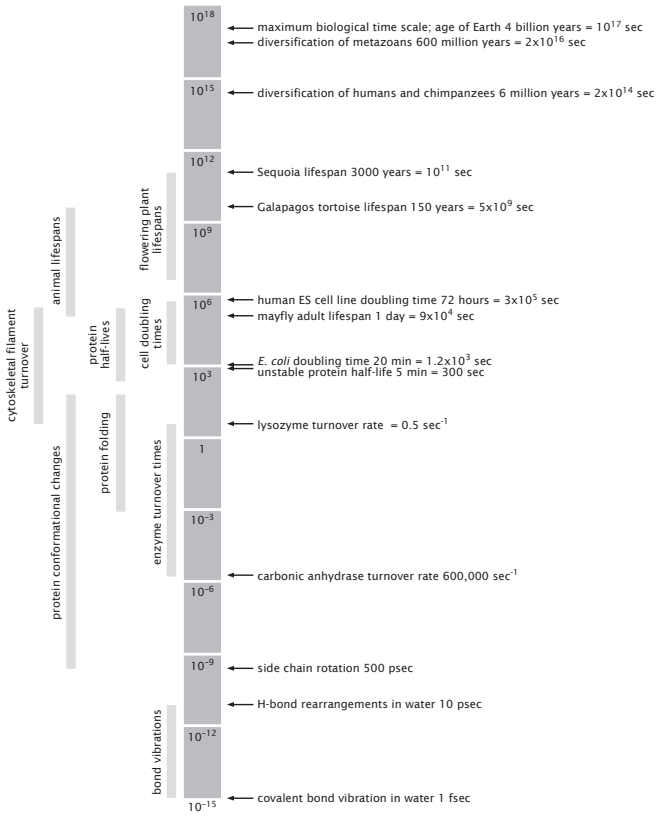


Figure 3.1: Logarithmic scale showing range of times scales associated with various biological processes.

