Chapter 8

Random Walks and the Structure of Macromolecules

"The journey of a thousand miles begins with a single step." - Lao Tsu from "Tao Te Ching"

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

There are many different ways of characterizing biological structures. A useful alternative to the deterministic description of structure in terms of well defined atomic coordinates is the use of statistical descriptions. 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 macro-

molecular 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

In the context of polymeric systems such as DNA, the notion of structure 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

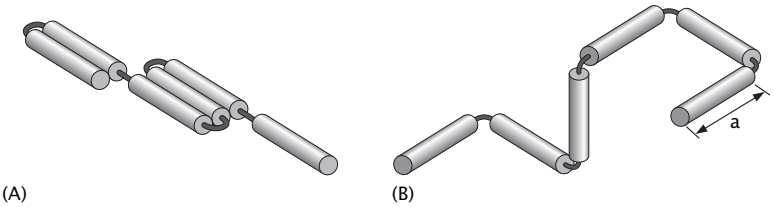

Figure 8.1: Random walk model of polymer. Schematic representation of a (A) one-dimensional random walk and a (B) three-dimensional random walk as an arrangement of linked segments of length a.

at the level of characterizing a blob of a given size and shape. Indeed, these considerations bring us immediately to the examination of statistical measures of structure. As hinted at in the title to this section, one 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. In the remainder of the chapter we show the calculable consequences of statistical descriptions 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. The one- and three-dimensional versions of this idea are shown in fig. 8.1. In the one-dimensional case the segments are at ± 180 degrees with respect to each other. We draw them as non-overlapping for clarity. For the three-dimensional case, we illustrate the situation 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

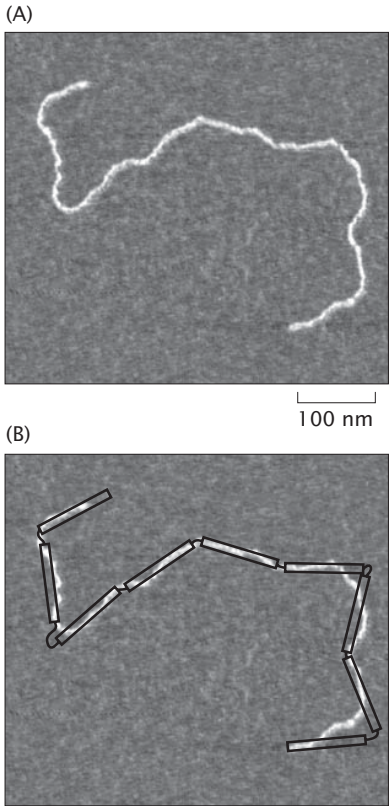

Figure 8.2: DNA as a random walk. (A) Structure of DNA on a surface as seen experimentally using atomic-force microscopy. (B) Representation of the DNA on a surface as a random walk. (Adapted from P. A. Wiggins *et al.*, *Nat. Nanotech.*, 1:37, 2006.)

in settings ranging from estimating the entropic cost of confining DNA to a bacterial cell, to the stretching of DNA by laser tweezers.

8.2.1 A Mathematical Stupor

In Random Walk Models of Polymers, Every Macromolecular Configuration Is Equally Probable

In this section we work our way up by degrees to some of the full beauty and depth of the random walk model. The aim of the analysis is to obtain a probability distribution for each and every macromolecular configuration and to use these probabilities to compute properties of the macromolecule that can be observed experimentally, such as the mean size of the macromolecule and

the free energy required to deform that molecule. Our starting point will be an analysis of the random walk in one-dimension, with our discussion being guided by the ways in which we will later generalize these ideas and apply them in what might at first be considered unexpected settings.

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

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

Given the spectrum of possible configurations and their corresponding probabilities, one of the most immediate questions we can pose concerns the mean distance of the walker from its point of departure as a function of the number of segments in the chain. In the context of biology, this question is tied to problems such as the cyclization of DNA, the likelihood that a tethered ligand and receptor will find each other and to the gross structure of plasmids and chromosomal DNA in cells. To find the end-to-end distance for the molecule of interest we can use both simple arguments as well as brute force calculation, and we will take up both of these options in turn. The simple argument notes that the expected value of the walker's 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 displacement 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 displacement of 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, and since $x_i = \pm a$, 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 Probability of a Given Macromolecular Configuration Depends Upon its 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. This kind of counting result was derived in the "Math Behind the Models" box on pg. 304. 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

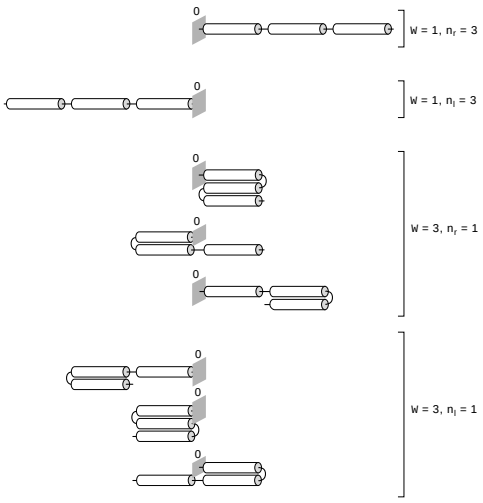

Figure 8.3: Random walk configurations. The schematic shows all of the allowed conformations of a polymer made up of three segments ($2^3 = 8$ conformations) and their corresponding degeneracies.

are poised to write down the probability of an overall departure n_r from the origin which is given by

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

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

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

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

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

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

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

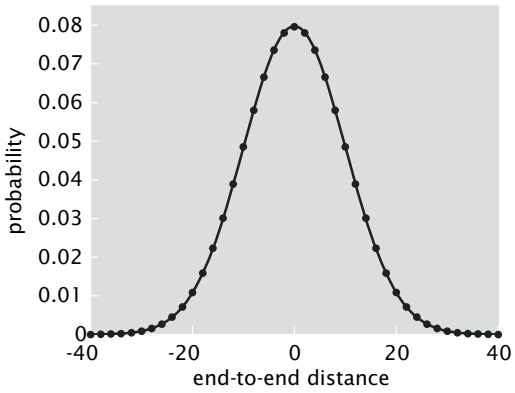

Figure 8.4: End-to-end probability distribution for a one-dimensional "macromolecule" with 100 segments. The figure shows a comparison of the Binomial distribution (dots) given in eqn. 8.10 and the approximate Gaussian distribution (curve) given in eqn. 8.16.

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. 280 and the problems at the end of chap. 5, $\ln n! \approx n \ln n - n + \frac{1}{2}\ln(2\pi n)$ for $n \gg 1$, and the Taylor expansion (pg. 273), $\ln(1 + x) \approx x - x^2/2$ for $x \ll 1$. Note that here we take the first three terms in the Stirling approximation, and keep terms up to x^2 in the Taylor expansion, in anticipation of the fact that the leading term in $\ln p(R; N)$ 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 Stirling 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 a bit of algebra (which is left as an exercise for the reader) we arrive at the formula

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

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

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

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

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

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 identically distributed independent 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, and independent of the allowed angles between subsequent steps of the walk, as long as each step is taken independently of the previous one.

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 the two identities

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

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

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

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,

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

Note that $P(\mathbf{R}; N)$ has units of inverse volume, or concentration, and has the nice intuitive interpretation as the concentration of one end of the random-walk polymer at position $\mathbf{R}$, with respect to the other end. Furthermore, $P(\mathbf{R}; N)$ is sharply peaked at $\mathbf{R} = 0$, and this property underlies the elasticity 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 a 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.

- **Estimate: End-to-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 circular DNA associated with an *E. coli* cell is roughly 5 million basepairs long. An open DNA chain of the same size can be modeled as a random walk of roughly $N = 15000$ steps since the Kuhn length (the length of the "rigid" segments in the chain model) 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 can be estimated from eqn. 8.15 and 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}. As discussed above, this overwhelming probability that R is close to zero is responsible for the elastic response of polymer chains due to an applied load.

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 $\langle \mathbf{t}(s) \cdot \mathbf{t}(u) \rangle$, where $\mathbf{t}(s)$ is the tangent vector evaluated at 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

$$\langle \mathbf{t}(s) \cdot \mathbf{t}(u) \rangle = e^{-\frac{|s-u|}{\xi_p}} \quad . \tag{8.24}$$

In chap. 10 we derive this equation in the context of a model where the polymer is thought of as a long and thin elastic beam. Furthermore, we note that eqn. 8.24 is not universally valid. For example if the tangents are kept fixed and equal at the ends of the polymer, say by laser tweezers, then $\langle \mathbf{t}(0) \cdot \mathbf{t}(s) \rangle$ will decay at first, but as s approaches the contour length of the polymer L it will necessarily increase, since $\mathbf{t}(0) \cdot \mathbf{t}(L) = 1$. Other constraints on the polymer, such as confinement by the cell wall, will also lead to deviations from eqn. 8.24. Still, for small enough separations $|s-u|$ the exponential law is expected to hold.

A good example of a long flexible polymer is provided by genomic DNA of viruses such as λ-phage with a contour length of 16.6 μm. 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 $\mathbf{t}(s)$,

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

As a result, we can write

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

where $\langle \cdots \rangle$ is an average over all polymer configurations. Using the average of

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

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

To obtain this result we made a change of variables $x = u - s$ in the second integral and then replaced the upper bound of integration $L - s$ by ∞, which is justified in the $L \gg \xi_p$ limit. Comparing the above formula to the result that follows from the random walk model, eqn. 8.4, $\langle \mathbf{R}^2 \rangle = aL$, we see that Kuhn length a is twice the persistence length, $a = 2\xi_p$. In rewriting the random walk result we made use of the relation between the length of the walk and the number of Kuhn segments, $L = Na$. We are now prepared to make estimates of the physical size of genomes in solution.

8.2.2 How Big is a Genome?

A simple estimate of the size of a polymer in solution can be obtained using the end-to-end distance,

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

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

$$\langle R_G^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle (\mathbf{R}_i - \mathbf{R}_{CM})^2 \rangle. \tag{8.30}$$

Roughly speaking it measures the average distance between the monomers and the center of mass of the polymer. The center of mass is defined as

$$\mathbf{R}_{CM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{R}_i. \tag{8.31}$$

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)

$$\sqrt{\langle R_G^2 \rangle} = \sqrt{\frac{L\xi_p}{3}}. \tag{8.32}$$

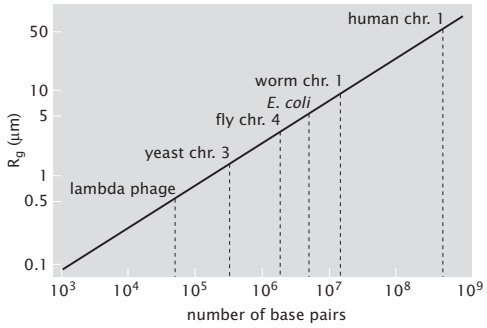

Figure 8.5: Size of genomic DNA in solution. Plot of the average size of a DNA molecule in solution as a function of the number of base pairs using the random walk model. The labels correspond to particular chromosomes from viruses, bacteria, yeast, flies, worms and humans.

