# Random Walks and the Structure of Macromolecules

# 8

**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 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.

> "I only went out for a walk and finally concluded to stay out till sundown, for going out, I found, was really going in."
>
> **John Muir**

## 8.1 What Is a Structure: PDB or $R_G$?

The study of structure is often a prerequisite to tackling the more interesting question of the functional dynamics of a particular macromolecule or macromolecular assembly. Indeed, this notion of the relation between structure and function has been elevated to the status of the true central dogma of molecular biology, namely, "sequence determines structure determines function" (Petsko and Ringe, 2004), which calls for uncovering the relation between sequence and consequence. The idea of structure is hierarchical and subtle, with the relevant detail that is needed to uncover function often occurring at totally disparate spatial scales. For example, in thinking about nucleosome positioning, an atomic-level description of the state of methylation of the DNA might be 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 versus 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 PDB and VIPER. In this world view, a structure is the set $(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, where $\mathbf{r}_i$ is the vector position $\mathbf{r}_i = (x_i, y_i, z_i)$ 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 that 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 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 of 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

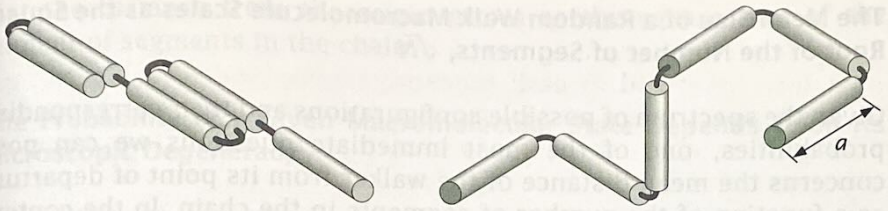


**Figure 8.1:** Random walk model of a polymer. Schematic representation of (A) a one-dimensional random walk and (B) a three-dimensional random walk as an arrangement of linked segments of length $a$.

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 Figure 8.1. In the one-dimensional case, the segments are at $\pm 180^\circ$ with respect to each other. We draw them as nonoverlapping for clarity. For the three-dimensional case, we illustrate the situation in which the links are restricted to $90^\circ$ 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).

Figure 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, Figure 8.2 shows a conformation of DNA on a surface. Using the discretization advocated above, we show how this same structure can be approximated using a series of rigid rods (the Kuhn segments) connected by flexible hinges. We will argue that this level of description can be useful in settings ranging from estimating the entropic cost of confining DNA to a bacterial cell to the stretching of DNA by laser tweezers.

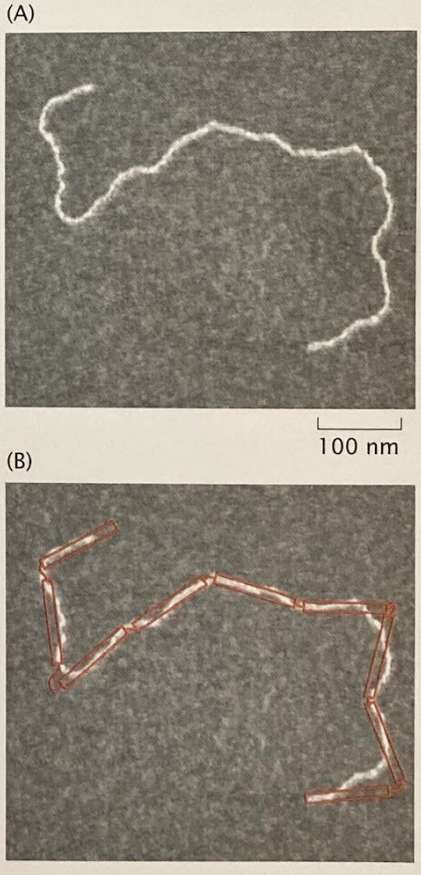


**Figure 8.2:** DNA as a random walk. (A) Structure of DNA on a surface as seen experimentally using atomic-force microscopy. (B) Representation of the DNA on a surface as a random walk. (Adapted from P. A. Wiggins et al., *Nat. Nanotech.* 1:37, 2006.)