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

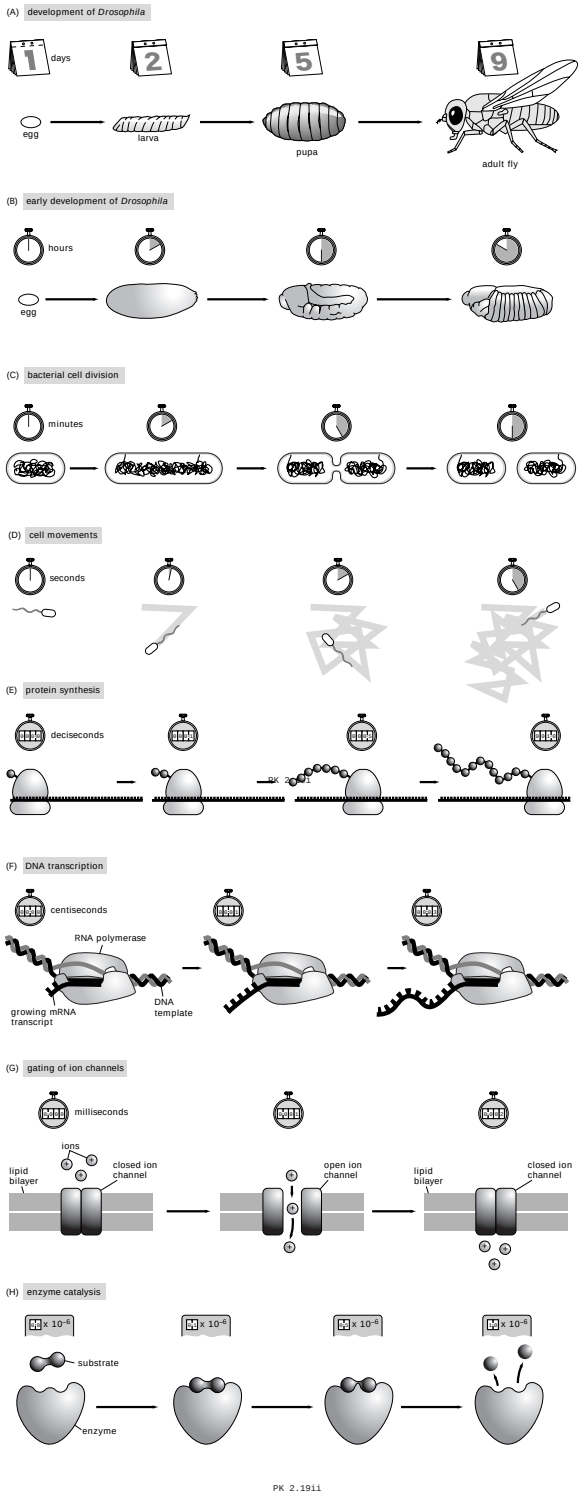


Figure 3.2: Hierarchy of biological time scales. Cartoon showing range of time scales associated with different biological processes.

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.

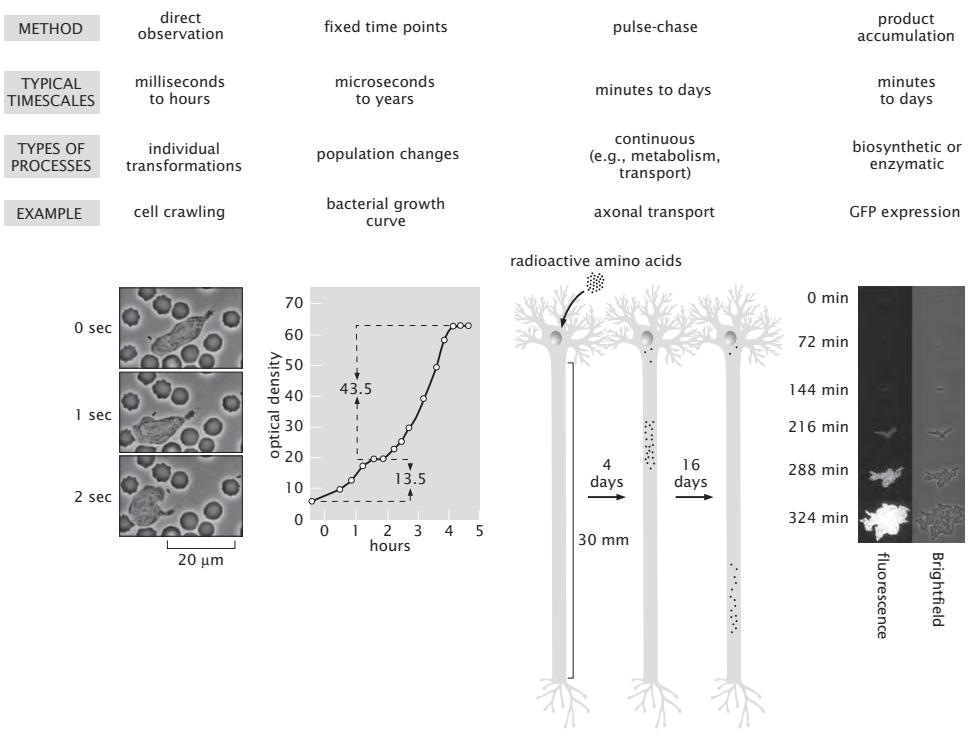


Figure 3.3: Experiments to measure the timing of biological processes. The figure summarizes four strategies for measuring biological rates. For *direct observation* the example shows three frames from a video sequence of a single white blood cell (neutrophil) pursuing a bacterium through a forest of 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