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 $L \approx 0.34 \text{ nm} \times N_{bp}$, and hence,

$$\sqrt{\langle R_G^2 \rangle} \approx \frac{1}{3}\sqrt{N_{bp}\xi_p} \text{ nm}. \tag{8.33}$$

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. 1.13 (pg. 48) 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. 8.32 to formulate an estimate of the size of the DNA. Consider fig. 1.13 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. Recalling that the persistence length is $\xi_p \approx 50$nm, eqn. 8.33 tells us that the mean size of the DNA seen in fig. 1.13 is $\sqrt{\langle R_G^2 \rangle} = 1/3\sqrt{150 \times 10^3 \times 50}$ nm ≈ 0.9 μm. This result is

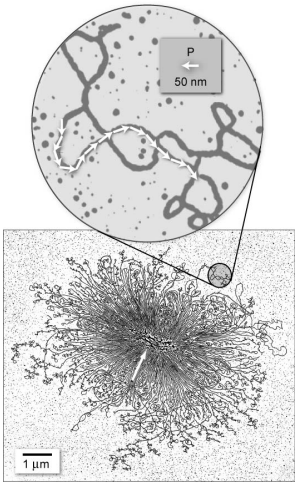

Figure 8.6: Illustration of the spatial extent of a bacterial genome which has escaped the bacterial cell. The expanded region in the figure shows a small segment of the DNA and has a series of arrows on the DNA, each of which have a length equal to the persistence length in order to give a sense of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

comparable to though larger than the length scale of the exploded DNA seen in fig. 1.13. Given the crudeness of the model and probably more importantly, the fact that the DNA seems to be constrained via links 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, $N_{bp} \approx 4.6 \times 10^6$ base pairs. Once again invoking eqn. 8.33 tells us that the mean size of the DNA seen in fig. 8.6 is $\sqrt{\langle R_G^2 \rangle} \approx 5\ \mu m$. 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 linear density of chromatin ν, which specifies the number of base pairs of DNA in a nanometer of chromatin fiber. For the 30 nm-fiber, shown in fig. 8.7(A), $\nu \approx 100$ bp/nm, while for the 10 nm-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 10 nm-fiber. We see that there are on average 2 nucleosomes for every 50 nm of fiber. We assume there are 200 bp per nucleosome (150 bp wound around the histones plus 50 bp of linker DNA) therefore $\nu \approx 2 \times 200$ bp/50 nm $= 8$ bp/nm. For comparison, for metaphase chromosomes $\nu \approx 30,000$ bp/nm.

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

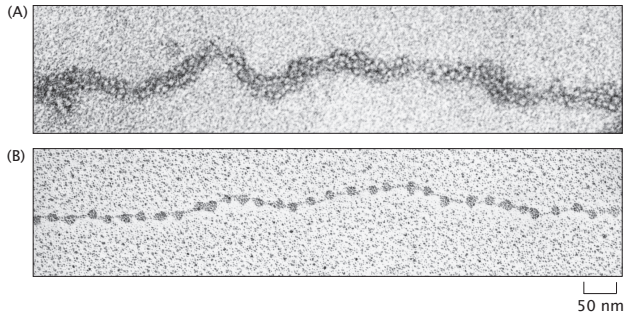

Figure 8.7: Electron microscopy images of chromatin. (A) Chromatin extracted from an interphase nucleus appears as a 30 nm thick fiber. (B) The 10 nm fiber structure shows individual nucleosomes. (Adapted from B. Alberts *et al.*, Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

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 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, but 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 observed nonetheless, 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. Tethering scenarios posit that molecules have particular physical locations because they are held there by tethering molecules. Possible tethering scenarios are shown in fig. 8.9.

- **Estimate: Chromosome Packing in the Yeast Nucleus.** Using polymer physics, here we examine the question of whether chromosomes in yeast are more likely to resemble spaghetti mixed in a bowl, or segre-

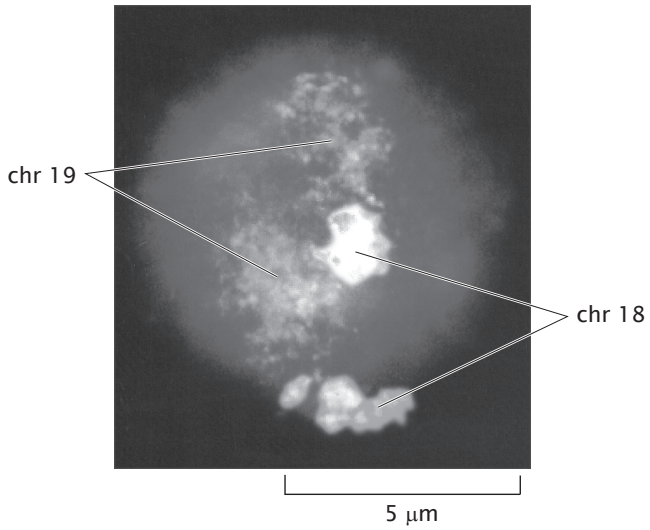

Figure 8.8: Fluorescently stained chromosomes 18 and 19 in a human cell. The chromosomes assume separate territories within the nucleus. (Adapted from B. Alberts *et al.*, Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

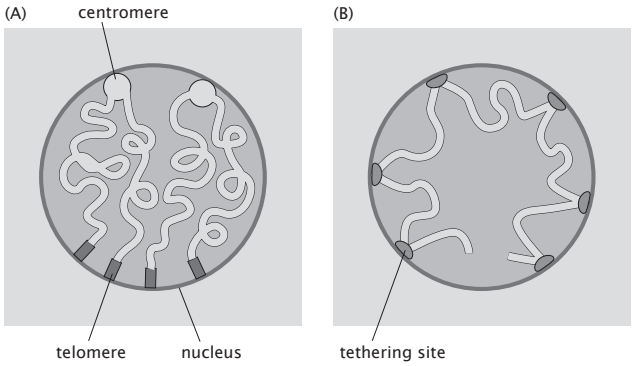

Figure 8.9: Cartoon representation of possible tethering scenarios of interphase chromosomes. (A) Tethering at the centromere and the two telomeres at the nuclear periphery. (B) Tethering at intermediate locations. (Adapted from W. F. Marshall, *Curr. Biol.*, 12:R185, 2002.)

gated blobs not unlike meatballs. The yeast cell has 16 chromosomes in its nucleus. The diameter of the interphase nucleus is about 2 μm. The chromosome size varies between 230 kb to 1500 kb, with a total genome size of 12 Mb. This gives a mean density of $c = 12\ \text{Mb}/(4\pi/3 \times 1\mu\text{m}^3) \approx 3\ \text{Mb}/\mu\text{m}^3$. We now 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 $c^* = N_{bp}/(4\pi/3R_g^3)$ where N_{bp} is the chromosome size in base pairs, and R_g is the radius of gyration of the polymer. If we take an average size of a yeast chromosome to be 12 Mb/16 = 750 kb and a packing density of 8bp/nm the length of this polymer is 750kb/(8bp/nm) = 94 μm. Using the *in vitro* measured value of the persistence length for a 10 nm-fiber, $\xi_P = 30$ nm, the estimate for the radius of gyration is, $R_g = 0.97$ μm. This then leads to a density for a "free" chromosome $c^* = 750\ \text{kb}/(4\pi/3\times(0.97\ \mu\text{m})^3) \approx 200\ \text{kb}/\mu\text{m}^3$, which is about 10 times smaller than the density of chromosomes in the nucleus. The same qualitative conclusion is reached assuming a 30 nm-fiber model for the chromosomes. Namely, using a packing density of 100 bp/nm and the reported persistence length of 200 nm, an average chromosome has a density of $c^* \approx 500\ \text{kb}/\mu\text{m}^3$. This indicates that the chromosomes in the yeast nucleus should typically be found in an entangled melt-like configuration since there is not enough room for them to adopt their preferred configurations without overlap. 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 experimental trick 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. Note that our use of the word "distance" depends upon context; in some cases we will be referring to the scalar distance between two points and in other cases to the displacement vector connecting them. We will pass freely back and forth between these two cases and their relation is explored in the problems at the end of the chapter. In the absence of tethering (or if there is a single tether present) a random walk model of chromatin predicts a Gaussian

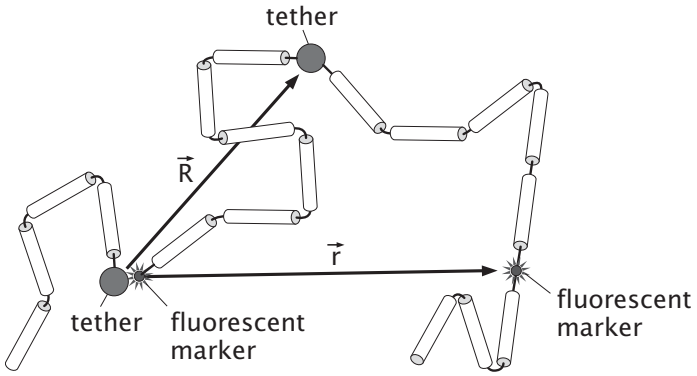

Figure 8.10: Simple configuration of a tethered chromosome. The two tethers are at fixed locations in space, and the second tether is at position $\mathbf{R}$ with respect to the first. The distribution of distances between the two fluorescent markers, one being at the same position on the chromosome as the tether, is a displaced Gaussian.

distribution of distances $\mathbf{r}$ between the two fluorescent markers,

$$P(\mathbf{r}) = \left(\frac{3}{2\pi N a^2}\right)^{3/2} \exp\left(\frac{-3\mathbf{r}^2}{2Na^2}\right) \ . \tag{8.34}$$

Here $a = 2\xi_p$ is the Kuhn or segment length of the polymer and N is the total number of Kuhn segments between the two markers.

The simplest tethering configuration that leads to a distance distribution different than that described above is one with two tethers, as shown in fig. 8.10. One tether is assumed to coincide with the location of one of the two fluorescent markers, and the other tether is at a position $\mathbf{R}$ between the two markers. This configuration of markers and tethers leads to a displaced Gaussian distribution of distances $\mathbf{r}$ between the markers,

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

where N' is now the number of Kuhn segments between the second tether and the second marker. This formula follows simply from eqn. 8.34 when applied to the distribution of distances $(\mathbf{r} - \mathbf{R})$ between the second tether and the second marker. It is interesting to note that mathematical properties of Gaussian distributions, like the one that says that a convolution of two Gaussian distributions is a Gaussian distribution, dictate that *any* tethering configuration will result in a displaced Gaussian distribution of distances.

The implicit assumption we have made in writing down eqns. 8.34 and 8.35 is that chromosomes configurations can be described by random walks. In light of

the dense packing of chromosomes in cells this might seem like an overly zealous use of a simple physical model. However, as we demonstrate using several examples later in this chapter, this model captures key features of experimental data on chromosomes and, more importantly, it makes falsifiable predictions suggesting new directions for experimentation. As a result, this model is a good starting point for quantitative investigations of chromosome geography. This idea is further bolstered by the Flory theorem which states that for dense polymer systems, such as chromosomes confined to cells, polymer configurations are described by random walks.

The contour length of the chromosome between the two tagged loci, Na, can be expressed in terms of the genomic distance between the two fluorescent markers as $Na = N_{bp}/\nu$, where ν is the linear packing density of DNA in chromatin. For example, two genomic loci $N_{bp} = 100$ kb apart would be separated by a 30 nm fiber which is 100 kb/100 bp/nm = 1 μm in contour length. Assuming that the chromatin structure is that of a 10 nm 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 $Na = N_{bp}/\nu$, the mean end-to-end distance squared can also be written as $\langle R^2 \rangle = N_{bp}a/\nu$. 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 labeled 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. 8.11. The fit to the data yields an estimate of $a/\nu = 2$ nm^2/bp, which is nothing but the initial slope of the linear portion of the data. The fact that the data levels off at large genomic distance can be attributed to the effect of chromosome confinement within the cell nucleus. Below we analyze this confining effect using a random walk model for chromosome configurations in the bacterium *V. cholerae*.