### 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 Figure 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 right and left steps are given as $p_r = p_\ell = 1/2$. 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 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 each 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

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

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, the probabilities are all equal). This result may be simplified by noting that the averaging operation represented by the brackets $\langle \cdots \rangle$ on the right-hand side of the equation can be taken into the summation symbol (that is, 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

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

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 Equation 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 right-hand 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

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

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 State 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 $n_\ell = N - n_r$ to the left)? Since the probability of each right or left step is given by $p_r = p_\ell = 1/2$, 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 $n_\ell$ 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 on p. 239. A particular example of this thinking to the case $N = 3$ is shown in Figure 8.3, where we see that there is one configuration in which all three segments are right-pointing, one configuration in which all three

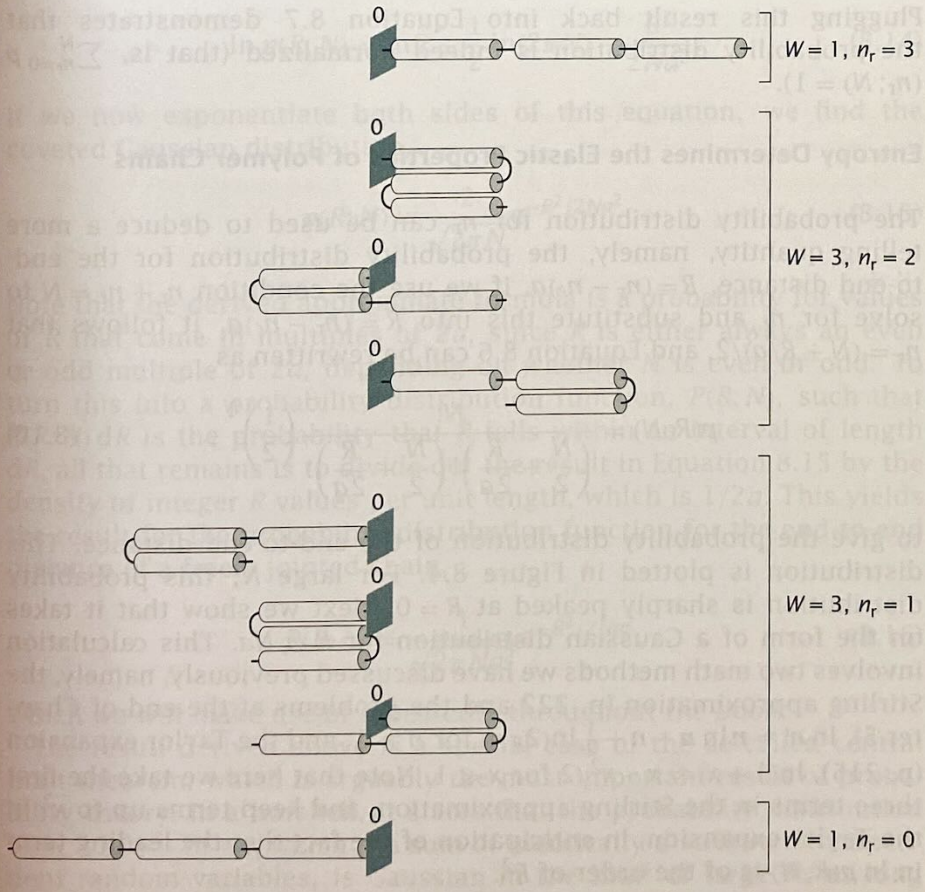


**Figure 8.3:** Random walk configurations. The schematic shows all of the allowed conformations of a polymer made up of three segments ($2^3 = 8$ conformations) and their corresponding degeneracies.

segments are left-pointing, and three configurations each for the cases in which $n_r = 2, n_\ell = 1$, and $n_r = 1, n_\ell = 2$.

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

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

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

### Entropy Determines the Elastic Properties of Polymer Chains

The probability distribution for $n_r$ can be used to deduce a more telling quantity, namely, the probability distribution for the end-to-end distance, $R = (n_r - n_\ell)a$. If we use the condition $n_r + n_\ell = N$ to solve for $n_\ell$ and substitute this into $R = (n_r - n_\ell)a$, it follows that $n_r = (N + R/a)/2$, and Equation 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 Figure 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, namely, the Stirling approximation (p. 222 and the problems at the end of Chapter 5), $\ln n! \approx n \ln n - n + \frac{1}{2}\ln(2\pi n)$ for $n \gg 1$, and the Taylor expansion (p. 215), $\ln(1 + x) \approx x - x^2/2$ 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 the order of $R^2$.

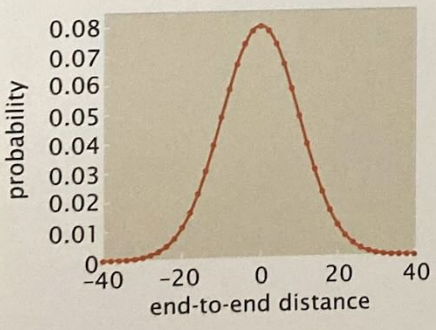


**Figure 8.4:** End-to-end probability distribution for a one-dimensional "macromolecule" with 100 segments. The figure shows a comparison of the binomial distribution (dots) given in Equation 8.10 and the approximate Gaussian distribution (curve) given in Equation 8.16.