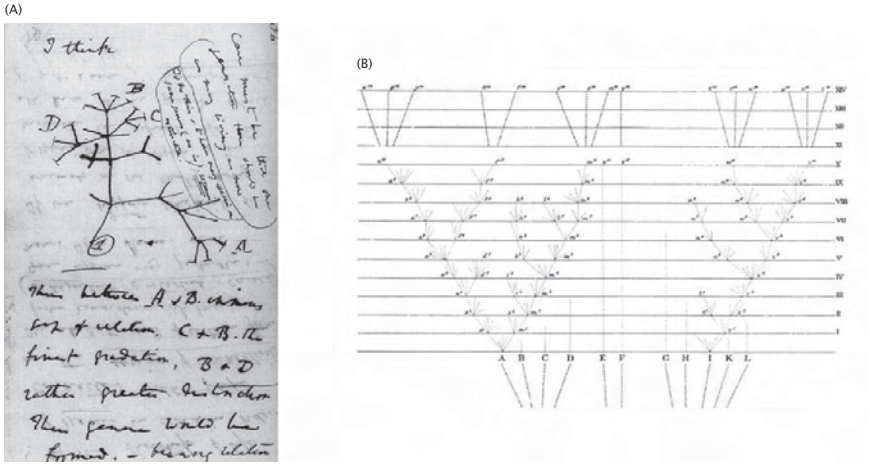


Figure 3.4: Two versions of Darwin's phylogenetic tree. (A) In his notebooks, Darwin drew the first version of what we now recognize as a common schematic demonstrating the relatedness of organisms. He introduced this speculative sketch with the words "I think" as his theory was beginning to take form. (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

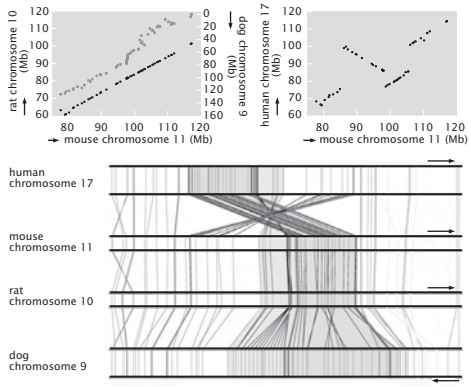
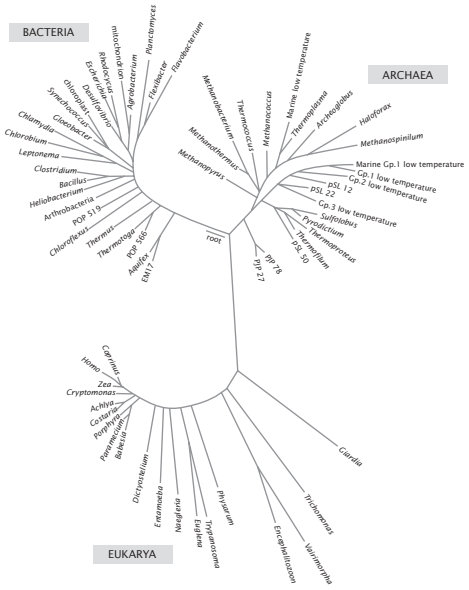


Figure 3.5: Inferring evolutionary relatedness by chromosome alignment. Equivalent regions of four chromosomes from mouse, rat, dog and human were compared to find the location of homologous genes. The graphs at top show the position of each gene in the rat, dog and human sequences as a function of their 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

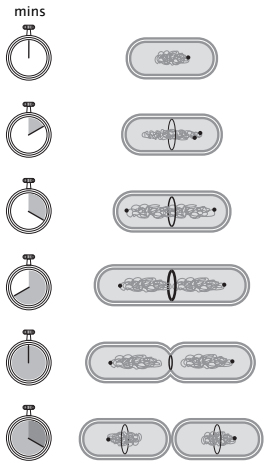


Figure 3.7: Schematic of an idealized bacterial cell cycle. A newborn cell shown at the top has a single chromosome with a single origin of replication marked by the dot. The cell cycle initiates with the duplication of the origin and DNA replication then proceeds in an orderly fashion around the circular 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 ($\approx$ 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*.