With a measurement of the chromatin material parameter a/ν in hand, we can compute the expected probability distribution of distances between fluorescently tagged loci on the chromosome. Typically, due to random orientations of cells in the microscope, experiments with tagged chromosomes only yield information about the magnitude r of the distance vector $\mathbf{r}$ between the two marked spots on the chromosome. Probability distributions for this quantity follow from eqns. 8.34 and 8.35 by integrating out the angular variables θ and ϕ associated with the vector $\mathbf{r}$. This procedure yields

$$P(r) = \left(\frac{3}{2\pi N a^2}\right)^{3/2} 4\pi r^2 \exp\left(\frac{-3r^2}{2Na^2}\right) , \tag{8.36}$$

for the free-polymer case, and

$$P(r) = \left(\frac{3}{4\pi N a^2}\right)^{1/2} \frac{r}{R}\left[\exp\left(\frac{-3(r-R)^2}{2N'a^2}\right) - \exp\left(\frac{-3(r+R)^2}{2N'a^2}\right)\right] \tag{8.37}$$

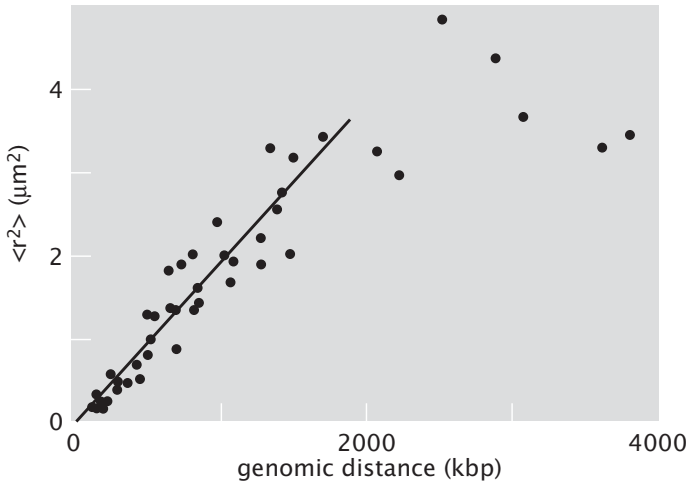

Figure 8.11: Physical distance between two fluorescently labeled loci on human chromosome four as a function of the genomic distance. The physical distance is measured in terms of the average squared distance between the two labels. (Adapted from G. van den Engh *et al.*, *Science*, 257:1410, 1992.)

when the polymer is tethered. Note that that tethering gives a different functional form for the distribution of distances. This provides us with a mathematical tool with which to detect tethering of chromosomes in cells.

Measurement of the distribution of distances between tagged regions on yeast chromosome III suggests that this difference in distributions can be observed *in vivo*. Namely, in fig. 8.12 we show the distance distribution measured between two fluorescent tags, one placed near the HML region of chromosome III of budding yeast and the other on the spindle pole body, which is at a fixed location on the nuclear periphery and essentially marks the location of the centromere. The measured distribution is poorly fitted by the free-polymer formula, eqn. 8.36, while the tethered polymer formula, eqn. 8.37 does the job nicely.

The fit to the tethered-polymer distribution yields two quantities that characterize the model, the mean squared distance between the tether and the fluorescent marker at HML, $N'a^2 = 0.5\ \mu\text{m}^2$, and $R \approx 0.9\ \mu\text{m}$, the distance from the spindle pole body to the tethering point. Note that in order to compute the genomic location of the putative tethering point we need the quantity a/ν which characterizes chromatin structure. For that, measurements like the ones leading to fig. 8.11 for human chromosome four are needed.

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

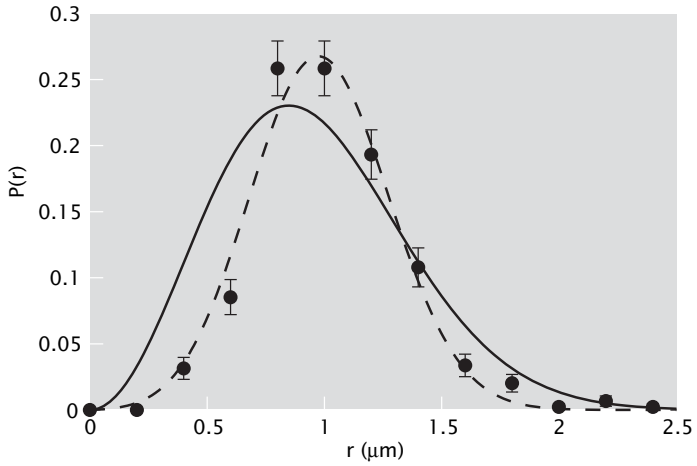

Figure 8.12: Statistics of yeast chromosome III. Distribution of distances between two fluorescent tags placed in proximity of the centromere and the HML region on yeast chromosome III. These two regions are separated by approximately 100 kb in genomic distance. The full line is a fit to the free-polymer distance distribution, eqn. 8.36, while the dashed line is a fit to the tethered-polymer formula, eqn. 8.37. (Courtesy of S. Gordon-Messer, J. Haber and D. Bressan.)

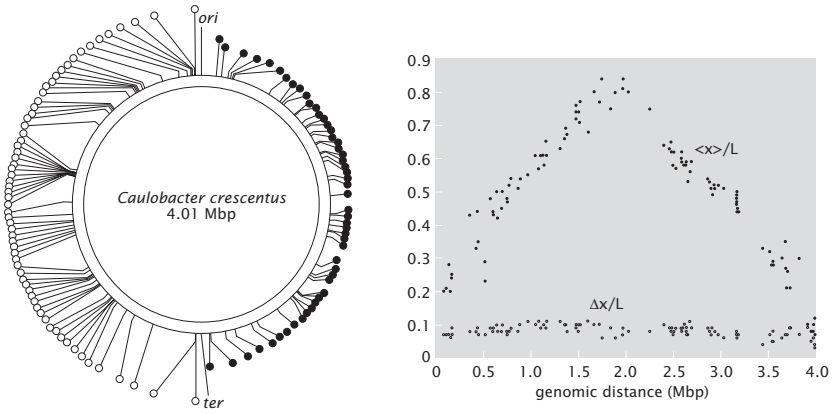

Figure 8.13: Chromosome geography in *Caulobacter crescentus*. Average positions (x/L) and the standard deviation ($\Delta x/L$) of the position along the long axis of the cell, for 112 different fluorescently tagged locations along the chromosome of *C.crescentus*. The locations of the fluorescent tags are shown on the diagram. (Adapted from P. H. Viollier *et al.*, *Proc. Nat. Acad. Sci.*, 101:9257, 2004.)

fluorescent markers placed at different genomic locations, as shown in fig. 8.13. 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 marker 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 to be expected assuming a simple model of the 4 Mbp 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. 8.13 the standard deviation of the position is independent of genomic distance from the origin of replication, and is approximately 0.2 μm (cell length $L \approx 2$ μm). 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 one example). 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 end-to-end distance, $r^2 = x^2 + y^2 + z^2$, is Na^2, the mean of x^2 is three times less (assuming a spherically symmetric distribution), or $Na^2/3$. Using the relation between genomic distance and the mean distance squared, $Na^2 = N_{bp}a/\nu$, 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 $(0.2\ \mu\text{m})^2 = Na^2/3 = N_{bp}/3(100/3)\ \text{nm}^2/\ \text{bp}$, $N_{bp} \approx 4$ kb, which means that the loop should be 8 kb 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 10 kb in size.

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

Another experiment placed 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, roughly 3 Mb and 1 Mb in size. In this case the position of the fluorescent marker 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. 8.14 for the origin of replication for the larger of the two chromosomes. For comparison, the length of the cell is about 3.2 μm, while its diameter is roughly 0.8 μm.

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 $x_0 = 0.6$ μm, 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, eqn. 8.16,

$$P(x) = \sqrt{\frac{1}{2\pi Na^2}} e^{-(x-x_0)^2/Na^2} \tag{8.38}$$

we extract the parameter $Na^2 = 0.16\ \mu\text{m}^2$. Assuming once again the Kuhn length of bare DNA, $a = 0.1$ μm, we conclude that the number of Kuhn segments between the fluorescent marker and the tethering point at $x_0 = 0.6$ μm, is $N = 16$. Taking $\nu = 3\ bp/nm$ 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 5 kb 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 used to collect distance data from cells whose orientation 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 as shown in fig. 8.15. Our goal is 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 in which polymer configurations are identified with trajectories of a

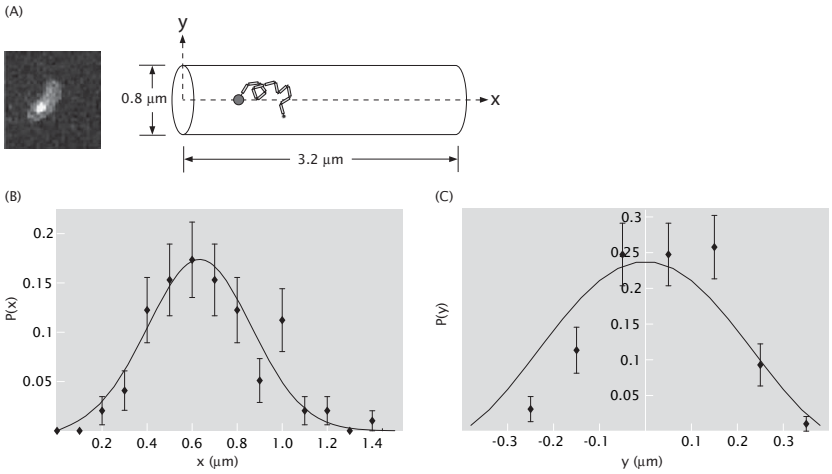

Figure 8.14: Chromosome position distributions *in vivo*. (A) The position of the fluorescently tagged origin of replication on the larger of the two *V.cholerae* chromosomes, is measured along the long axis of the cell (x-direction) and perpendicular to it (y-direction). The cell can be modeled as a 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. $P(x)$ is the Gaussian distribution characteristic of a free random walk polymer, while $P(y)$ is non-Gaussian due to effects of confinement by the cell walls. Calculation of $P(y)$ for a random walk polymer confined to a cylinder is left as a homework assignment. (Courtesy of A. Fiebig and J. Theriot, unpublished data.)

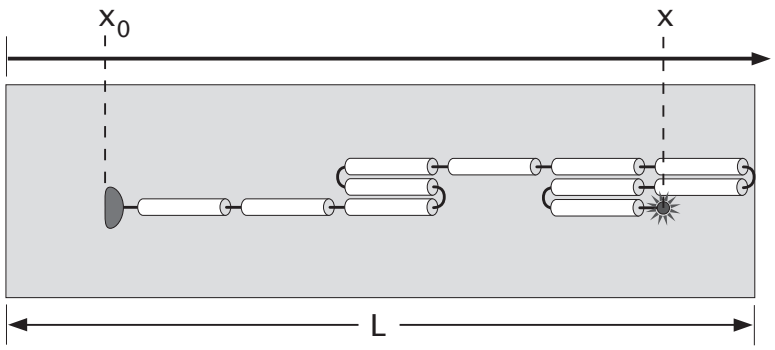

Figure 8.15: Simplified one-dimensional model of a chromosome confined to a cell of size L and tethered at position x_0. The model makes a prediction for the distribution of distances to the fluorescent marker, $P(x)$.

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 cell, we impose absorbing boundary conditions. In other words we demand that $P(x;N)$ vanishes for $x = 0$ and $x = L$ and for any N. This guarantees that any walk that crosses the boundary of the cell 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 cell is then $G(x;N)$. This quantity satisfies the diffusion equation,

$$\frac{\partial G(x;N)}{\partial N} = \frac{a^2}{2}\,\frac{\partial^2 G(x;N)}{\partial x^2}\ . \tag{8.39}$$

This connection between random walks and diffusion leading to the above equation is explored in chap. 13 (see discussion on pg. 685) and in the problems at the end of this chapter.

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

$$P(x;N) = \frac{G(x;N)}{\int_0^L G(x;N)dx}\ . \tag{8.40}$$

Therefore, to obtain the probability distribution $P(x;N)$ we must first solve eqn. 8.39 with boundary conditions $G(0;N) = G(L;N) = 0$ and the initial condition $G(x;0) = \delta(x - x_0)$ which says where the polymer walk begins. The delta function $\delta(x - x_0)$ is sharply peaked at x_0 indicating that the random walker starts at this position.