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

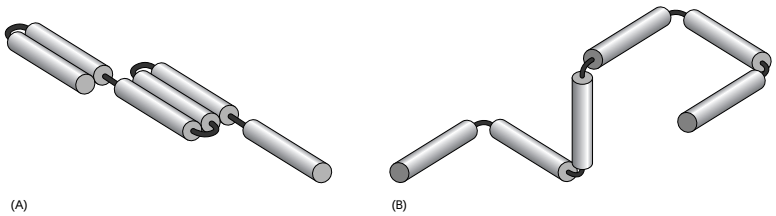


Figure 8.1: Random walk model of polymer. Schematic representation of a (A) one-dimensional random walk and a (B) three-dimensional random walk as an arrangement of linked segments of length a.

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.

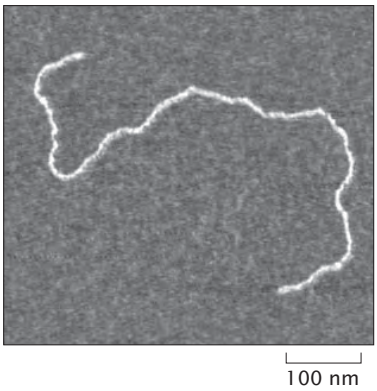


Figure 8.2: Structure of DNA on a surface as seen experimentally using atomic-force microscopy.

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 prob-

abilities, 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

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

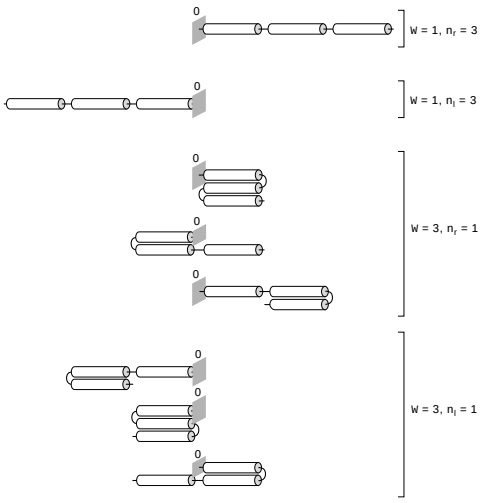


Figure 8.3: Random walk configurations. The schematic shows all of the allowed conformations of a polymer made up of three segments ($2^3 = 8$ conformations) and their corresponding degeneracies.

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

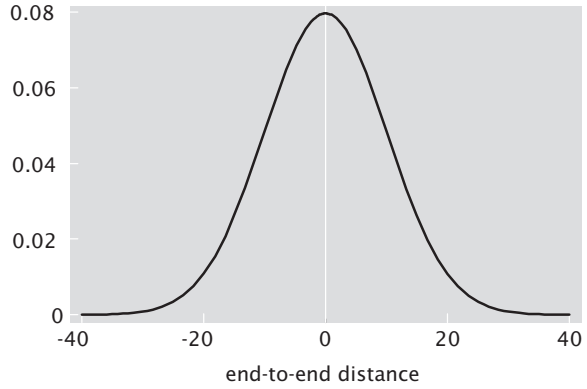


Figure 8.4: End-to-end probability distribution for a one-dimensional "macromolecule" with 100 segments. RP: Fix the figure so that it shows a comparison of the Binomial distribution and the approximate Gaussian for different values of N.

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,

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

In the next step we rewrite the logarithms,

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

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

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

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,

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

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

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

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)

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

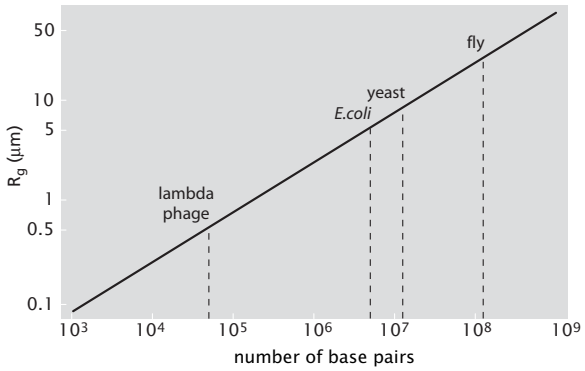


Figure 8.5: Plot of the average size of a DNA molecule in solution as a function of the number of base pairs using the random walk model.