To solve eqn. 8.39 we expand the function $G(x;N)$ into a Fourier series (see "The Math Behind the Models" on pg. 420),

$$G(x;N) = \sum_{n=1}^{\infty} A_n(N)\sin\left(\frac{n\pi}{L}x\right). \tag{8.41}$$

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

$$\delta(x - x_0) = \sum_{n=1}^{\infty} A_n(0)\sin\left(\frac{n\pi}{L}x\right) \tag{8.42}$$

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

$$\sum_{n=1}^{\infty} A_n(0)\int_0^L \sin\left(\frac{n\pi}{L}x\right)\sin\left(\frac{m\pi}{L}x\right)dx = A_m(0)\frac{L}{2} \tag{8.43}$$

where we have used the orthogonality property of sine functions given by

$$\int_0^L \sin\left(\frac{n\pi}{L}x\right)\sin\left(\frac{m\pi}{L}x\right)dx = \delta_{n,m}\frac{L}{2}\ . \tag{8.44}$$

Putting the results of integration of the left and right hand side of eqn. 8.42 together, we find

$$A_m(0) = \frac{2}{L} \sin\left(\frac{m\pi}{L} x_0\right) \ . \tag{8.45}$$

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. 8.39, is satisfied. To figure this out we simply substitute the Fourier expansion of $G(x;N)$ into the differential equation. This yields

$$\sum_{n=1}^{\infty} \frac{\partial A_n(N)}{\partial N} \sin\left(\frac{n\pi}{L} x\right) = -\frac{a^2}{2} \sum_{n=1}^{\infty} A_n(N) \left(\frac{n\pi}{L}\right)^2 \sin\left(\frac{n\pi}{L} x\right) \ . \tag{8.46}$$

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 $A_m(N)$ given by

$$\frac{\partial A_m(N)}{\partial N} = -\frac{a^2}{2} \left(\frac{m\pi}{L}\right)^2 A_m(N) \ . \tag{8.47}$$

The solution to this equation is an exponential function,

$$A_m(N) = A_m(0) \exp\left(-\left(\frac{m\pi}{L}\right)^2 \frac{a^2}{2} N\right) \ , \tag{8.48}$$

where the coefficient $A_m(0)$ was determined above (eqn. 8.45) from the initial condition.

Finally, the solution to eqn. 8.39 that satisfies the initial condition that all walkers start at x_0 and the absorbing boundary conditions at the cell boundaries, is

$$G(x;N) = \sum_{n=1}^{\infty} \frac{2}{L} \sin\left(\frac{n\pi}{L} x_0\right) \sin\left(\frac{n\pi}{L} x\right) \exp\left(-\left(\frac{n\pi}{L}\right)^2 \frac{a^2}{2} N\right) \ . \tag{8.49}$$

To turn this quantity into the probability distribution for the end-to-end distance of a polymer confined in a cell, we make use of eqn. 8.40, to yield

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

This probability distribution is plotted in fig. 8.16(A) for DNA ($a = 100$ nm) confined to a cell 2 μm in length, for DNA lengths ranging from 0.5 μm to 10 μm. Note that for the shortest chain the confining cell has no effect and the end-to-end distance distribution is a simple Gaussian function, eqn. 8.38. For the intermediate chain length, $Na = 2$ μm, the effect of the cell is to skew the distribution owing to the fact that the tethering point, $x_0 = 0.75$ μm, was chosen closer to the left cell boundary. Finally, for very long DNA lengths the

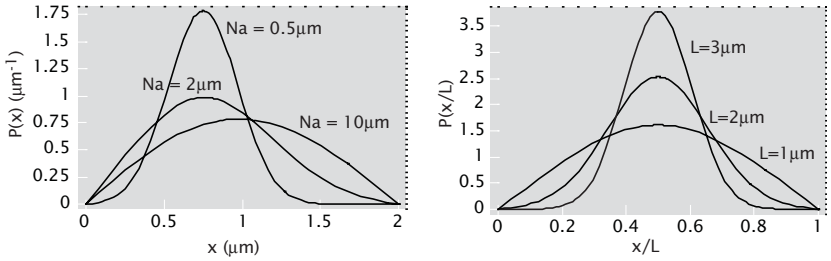

Figure 8.16: Distributions for confined polymers. (A) The distribution of distances to the fluorescent marker for the one-dimensional chromosome model for different contour lengths of the chromatin fiber between the tethering point (at $x_0 = 0.75$ μm) and the fluorescent marker. The cell size is $L = 2$ μm, and the packing density and Kuhn length are that of bare DNA. (B) Same as in A, for a 1 μm long chromatin fiber confined to cells of different size and tethered in the middle of the cell.

distribution is once again symmetric, with all memory of the tethering point lost.

The confined random walk model provides us with the quantitative intuition that allows us to conclude that the observed distribution of average positions of markers along the *C.crescentus* chromosome shown in fig. 8.13 is inconsistent with a model of a polymer confined to the cell interior and tethered only at the poles of the bacterium. This tethering configuration would lead to the average position for most markers (except ones close to the origin and terminus of replication which are thought to be co-localized with the poles) being at the midpoint of the cell. Therefore, further constraints need to be imposed on the chromosome to establish the observed chromosome geography.

In fig. 8.16(B) we once again plot the end-to-end distance distribution using eqn. 8.50, but this time for a DNA molecule that has a length of $Na = 1$ μm ($a = 100$ nm), tethered at the center of the confining box, for box sizes ranging from 1 μm to 3 μm. 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 $y = 0$ (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. 8.14 where we model the cell

as a cylinder and take into account the fact that the experimentally measured y-position of the fluorescent marker is the projection of its radial distance onto the plane of the cover-slip, on which the cells rest. The details of this calculation are left as a homework exercise.

- **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 given by

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n}{L}\, x\right) + b_n \sin\left(\frac{2\pi n}{L}\, x\right) \ . \tag{8.51}$$

Here a_n and b_n are Fourier coefficients, numbers that need to be computed for a given function f and that encode the special features of the function of interest. The above equality is true for all points on the interval with the possible exception of $x = 0$ and $x = L$. 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. In particular, 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. We can restate this mathematically as

$$\begin{aligned}
\int_0^L \sin\left(\frac{2\pi n}{L}\, x\right) \cos\left(\frac{2\pi m}{L}\, x\right) dx &= 0 \\
\int_0^L \sin\left(\frac{2\pi n}{L}\, x\right) \sin\left(\frac{2\pi m}{L}\, x\right) dx &= \delta_{n,m}\, \frac{L}{2} \\
\int_0^L \cos\left(\frac{2\pi n}{L}\, x\right) \cos\left(\frac{2\pi m}{L}\, x\right) dx &= \delta_{n,m}\, \frac{L}{2},
\end{aligned} \tag{8.52}$$

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. 8.51 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

$$\begin{aligned} \int_0^L f(x)dx &= \frac{a_0}{2}L \\ \int_0^L f(x)\cos\left(\frac{2\pi n}{L}\,x\right)dx &= a_n\frac{L}{2} \\ \int_0^L f(x)\sin\left(\frac{2\pi n}{L}\,x\right)dx &= b_n\frac{L}{2} \end{aligned} \tag{8.53}$$

from which we can compute the Fourier coefficients

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

To illustrate the procedure of expanding a function into a Fourier series, 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 $L/2 < x < L$ (i.e. a square wave). Fourier coefficients are computed using eqn. 8.54, 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. 8.17.

8.2.4 DNA Looping: From Chromosomes to Gene Regulation

The organization of genomes occurs at many different scales. A shorter scale phenomenon of widespread significance is the formation of loops of various kinds in both genomic DNA and RNAs as well. Fig. 8.18 shows how nucleic acids form "loops" in a wide variety of different settings. For example, as illustrated in fig. 8.18(A), melting of DNA results in bubbles of single stranded fragments and the meandering of the single-stranded fragments can be evaluated as a problem in random walks. Similar ideas are relevant in evaluating the propensity of RNA to form hairpin loops which are an important element of RNA secondary structure. Another favorite example involves the formation of DNA loops by transcription factors as part of the process of gene regulation. Yet another example shown in fig. 8.18(D) involves genetic recombination in which distant parts of chromosomal DNA find one another as a precursor to the recombination event itself. These events are important in situations ranging from mating type switching in yeast to V(D)J recombination in B cells, to the stochastic decision making that attends olfactory receptor selection.

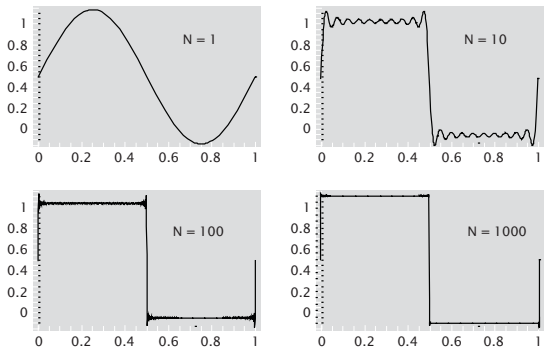

Figure 8.17: Fourier series representation of a square wave. Different graphs correspond to the Fourier series representation of the square wave function where the first N terms have been retained in the sum on the right hand side of eqn. 8.51.

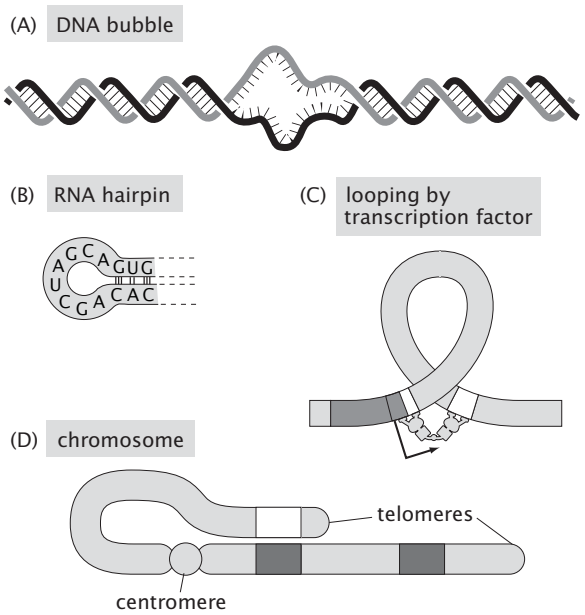

Figure 8.18: Examples of looping. (A) bubble formation in a double-stranded DNA helix, (B) hairpin loop in RNA secondary structure, (C) DNA looping due to a transcription factor, (D) long distance DNA looping of chromosomal DNA.

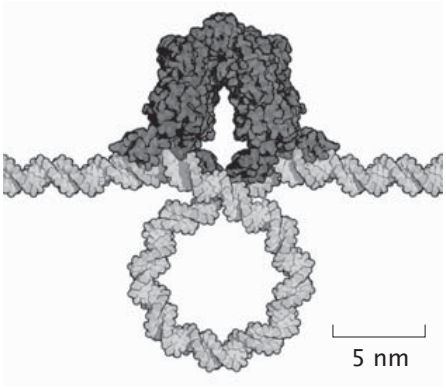

Figure 8.19: Model for DNA loop formation by the Lac repressor. The interface between the protein and the DNA was determined by x-ray crystallography, but the overall position and shape of the DNA in the loop is an artist's rendition. (Courtesy of David Goodsell.)

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

In fig. 4.13 (pg. 200) and section 4.4.3 (pg. 202), we introduced the *lac* operon as a particularly notable example of gene regulation. One part of the *lac* operon story is how the genes of this operon are repressed by the Lac repressor protein as shown in fig. 8.19. Thus far, our description of Lac repressor has been largely schematic without particular reference to the mechanical actions responsible for repression. The actual story of the action of Lac repressor is more complicated than that illustrated in fig. 4.15 (pg. 204). In fact, there are several other operator sites (O_2 and O_3) in addition to the primary operator site (O_1) described there where the repressor can bind resulting in a DNA loop like that shown in fig. 8.19. The effectiveness of repression is highest when the Lac repressor tetramer (built up from four copies of the *lacI* gene) binds to two operators simultaneously.

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

In order for a protein molecule such as the Lac repressor to spontaneously form a loop in the DNA, the DNA and protein must together suffer a fluctuation that brings all of the pieces into physical proximity. As will be shown in chap. 10, for the DNA to bend in this way costs elastic energy. However, there is also a contribution to the free energy of looping from entropy since when the DNA is looped, there are fewer conformations available to the system and hence a reduction in the entropy.

As a warm-up exercise to evaluate the entropic cost of loop formation we

consider a one-dimensional model and examine the fraction of conformations which close on themselves. The probability, $p_\circ$, of loop formation is the probability that the one-dimensional random walker returns to the origin. Using eqn. 8.10 for $R = 0$, we conclude

$$p_\circ = \frac{\text{number of looped configs.}}{\text{total number of configs.}} = \frac{\frac{N!}{(\frac{N}{2})!(\frac{N}{2})!}}{2^N} \tag{8.55}$$

where N is the number of Kuhn segments. Here we are interested in the long chain limit, which corresponds to $N \gg 1$. This is also the limit in which the random walk model can be applied to DNA conformations, as discussed previously. To further simplify eqn. 8.55 we make use of our trusty Stirling formula (pg. 280), $N! \approx (N/e)^N\sqrt{2\pi N}$, which holds for $N \gg 1$ and implies

$$p_\circ \approx \sqrt{\frac{2}{\pi N}}. \tag{8.56}$$

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

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

$$P(R;N) \approx \frac{1}{\sqrt{2\pi Na^2}} \tag{8.57}$$

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

$$p_\circ = \int_{-\delta}^{\delta} \frac{1}{\sqrt{2\pi Na^2}}\, dR = \sqrt{\frac{2}{\pi N}}\frac{\delta}{a} \tag{8.58}$$

which, as expected, is the same as eqn. 8.56 for $\delta = a$.

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

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

$$p_\circ = \int_0^\delta 4\pi R^2 P(R;N) dR = \int_0^\delta 4\pi R^2 \left(\frac{3}{2\pi N a^2}\right)^{\frac{3}{2}} e^{-\frac{3R^2}{2Na^2}} dR \ . \tag{8.59}$$

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

$$p_\circ = \int_0^\delta 4\pi R^2 \left(\frac{3}{2\pi N a^2}\right)^{\frac{3}{2}} dR = \left(\frac{6}{\pi N^3}\right)^{\frac{1}{2}} \left(\frac{\delta}{a}\right)^3 \ . \tag{8.60}$$

The main conclusion that follows from this calculation is that the cyclization probability decays as the number of Kuhn segments of the chain to the power $-3/2$. In section 10.3 (pg. 508), we will finish these arguments by showing how to link the entropic and energetic description of DNA looping. These ideas will then be applied to compute the probability of gene expression in section 19.2.5 (pg. 1028).

8.2.5 PCR, DNA Melting and DNA Bubbles

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

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

The melting process is a competition between entropy, which favors the melted state, and energy, which is minimized when all the bases are paired up and hydrogen bonds are formed between them. As a result, melting can be induced by an increase in temperature, which changes the relative weights of entropy and energy in the DNA free energy, or, for example, by changing salt concentrations which change the energetics of hydrogen bonding. When a cell needs to melt its DNA helix, it does not change the temperature or salt concentration, but rather uses an energy-consuming enzyme called a helicase to pay the energetic penalty of separating the two DNA strands.

The polymerase chain reaction (PCR) has been a revolution within the revolution of molecular biology. PCR permits the amplification of DNA fragments so that these fragments can be used for processes such as cloning genes for expressing insulin in bacteria, finding rare mutations in a population, identifying

the origin of a blood sample at a crime scene and comparing the sequence of human vs. neanderthal. The basic idea is shown schematically in fig. 8.20. The goal of the PCR reaction is to take some fragment of a DNA molecule and make a huge number of copies of it. In fig. 8.20, it is seen that the reaction consists of the template DNA (the piece to be copied), "primers" which are small (approximately 20bp) DNA fragments that are complementary to sites on the DNA adjacent to the region of interest, DNA polymerase which is the molecular xerox machine that makes the copies and a host of nucleotides (the As, Gs, Ts and Cs) that are the raw material for constructing new DNA molecules.

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

DNA Melting Temperatures Can Be Estimated Using a Random Walk Model

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

To examine this competition quantitatively we consider a simplified version of the so-called Poland-Scheraga model where we allow the formation of only one bubble as shown in fig. 8.21. This is a reasonable assumption for a DNA strand of moderate length ($100 - 1000$bp) as the energy penalty for initiating a bubble is considerably larger than for elongating the bubble. For short strands the entropy gained by having more than one bubble will not be enough to overcome this energy penalty for bubble initiation.

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

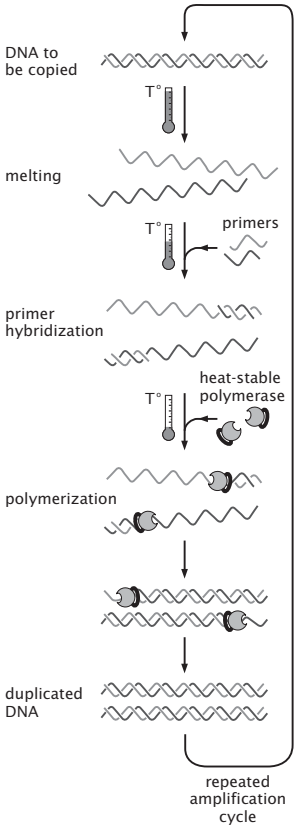

Figure 8.20: Schematic of the polymerase chain reaction (PCR) and its dependence upon DNA melting. The thermometer icons show how the temperature is varied at each step during a cycle of the polymerase chain reaction.

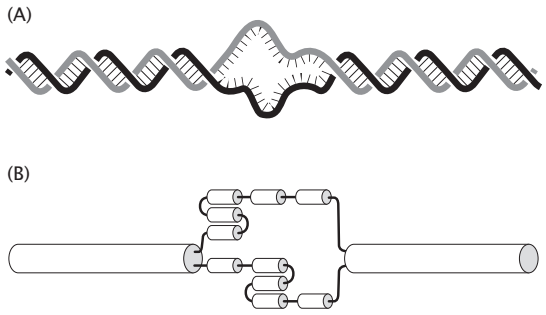

Figure 8.21: One-bubble Poland-Scheraga model. The possible states of a DNA strand of length N base pairs are labeled by the length of the single bubble, $1 \leq n \leq N$. (A) Schematic of a single bubble in the DNA. (B) One-dimensional random walk picture of the DNA with a bubble. The significance of the lengths of the cylinders is to characterize the difference in persistence length between the stiff dsDNA and the much more flexible ssDNA.

that this probability is given by

$$p_1(n) = \frac{e^{-\Delta G_1(n)/k_B T}}{Z} \tag{8.61}$$

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

$$Z = \sum_{n=1}^{N} e^{-\Delta G_1(n)/k_B T} \tag{8.62}$$

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

$$\Delta G_1(n) = E_{\text{in}} + nE_{\text{el}} - k_B T \ln \left(\Omega_\circ(n)(N-n)\right) \tag{8.63}$$

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

$$\Omega_\circ(n) = 2^{2n} p_\circ(2n) \tag{8.64}$$

which is nothing but the number of random walks of total $2n$ steps that return to the origin, introduced in eqn. 8.55. This reduces to

$$\Omega_\circ(n) = \frac{2^{2n}}{\sqrt{\pi n}} \tag{8.65}$$

for $n \gg 1$, where we have made use of eqn. 8.56 for the cyclization probability, $p_\circ(2n)$.

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

$$\frac{\Delta G_1(n)}{k_B T} = (\epsilon_{\rm el} - 2\ln 2)\, n + \frac{1}{2}\ln n - \ln(N-n)\ , \tag{8.66}$$

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

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

$$(\epsilon_{\rm el} - 2\ln 2) + \frac{1}{2n} + \frac{1}{N-n} = 0 \tag{8.67}$$

whose solutions are

$$n^*_\pm = N\frac{1 + \Delta\epsilon \pm \sqrt{1 + 6\Delta\epsilon + \Delta\epsilon^2}}{\Delta\epsilon} \tag{8.68}$$

where we have introduced a new variable

$$\Delta\epsilon \equiv 2(\epsilon_{\rm el} - 2\ln 2)\ . \tag{8.69}$$

Consider first the situation when $\Delta\epsilon > 0$. In this case both solutions, $n^*_\pm$ are not of interest as they do not correspond to bubbles whose length is positive and smaller than N. This means that on the interval $0 < n \leq N$ the free energy is monotonically increasing and therefore we expect that the state with no bubble wins out as one with the lowest free energy. (Note that this conclusion is not 100% guaranteed because the Stirling approximation gets worse as n becomes smaller.) Going back to the original parameters in the model, this means that for temperatures low enough so that $E_{\rm el}/k_B T > 2\ln 2$, the no-bubble state wins out. At higher temperatures, when $\Delta\epsilon < 0$, the situation is very different. In this case both solutions $n^*_\pm$ are of interest as they are both positive and less than N. One of the solutions is typically small compared to N and is a local maximum while the other is close in value to N and is a local minimum. In fig. 8.22 we show plots of the reduced free energy as given in eqn. 8.66 (i.e. without making the Stirling approximation) for values of $\epsilon_{\rm el}$ close to $2\ln 2 \approx 1.39$, which explicitly demonstrate this behavior. It is interesting to note that for $E_{\rm el}/k_B T < 2\ln 2$ even though the no-bubble configuration has the lowest free energy one should observe fluctuations into the one-bubble states with a typical bubble size that will depend on temperature. Also, close to the critical value of temperature the free energy as a function of bubble size becomes fairly flat so the prediction of the model is that one should observe bubbles of varying sizes appear simply due to thermal fluctuations.

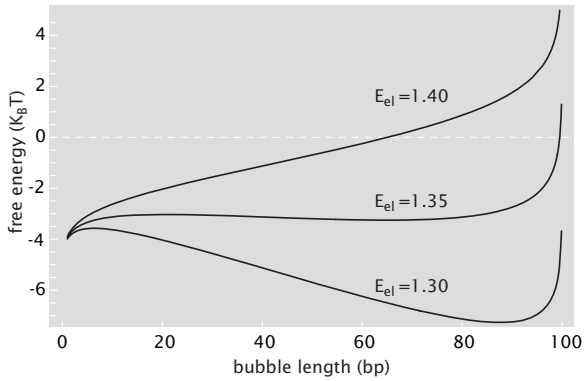

Figure 8.22: Free energy of the one-bubble model as a function of the bubble size. As the temperature is increased the reduced energy for bubble elongation ϵ_{el} becomes smaller and smaller. For small temperatures the most likely bubble size is zero, while at high temperatures it is close to the DNA length (here chosen to be $N = 100$ base pairs), and the chain is completely melted. At intermediate temperatures, the model predicts strong fluctuations of the bubble size indicated by the rather flat free energy landscape.

8.3 The New World of Single Molecule Mechanics

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

Single Molecule Measurement Techniques Lead to Force Spectroscopy

There are a number of different ways of applying forces to individual macromolecules. Several of these techniques are represented in schematic form in fig. 8.23. One such technique shown in fig. 8.23(A) involves the use of micron-sized cantilevers which are attached to a macromolecule which is, in turn, tethered to a surface. Through control of the height of the surface to which the molecule is tethered, for example, the cantilever will suffer a deflection which can be measured using reflected laser light. A second example shown in fig. 8.23(B) is optical tweezers which permit the application of forces of order 1-50 pN on macromolecules of interest. In this case, the key idea is that by attaching a macromolecule to a micron-sized bead, it is possible to pull on the bead (and hence the molecule) by shining laser light on the bead and using the resulting radiation pressure from the laser light to manipulate the bead. The same concept is similarly played out in the context of the magnetic tweezers shown in fig. 8.23(C) where the bead is manipulated by magnetic fields rather than laser

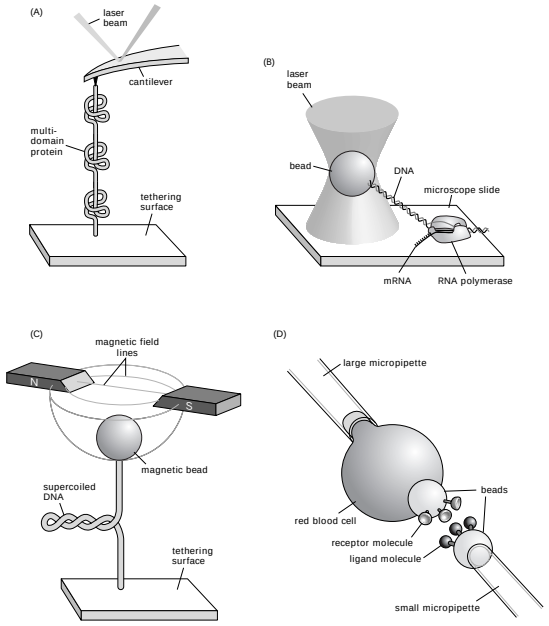

Figure 8.23: Schematic showing a variety of single molecule techniques. (A) single molecule atomic-force microscopy being used to stretch a multi-domain protein, (B) optical tweezers being used to measure the rate of transcription, (C) magnetic tweezers being used to measure the torsional properties of DNA and (D) pipette-based force apparatus being used to measure ligand-receptor adhesion forces.

light. One of the interesting variations on the forcing scheme provided by the magnetic tweezer is the opportunity to apply torsional forces which examine the response of molecules to twist. The final example shown in fig. 8.23(D) is the use of a pipette-controlled force apparatus in which the strengths of ligand receptor interactions as well as the mechanical response of lipid bilayer vesicles can be examined. Our main point in this discussion is to alert the reader to the emergence of single-molecule techniques that complement the tools of traditional solution biochemistry and permit the measurement of not only the average properties of the various macromolecules of biological interest, but also the fluctuations about this average response.