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,

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

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

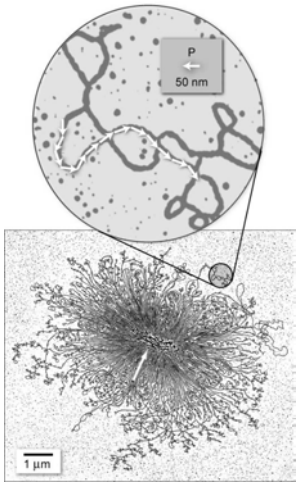


Figure 8.6: Illustration of the spatial extent of a bacterial genome which has escaped the bacterial cell. The expanded region in the figure shows a small segment of the DNA and has a series of arrows on the DNA, each of which have a length equal to the persistence length in order to give a sense of 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, $L \approx 4.6 \times 10^6 \times 3.4\ \mathring{A} \approx 1.5 \times 10^7\ \mathring{A} \approx 1600\ \mu m$. 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

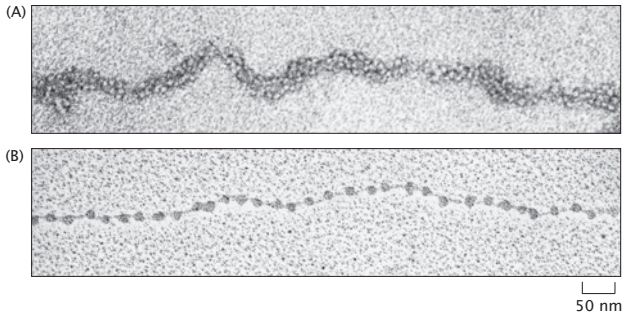


Figure 8.7: Chromatin under the electron microscope. (A) Chromatin extracted from an interphase nucleus appears as a 30nm thick fiber. (B) The 10nm fiber structure shows individual nucleosomes.

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 $c^* = N_G/(4\pi/3R_g^3)$

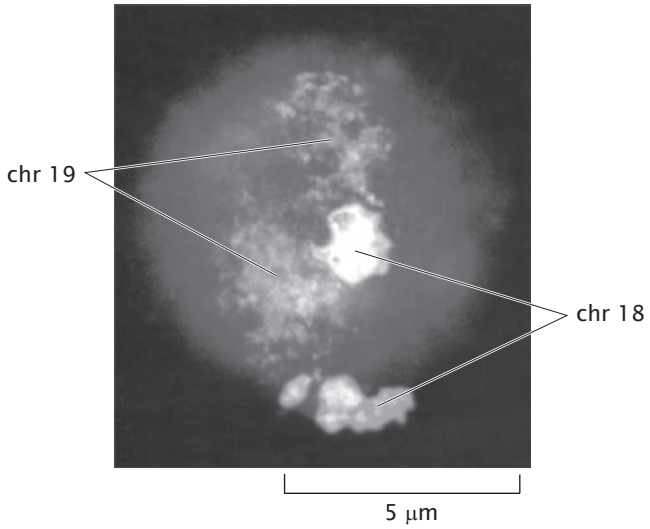


Figure 8.8: Fluorescently stained chromosomes 18 and 19 in a human cell. The chromosomes assume separate territories within the nucleus.