8.3.1 Force-Extension Curves: A New Spectroscopy

Different Macromolecules Have Different Force Signatures When Subjected to Loading

The techniques introduced above permit the explicit measurement of the

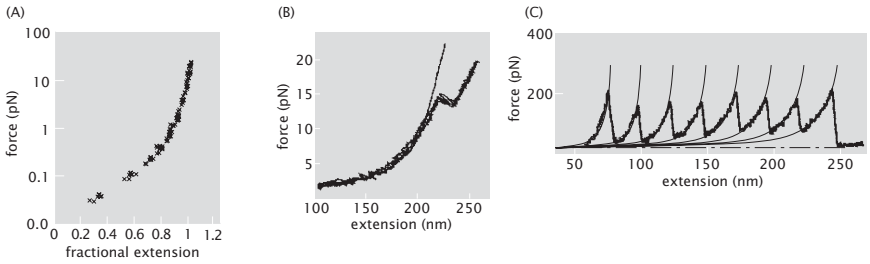

Figure 8.24: Force-displacement curves for a variety of different macromolecules: (A) double-stranded DNA, (B) RNA and (C) protein made of repeats of Ig module 27 of the I band of human cardiac titin. The measured curves illustrate the sense in which single molecule experiments serve as the basis of force spectroscopy. Non-monotonic features seen in the plots correspond to changes in structure due to an applied force. (A, adapted from C. Bustamante *et al.*, *Curr. Op. Struc. Biol.*, 10:279, 2000; B, adapted from J. Liphardt *et al.*, *Science*, 292:733, 2001; C, adapted from M. Carrion-Vazquez *et al.*, *Proc. Nat. Acad. Sci.*, 96:3694, 1999.)

force-extension characteristics of a range of different molecules. Fig. 8.24 shows the force-extension properties of several characteristic examples ranging from DNA to proteins. In particular, fig. 8.24(A) shows the force-extension characteristics of a single DNA molecule subjected to loading (a similar example was shown in fig. 5.14, pg. 265). Note that the same characteristic force-extension signature will be found for a given DNA molecule regardless of which of the various techniques is used to measure it, and further, that this curve provides a unique fingerprint which serves as the basis of *force spectroscopy* of macromolecules. Fig. 8.24(B) shows a plot of the force-extension properties of a particular RNA molecule. Note that the character of the secondary structure associated with a given RNA molecule is translated, in turn, into the character of the force-extension curve, illustrating the idea that the force-extension curve provides a spectroscopic fingerprint of different macromolecules. Fig. 8.24(C) shows yet a third example of the intriguing diversity of force-extension curves associated with different macromolecules, this time revealing how the multidomain protein titin unfolds in the presence of force. One immediate statement that can be made in this example is that the number of load drops in the curve corresponds to the number of unfolded domains in the protein. We emphasize that these three examples are but a tiny representation of the broad class of measurements that have been made on polysaccharides, lipids, proteins and nucleic acids as well as their assemblies.

8.3.2 Random Walk Models for Force-Extension Curves

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

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

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

$$\text{force} = -\frac{\partial G}{\partial L}, \tag{8.70}$$

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

We begin with a one-dimensional rendition of the freely-jointed chain model. We imagine a polymer of overall length $L_{\text{tot}} = Na$, where N is the number of monomers and a is the length of each monomeric segment. The basic thrust of our argument will be to construct the free energy $G(L)$ as a function of the length $L = (n_r - n_l)a$ from which the force necessary to arrive at that extension is given by eqn. 8.70. As before, we use the notation n_r and n_l to signify how many of the total links are right pointing (n_r) and how many are left pointing (n_l). In order to proceed, we need an explicit formula for the free energy. As noted above, in this simplest of models we ignore any enthalpic contributions to the free energy, with the entirety of the free energy of the molecule taking the form

$$G(L) = -k_B T \ln W(L; L_{\text{tot}}), \tag{8.71}$$

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

As shown in fig. 8.25, we are interested in the equilibrium of our random walk representation of the polymer when it is subjected to external forcing such as can be provided by an optical tweezers setup. A particularly transparent way to imagine this problem is to think of weights being dangled from the ends of the polymer as shown in fig. 8.26 (this idea of representing the energy of the loading device via weights was introduced in fig. 5.12 (pg. 263)). In this case, the free energy of eqn. 8.71 must be supplemented with a term of the form $U_{weights} = -2mgL$. What this term says physically is that the more the molecule is stretched, the lower the weights will dangle with the result that their potential energy is decreased.

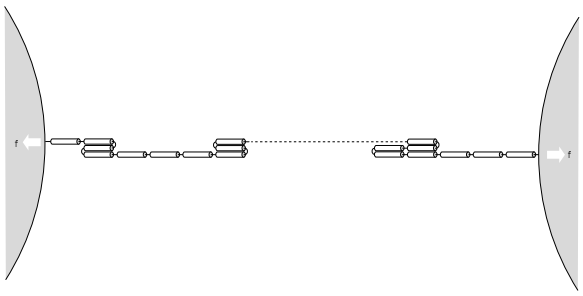

Figure 8.25: Model polymer subjected to loading. Schematic of a model one-dimensional polymer subjected to external forcing by optical tweezers.

Figure 8.26: Polymer subject to external load. Schematic of a model one-dimensional polymer subjected to external forcing through the attachment of weights on the end. This scenario is a pedagogical device to illustrate how to include the forcing in the overall free energy budget.

Putting together this term with the contribution from eqn. 8.71, we have for the total free energy of the system

$$G(L) = \underbrace{-2mgL}_{\text{contribution from weights}} - \underbrace{k_BT\ln W(L;L_{\text{tot}})}_{\text{entropic contribution of polymer conformations}}. \tag{8.72}$$

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

$$W(n_R;N) = \frac{N!}{n_R!(N-n_R)!}, \tag{8.73}$$

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

Given the free energy, our task now is to minimize it with respect to length (or n_R). To that end, we first invoke the Stirling approximation (pg. 280), which

we remind the reader allows us to replace $\ln N!$ by $N \ln N - N$. In light of this approximation, the overall free energy may be written as

$$G(n_R) = -2mgn_R a + k_B T \left(n_R \ln n_R + (N - n_R) \ln (N - n_R)\right). \tag{8.74}$$

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

$$\frac{\partial G}{\partial n_R} = -2mga + k_b T \ln n_R - k_B T \ln (N - n_R) = 0 \ . \tag{8.75}$$

Solving this equation for the quantity n_R/n_L we obtain

$$\frac{n_R}{n_L} = e^{\frac{2mga}{k_B T}} \tag{8.76}$$

which we use to obtain a simple relation for the extension

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

The construct of using weights to load the molecule was a convenient pedagogical device to provide a concrete mechanism for seeing how the energy of the loading device can be included in the free energy budget. More generally, the two ends are subjected to a force f with the result that $z = \tanh(fa/k_B T)$. This force-extension relation is shown in fig. 8.27. To gain further insight into the quantitative aspects of the model we consider the limiting case of a small force, i.e. $fa \ll k_B T$. For a dsDNA molecule in physiological conditions ($a \approx 100$nm) this corresponds to $f \ll 40$ fN while for the much more flexible ssDNA ($a \approx 1.5$nm) the small force regime is obtained for $f \ll 3$ pN. In the small force limit the force-extension curve is linear (as shown in the problems at the end of the chapter),

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

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

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

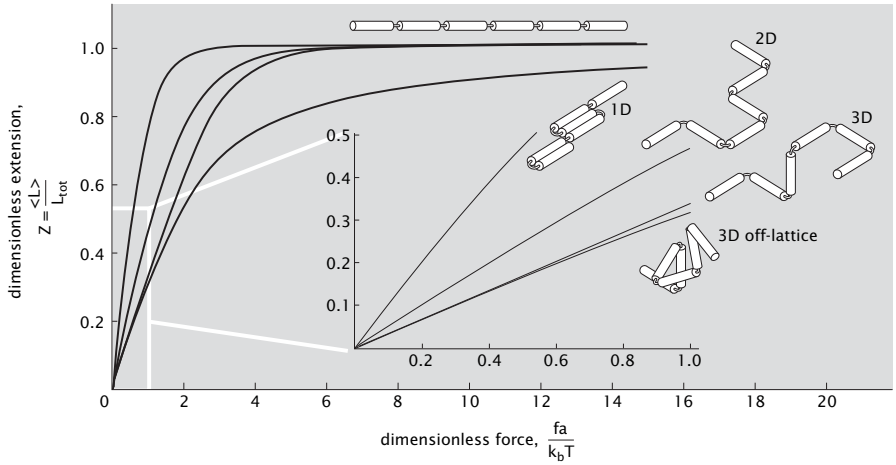

Figure 8.27: Relation between force and extension as obtained using the freely jointed chain model. Results for one-, two- and three-dimensions are shown and the three-dimensional case is shown for both the version in which the monomers can only point in the Cartesian directions and for the case in which they can point in any direction. The curves are related to their corresponding model by the cartoon showing the random-walk chain. A comparison of this model to the data was shown in fig. 5.14, pg. 265

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

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

and leave the details as an exercise for the reader.

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

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

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

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

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

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

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

At large forces when the polymer approaches full extension, the force-extension formula, eqn. 8.82, derived from the freely jointed chain model no longer adequately describes experimental data obtained by pulling on dsDNA. In that regime the elastic properties of dsDNA begin to matter and a more sophisticated model, which incorporates bending stiffness, describes the experimental data much better. This so-called worm-like chain model is taken up in chap. 10.

8.4 Proteins as Random Walks

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

Globular proteins in their native state form compact structures which are quite different from the open configurations implied by the random walk model. Therefore, we might be tempted to conclude that the random walk model has no business commenting on proteins. Instead we consider a modification of the random walk model we have employed so far by explicitly accounting for the compact nature of proteins.

The compact random walk model we employ in this section is defined on a lattice, meaning that the random walker, whose trajectories represent polymer configurations, jumps from one lattice site to the next. Usually when representing the polymer by a random walk on a lattice, the sites not occupied by the monomers (or, equivalently, those sites not visited by the random walker) are thought of as representing the solvent molecules. Simple random walks described in the previous sections are open structures with the monomer sites typically surrounded by solvent sites. As remarked above, this is inadequate for describing protein conformations which are compact with solvent typically making contact only with amino-acids at the surface of the protein. To mimic this property of proteins we invoke compact random walks (also referred to as Hamitonian walks) which are self-avoiding random walks that visit every site of the lattice, usually taken to be cubic, as depicted in fig. 8.28. By virtue of covering all the lattice sites by monomers, all the solvent sites are pushed to the surface. These compact random walks, are a very coarse grained model of proteins and, as with all coarse-grained models, one is limited in scope and precision of the questions that the model is equipped to address. The rewards on the other hand come in the form of simplicity and generality of the answers obtained. Furthermore, as with any good model, compact random walks reveal new questions and sharpen old ones about the structure of naturally occurring proteins.