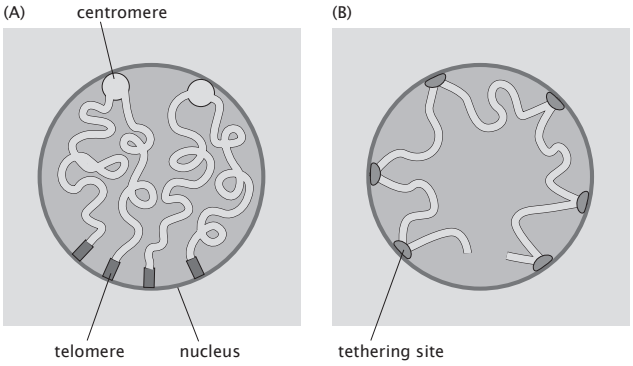


Figure 8.9: Cartoon representation of possible tethering scenarios of interphase chromosomes. The left panel shows tethering at the centromere and the two telomeres at the nuclear periphery. The right panel shows tethering at intermediate locations.

where N_G is the chromosome size in base pairs, and $R_{\rm g}$ 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, $\xi_{\rm P} = 30$nm, the estimate for the radius of gyration is, $R_{\rm g} = 0.97\mu$m. This then leads to a density for an "free" chromosome of $c^* = 750{\rm kb}/(4\pi/3 \times (0.97\ \mu{\rm m})^3) \approx 200\ {\rm kb}/\mu{\rm m}^3$ 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 $c^* \approx 500\ {\rm kb}/\mu{\rm m}^3$. 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,

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

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

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

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 =

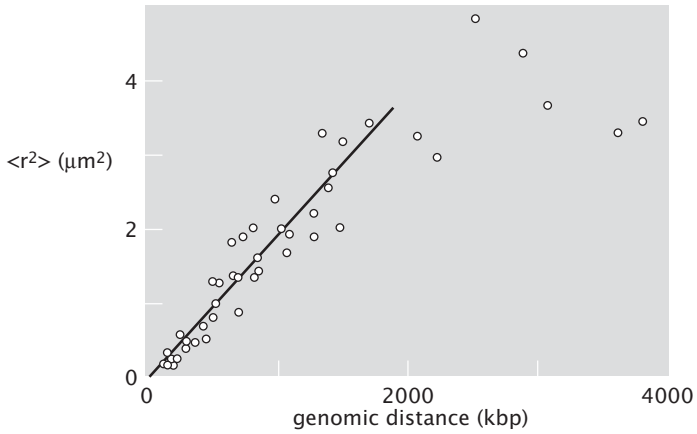


Figure 8.10: Physical distance between two fluorescently labeled loci on human chromosome four as a function of the genomic distance. The physical distance is measured in terms of the average squared distance between the two labels.

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

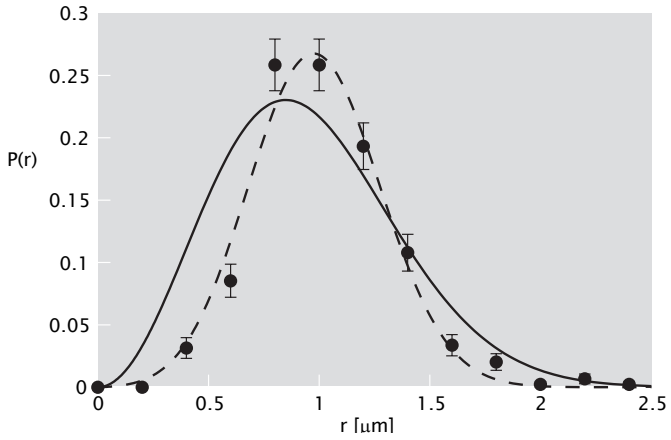


Figure 8.11: Statistics of yeast chromosome III. Distribution of distances between two fluorescent tags placed in proximity of the cetromere and the HML region on yeast chromosome III. These two regions are separated by approximately 100kb in genomic distance.

distance vector $\mathbf{r}$ 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 $\mathbf{r}$. 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 $R \approx 0.9\mu\text{m}$,