8.4.1 Compact Random Walks and the Size of Proteins

The Compact Nature of Proteins Leads to an Estimate of their Size

Possibly the simplest property of a globular protein is its size, as measured by its linear dimensions, or more precisely, its radius of gyration. Examination of representative proteins from the Protein Data Bank reveals a systematic dependence of the protein size on its mass. In particular, for globular proteins the radius of gyration scales roughly with the cube root of the mass. The

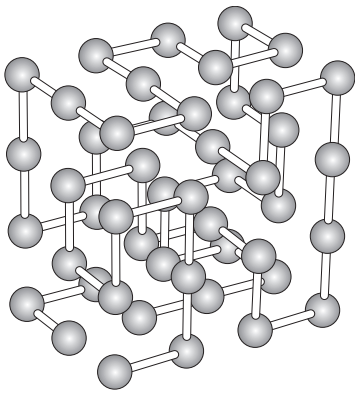

Figure 8.28: Compact polymer configuration on a 4x4x3 cubic lattice. Each ball represents an amino acid. (Adapted from K. A. Dill *et al.*, *Protein Sci.*, 4:561, 1995.)

relation between the physical size of proteins and their sequence size is shown in fig. 8.29.

The observed scaling is a simple consequence of the compact nature of proteins, and is thus also a property that is captured by compact random walks. Since a compact random walk completely fills the lattice (see fig. 8.28), its linear size will scale with the linear dimension of the lattice or with the cube root of the number of lattice sites, given that we have in mind a three-dimensional lattice. If we associate a single residue with each site, and take these to be of roughly equal mass, we arrive at the scaling law observed for many real proteins. Compactness implies that all the space occupied by proteins is filled, with no holes present. Therefore, the volume occupied by the protein, which necessarily scales as the cube of its linear dimension, is proportional to the mass. For proteins in the unfolded state, the structures are better described as random walks. The size of a random walk polymer, unlike compact polymers, scales as the 1/2 power of the mass. If one were to examine random self-avoiding walks (random walks with the additional constraint of no self-intersections), an argument due to Flory predicts scaling of the linear size with mass to the 3/5 power, indicating a structure which is even more expanded that that of a simple random walk.

8.4.2 Hydrophobic and Polar Residues: The HP Model

One of the challenges brought in on the heels of the successes of the great genomic sequencing initiatives is that of figuring out the structural and functional implications of these vast libraries of genes. One step in unraveling the meaning of all of this genomic data is to figure out how to go from a particular protein sequence to the corresponding structure. The problem is that when confronted with some new gene sequence, one would like to be able to state what proteins

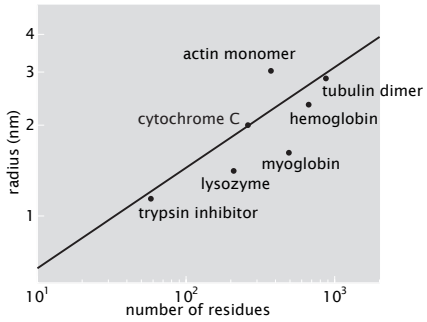

Figure 8.29: Scaling of protein size as a function of the number of amino acid residues. The line has a slope of 1/3 corresponding to a space-filling packing.

are implied by the various sequences and what structures these proteins have. Like for the analysis of protein-ligand binding in chap. 7, here too we will find that the use of internal-state variables to characterize the amino acid identity of a given residue is extremely powerful.

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

The above estimate tells us that the folding of a protein into its native structure is most certainly not a random process. The hydrophobic interaction between amino-acid residues and the water molecules that surround them leads to a collapse of the chain as was illustrated in fig. 5.8 (pg. 258). As a result the hydrophobic residues are sequestered to the interior of the protein, while the surface is populated by polar residues. Thus hydrophobicity is one force that can steer the protein to a folded state avoiding a random search of configuration space. Indeed, the spirit of the class of models introduced here is that collapse

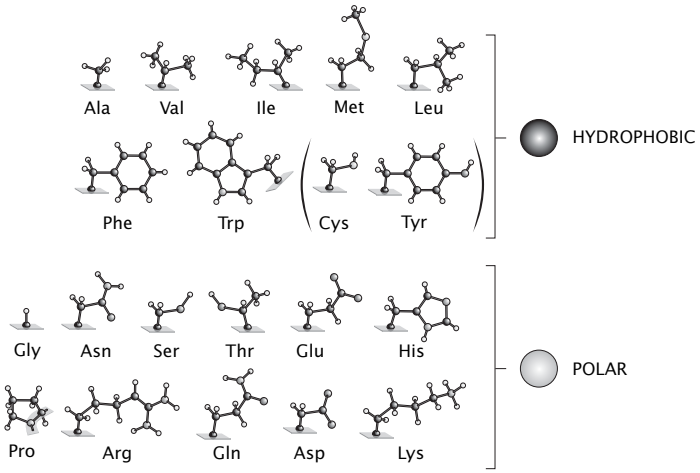

Figure 8.30: Mapping of the amino acids onto an HP alphabet. The 20 amino acids are coarsely separated into two categories, hydrophobic (H) or polar (P).

induced by hydrophobic effects drives the formation of secondary structure as opposed to an alternative view in which the formation of the hydrogen bonds that define secondary structure lead to collapse.

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

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

In the remainder of the book, we will use the HP model introduced here as the basis of a variety of different discussions. Our reasoning is that classifying amino acids according to just these two broad categories allows us to take otherwise analytically intractable problems and to render them tractable. For example, in section 18.4.1 (pg. 988), we will consider an HP model of translation and kinetic proofreading featuring only two species of tRNA. This simplification will allow us to carry out the analysis completely. Similarly, the entirety of chap. 18 on bioinformatics will be based on sequence alignments using only the HP alphabet. Though we compromise on biological realism, our sense is that

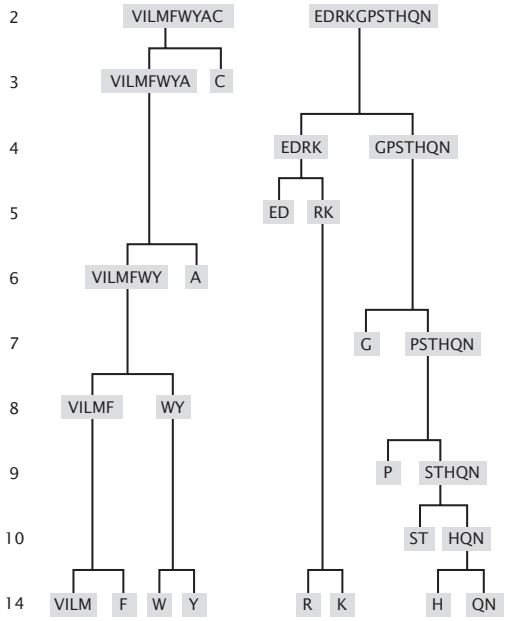

Figure 8.31: Hierarchy of amino-acid classifications. Groupings of amino acids into "classes" with similar properties. At the top of the figure, the amino acids are grouped into two categories, hydrophobic (H) on the left and polar (P) on the right. At each level the number of distinct classes is shown by the integer on the left. (Adapted from K. A. Dill and P. D. Thomas, *Proc. Nat. Acad. Sci.*, 93:11628, 1996.)

the pedagogical payoff is worth it.

8.4.3 HP Models of Protein Folding

The protein folding problem of finding the native structure given the amino acid sequence of a protein is one of a class of problems concerning the relationship between the amino-acid sequence space and the space of three-dimensional protein structures. Just as introducing a two letter alphabet greatly reduces the sequence space, constraining the space of structures to compact random walks on a lattice makes the exploration of structure space more tractable. In particular the number of compact polymer structures on a $3 \times 3 \times 3$ lattice, often used in numerical studies, is 103,346, while the number of possible sequences is $2^{27} = 134,217,728$.

To gain intuition about lattice HP models we will investigate the toy model that consists of 6 monomers on a 2×3 lattice. The number of possible *sequences* is $2^6 = 64$ while the number of compact structures that are unrelated by lattice rotations, translations or reflections is only 3. These are shown in fig. 8.32(A). Beside the list of sequences and structures, the other ingredient of the model is the hydrophobic energy which measures the extent to which the H-monomers make energetically unfavorable contacts with the solvent and with P-monomers. (Solvent molecules are the lattice lattice sites on the outside surface of the six-mer.) A simple model of this interaction is to assign a free energy penalty ϵ for every contact an H monomer makes with either a solvent molecule or a P monomer. These unfavorable contacts are shown as dashed lines in fig. 8.32(B). A more refined model might distinguish the interaction energy associated with an H-solvent and an H-P contact.

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

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

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

the denominator is nothing but the partition function for the three possible structures. The probability of this toy protein adopting the folded state as a

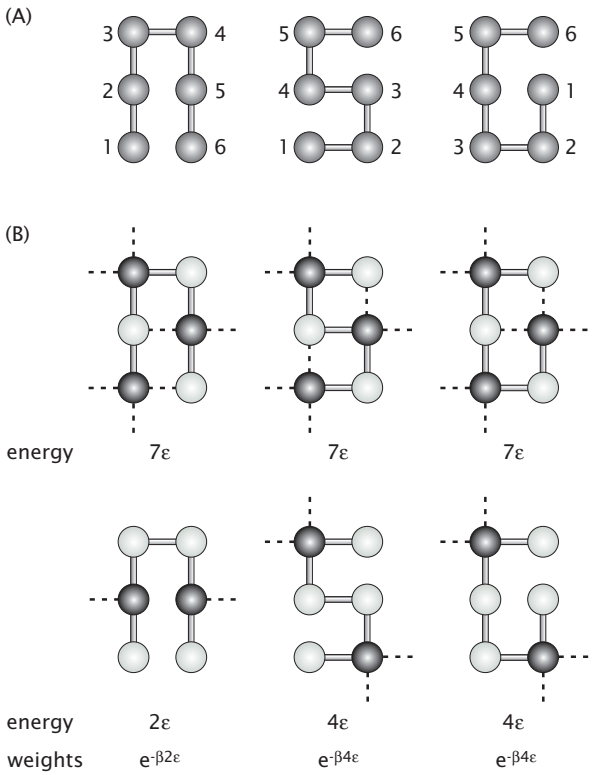

Figure 8.32: HP model of protein folding. (A) Possible compact structures of an HP six-mer on a 2×3 lattice, unrelated by symmetries. (B) The hydrophobic energy of an HP six-mer in a particular compact structure depends on its sequence. The energy function assigns a cost ϵ for every contact, represented here by a dashed line, between an H-monomer and either a P-monomer or a solvent molecule. The sequence in the top panel, HPHPHP, has the same energy in all three compact structures, while PHPPHP has one structure as its unique lowest energy state, which is characteristic of protein-like sequences.

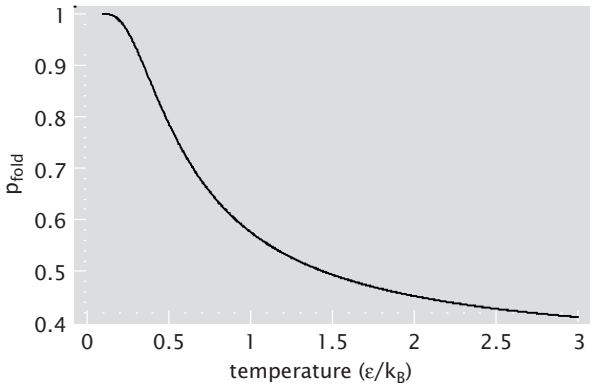

Figure 8.33: Probability of finding the PHPPHP polymer in its native state.

function of temperature is shown in fig. 8.33. Note the sigmoidal character of the plot which is characteristic of many real proteins.

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

In the context of our toy model we can address this question systematically as there are only 64 sequences to go through. For every sequence we would need to determine its energy in each one of the three possible compact structures, in order to identify the protein-like sequences with a unique lowest-energy structure.

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

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

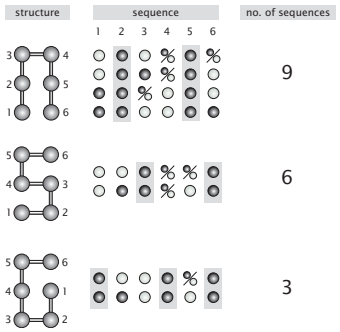

Figure 8.34: Protein-like sequences fold into a unique compact structure. The number of protein-like sequences varies from compact structure to compact structure. The structures with a particularly large number of protein-like sequences associated with them are highly *designable*.