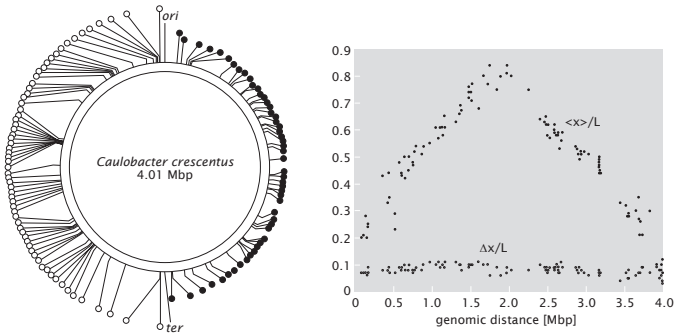


Figure 8.12: Chromosome geography in *Caulobacter crescentus*. Average positions (x/L) and the standard deviation $(\Delta x/L)$ of the position along the long axis of the cell, for 112 different fluorescently tagged locations along the chromosome of *C.crescentus*. The locations of the fluorescent tags 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 0.2μm (cell length $L \approx 2\mu$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 $(0.2\mu\text{m})^2 = Na^2/3 = N_G/3(100/3)\text{nm}^2/\text{bp}$, $N_G \approx 4$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 $Na^2 = 0.16\mu\text{m}^2$. 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 $16 \times 0.1\mu\text{m} \times 3\text{bp/nm} = 4.8$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

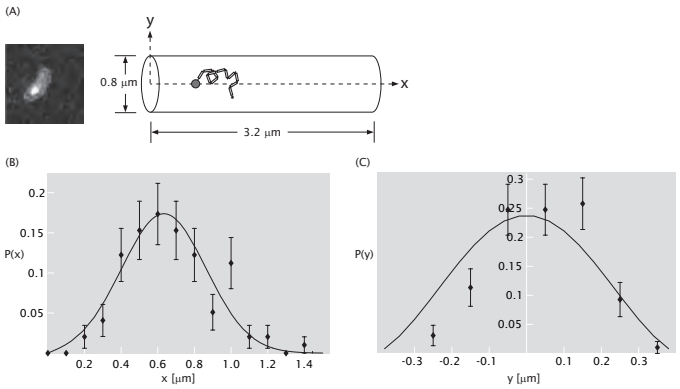


Figure 8.13: Chromosome position distributions *in vivo*. (A) The position of the fluorescently tagged origin of replication on the larger of the two *V.cholerae* chromosomes, is measured along the long axis of the cell (x-direction) and perpendicular to it (y-direction). The cell can be modeled as a 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

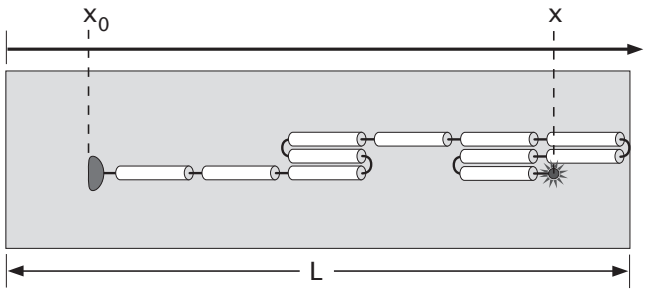


Figure 8.14: Simplified one-dimensional model of a chromosome confined to a cell of size L and tethered at position x_0. The model makes a prediction for the distribution of distances to the fluorescent marker $P(x)$.

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*

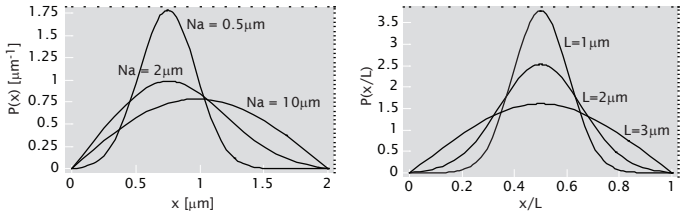


Figure 8.15: A. The distribution of distances to the fluorescent marker for the one-dimensional chromosome model for different contour lengths of the chromatin fiber between the tethering point (at $x_0 = 0.75\mu$m) and the fluorescent marker. The cell size is $L = 2\mu$m, and the packing density and 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.

The complete chapter can be found on the School web site.