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

8.5 Summary and Conclusions

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

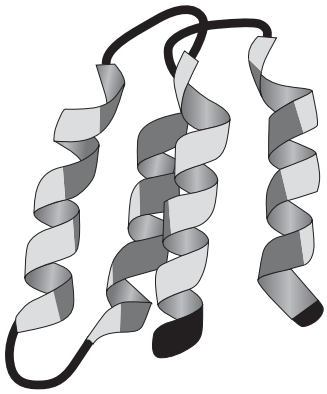

Figure 8.35: The four-helix bundle designed by using an HP sequence strategy. The HP sequence is chosen to conform to the 3.6 amino-acids per turn of an alpha-helix. The hydrophobic residues are sequestered in the interior of the bundle, while the polar ones are on the surface facing the solvent. (Adapted from M. H. Hecht *et al.*, *Protein Sci.*, 13:1711, 2004.)

8.6 Problems

1. **Gaussian chain in 3D.** Perform all the steps that lead to eq. 8.14.

2. **1-dimensional random walk.** Show that the Gaussian distribution of R for a 1-dimensional random walk given in eqn. 8.16 indeed has the required mean and variance.

3. **Radius of gyration.** Prove the relation between the contour length and the radius of gyration given by eqn. 8.32.

4. **Mean departure from the origin.** Compute the mean departure of a one dimensional random walker from its starting point. In particular, use the fact that the mean excursion can be written as $\langle R \rangle = (\langle n_r \rangle - \langle n_l \rangle)a$ and that the probability distribution for n_r right steps out of a total of N steps is given by the binomial distribution.

5. **Diffusion and master equations.** Eqn. 8.39 characterizes the probability distribution for random walkers. Derive this equation by using the fact that the probability that the walker will be at position x at step N implies that the walker was either at $x - a$ or $x + a$ at step $N - 1$. In particular, write an equation for $p(x, N)$ in terms of $p(x \pm a, N - 1)$ and by Taylor expanding $p(x \pm a, N - 1) \approx p(x, N) +$ derivative terms.

6. Random walk in a cylinder. Use a generalization of eqn. 8.39 to three dimensions to solve for the probability distribution for the end-end distance for a polymer tethered along the axis of a cylinder at position x_0. Use your results to compare to the data shown in fig. 8.14.

7. Chromosome tethering. (a) Given $P(\mathbf{R})$, the probability density for the vector $\mathbf{R}$ which characterizes the probability that the random walk will end in a little volume element at $\mathbf{R}$, find the probability density for $P(R)$ - the probability density that the random walk will end at a distance R from the origin. and then write an expression for the probability that the end to end distance is R.
(b) Obtain a careful derivation of the result given in 8.35.

8. 3D random walk and polymer cyclization. Calculate the cyclization probability of a discrete random walk in one, two and three dimensions.

9. Bubbles on DNA Compare the free energy of the one bubble state vs the two bubble state as a function of the chain length N. The point is to show that for short DNA strands the one-bubble state is dominant over the two bubble state.

10. Freely Jointed Chain in 3D (a) Derive eqn. 8.79, use the result to derive the relation between force and extension and make a plot of the resulting function.

(b) In the small force limit the force-extension curve is linear, i.e. in this regime the polymer behaves like an ideal Hookean spring with a stiffness constant $k \propto k_B T / L_{\text{tot}} a$. Demonstrate this claim and deduce the numerical factors that replace the proportionality with a strict equality.

11. Force-induced unfolding of multidomain proteins can be modeled using the random walk model. We can generalize the discussion of force-extension curves to the case of multidomain proteins. The data relevant to the particular case of the muscle protein titin has already been shown in fig. 8.24(C). The idea of the analysis we will bring to bear on this problem is shown in fig. 8.36, where it is seen that the overall contour length of the chain increases in a systematic and calculable way as a function of the number of domains that have unfolded. Use a model like that suggested in fig. 8.36 to compare to the data shown in fig. 8.24(C).

12. Transition between B and S form DNA. DNA subjected to a stretching force exceeding 60pN undergoes a structural transition from the usual B-form to the so-called S-form ("S" for stretch). Here we examine a simple model of this transition based on the freely-jointed chain model of DNA and compare it to experimental data.

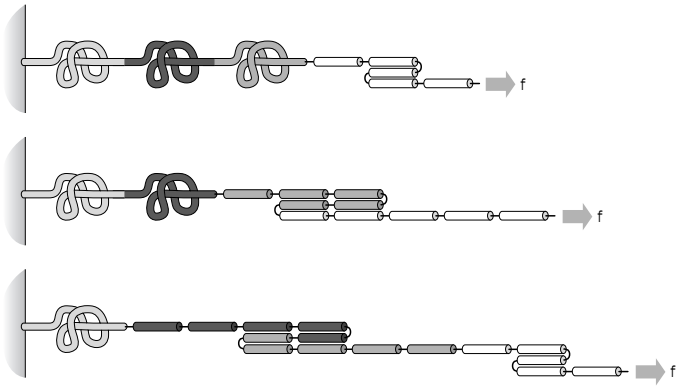

Figure 8.36: Random walk model for elastic parts of the titin force- extension curve.

(a) Consider the freely-jointed chain model in one dimension. Each link of the polymer points in the +x or the -x direction. There is a force F in the +x direction applied at one of the ends (see Fig.b). To account for the B to S transition we assume that links are of length b (B state) or a (S state), with $a > b$. Furthermore, there is an energy penalty ϵ of transforming the link from a B state to an S state. (This is the energy, presumably, for unstacking the base pairs.) Write down the expressions for the total energy and the Boltzmann factor for each of the four state of a single link, shown in fig. 8.37.

(b) Compute the average end-to-end distance for one link. The average end-to-end distance for a chain of N links is N times as large.

(c) Plot the average end-to-end distance normalized by Na (ie. the relative extension) as a function of force using the numbers appropriate for DNA: $b =$ 100 nm, $a = 190$ nm. To estimate ϵ take the *energy per base pair* for transforming B-DNA to S-DNA to be 5 k_BT (the length of one base pair is 1/3 nm long). How does your plot compare to fig. 8.37?

13. Scaling of Protein Size The scaling of a polymer's size as a function of the number of monomers is one of the central results to emerge from simple lattice models of polymers. The goal of this problem is to investigate the extent to which such arguments are in fact appropriate for biological polymers, and in particular, proteins. To that end, use the Protein Databank in order to download the coordinates for a variety of globular proteins, including myoglobin, hemoglobin, bovine pancreatic trypsin inhibitor (BPTI), lysozyme, cytochrome c, G-actin and tubulin. In each case, compute the radius of gyration and then make a single plot which shows the radius of gyration for each of these proteins as a function of the number of residues in the protein. The goal of the problem is to reproduce fig. 8.29.

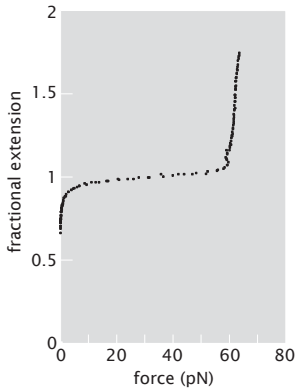

Figure 8.37: Force-extension curve for dsDNA. (Adapted from C. Bustamante *et al.*, *Nature*, 421:423, 2003.)

8.7 Further Reading

A. Y. Grosberg and A. R. Khokhlov, **Giant Molecules**, Academic Press, San Diego, California, 1997. This book is a thoughtful discussion of polymer physics that is pleasing to novices and professionals alike. Interested readers should also see their more advanced **Statistical Physics of Macromolecules**, American Institute of Physics, Woodbury: New York, 1994.

G. B. Benedek and F. M. H. Villars, **Physics With Illustrative Examples from Medicine and Biology: Statistical Physics**, Springer-Verlag, Inc., New York: New York, 2000. We have referred to the series by Benedek and Villars throughout the book - as always, a great source of interesting material.

H. Berg, **Random Walks in Biology**, Princeton University Press, Princeton: New Jersey, 1993. Berg's book is an enlightening classic. The discussion on random walks pertains to diffusion, but the understanding garnered in that setting can be used to think about polymers.

M. Doi, **Introduction to Polymer Physics**, Oxford University Press, Oxford: England, 1995. and M. Doi and S. F. Edwards, **The Theory of Polymer Dynamics**, Oxford University Press, Oxford: England, 1986. These books give the interested reader insights into the statistical physics of polymers.

S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy", Rev. Mod. Phys. **15**, 1 (1943). Chandrasekhar's amazing article is a compendium of elegant and useful results pertaining to random walks and more general ideas on stochastic processes.

P. Nelson, **Biological Physics: Energy, Information, Life**, W. H. Freeman and Company, New York: New York, 2004. Nelson's treatment of the elasticity and force-extension properties of DNA is excellent.

D. Poland and H. A. Scheraga, **Theory of helix-coil transitions in biopolymers; statistical mechanical theory of order-disorder transitions in biological macromolecules**, Academic Press, New York: New York, 1970. This book illustrates the use of simple lattice models like that we used for DNA melting applied also to problems in protein folding.

P. G. de Gennes, **Scaling Concepts in Polymer Physics**, Cornell University Press, Ithaca: New York, 1979. One of the great classics in the field of polymer physics. de Gennes' approach to intuitive models and simple arguments should inspire the next generation of physical biologists.

A. Fiebig, K. Keren and J.A. Theriot, "Fine-scale time-lapse analysis of the biphasic, dynamic behaviour of the two *Vibrio cholerae* chromosomes", Molecular Microbiology, **60**, 1164 (2006). An interesting example of the experimental study of chromosome geography.

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

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

K. Rippe, "Making contacts on a nucleic acid polymer", Trends Biochem. Sci. **26**, 733 (2001). This article demonstrates some of the interesting ways that polymer physics can be used to study biological problems pertaining to chromosome structure and organization.

8.8 References

C. Bustamante, S. B. Smith, J. Liphardt and D. Smith, "Single-molecule studies of DNA mechanics", Curr. Op. Struc. Biol. **10**, 279 (2000).

C. Bustamante, Z. Bryant and S. B. Smith, "Ten years of tension: single-molecule DNA mechanics", Nature **421**, 423 (2003).

M. Carrion-Vazquez, A. F. Oberhauser, S. B. Fowler, P. E. Marszalek, S. E. Broedel, J. Clarke and J. M. Fernandez, "Mechanical and chemical unfolding of a single protein: A comparison", Proc. Nat. Acad. Sci. **96**, 3694 (1999).

J. Liphardt, B. Onoa, S. B. Smith, I. Tinoco Jr. and C. Bustamante, "Reversible Unfolding of Single RNA Molecules by Mechanical Force", Science **292**, 733 (2001).

W.F. Marshall, "Order and disorder in the nucleus", Current Biology **12**, R185 (2002).

G. A. Petsko and D. Ringe, **Protein Structure and Function**, New Science Press, Ltd., London: England, 2004.

S. B. Smith, Y. Cui and C. Bustamante, "Overstretching B-DNA: The Elastic Response of Individual Double-Stranded and Single-Stranded DNA Molecules, Science **271**, 795 (1996).

P. D. Thomas and K. A. Dill. "An iterative method for extracting energy-like quantities from protein structures", Proc. Natl. Acad. Sci. **93** 11628 (1996).

G. van den Engh, R. Sachs and B.J. Trask, "Estimating genomic distance from DNA sequence location in cell nuclei by a random walk model", Science **257**, 1410 (1992).

P.H. Viollier, M. Thanbichler, P. T. McGrath, L. West, M. Meewan, H. H. McAdams and L. Shapiro, "Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication", Proc. Nat. Acad. Sci. **101**, 9257 (2004).

P. A. Wiggins, T. van der Heijden, F. Moreno-Herrero, A. Spakowitz, R. Phillips, J. Widom, C. Dekker and P. C. Nelson, "High flexibility of DNA on short length scales probed by atomic force microscopy", Nat. Nanotech. **1**, 137 (2006).