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

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

In the next step, we rewrite the logarithms,

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

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

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

which we substitute repeatedly in Equation 8.11. After a bit of algebra (which is left as an exercise for the reader), we arrive at the formula

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

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

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

Note that the derived approximate formula is a probability for values of $R$ that come in multiples of $2a$, since $R$ is either always an even or odd multiple of $2a$, 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 Equation 8.15 by the density of integer $R$ values per unit length, which is $1/2a$. This yields the result for the probability distribution function for the end-to-end distance of a freely jointed chain,

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

which we will make use of repeatedly throughout the book.

The result derived above is a special case of the so-called central limit theorem, which is arguably the most important result of probability theory. In a nutshell, it states that the probability distribution of $x_1 + x_2 + \cdots + x_N$, which is a sum of identically distributed independent random variables, is Gaussian in the limit of large $N$, as long

as the mean and variance of each individual $x_i$ are finite. Since the individual displacements of the random walker satisfy this condition, it immediately follows that for a large number of steps $N$, the total displacement $R$ is Gaussian-distributed, with mean $\langle \mathbf{R} \rangle = 0$ and variance $\langle \mathbf{R}^2 \rangle = Na^2$. Note that this holds regardless of whether the walk is executed in one, two, or three 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 one-dimensional walk given in Equation 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 three-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 $\kappa$ are to be determined from the two identities

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

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

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

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 $\kappa$, we can divide the second of Equations 8.20 by the first to give

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

Substituting this result into the integral in the first of Equations 8.20 gives us

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

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

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

Note that $P(\mathbf{R}; N)$ has units of inverse volume, or concentration, and has an intuitive interpretation as the concentration of one end of the random-walk polymer at position $\mathbf{R}$ in the vicinity of 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 nonzero, 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 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 the pressure of an ideal gas, another example of an entropic force.

**Estimate: End-to-End Probability for the *E. coli* genome**

ESTIMATE

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 base pairs long. An open DNA chain of the same size can be modeled as a random walk of roughly $N = 15{,}000$ steps since the Kuhn length (the length of the "rigid" segments in the chain model) for bare DNA is roughly 300 bp. The probability that the end-to-end distance is zero for a one-dimensional walk of this many steps can be estimated from Equation 8.15 and is $7 \times 10^{-3}$. The probability that $R = 500a$ is $2 \times 10^{-6}$, while for $R = 1000a$ the probability drops all the way down to $2 \times 10^{-17}$. 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 $\xi_p$, 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. Figure 8.5 conveys this idea in the case of DNA by illustrating the length scale over which the genomic DNA of a bacterium meanders.

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 that 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

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

In Chapter 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 Equation 8.24 is not universally valid. For example,

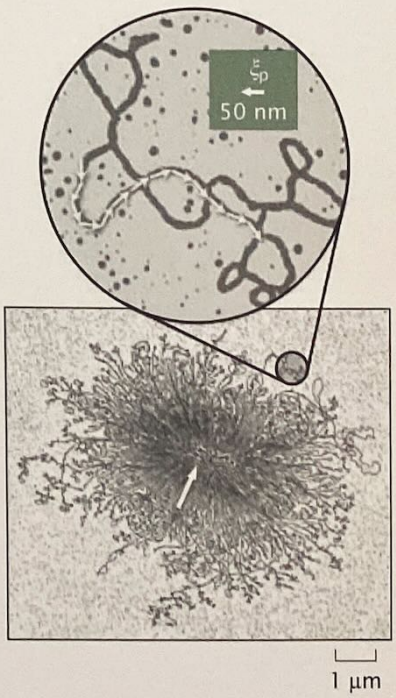


**Figure 8.5:** Illustration of the spatial extent of a bacterial genome that has escaped the bacterial cell. The expanded region in the figure shows a small segment of the DNA and has a series of arrows on the DNA, each of which has a length equal to the persistence length in order to give a sense of the scale over which the DNA is stiff. (Adapted from an original by Ruth Kavenoff.)

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 Equation 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 $16.6\,\mu\text{m}$. This should be compared with the persistence length $\xi_p \approx 50\,\text{nm}$ 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 that take random orientations with respect to each other. This is the logic that 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 with 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

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

where $\langle \cdots \rangle$ is an average over all polymer configurations. Using the average of the tangent–tangent correlation function, Equation 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 result is obtained by splitting up the integration over the $L \times L$ box in $s$–$u$ space into integrals over the two triangles, one with $s < u$ and the other with $s > u$, which give equal contributions (thus the factor of 2). In the limit $L \gg \xi_p$ we are considering here, we have

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

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 $\infty$, which is justified in the $L \gg \xi_p$ limit. Comparing the above formula with the result that follows from the random walk model, Equation 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 $\xi_p$ can be written as (proven by the reader in the problems at the end of the chapter)

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

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

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

This relation between the radius of gyration of DNA in solution and the number of base pairs is plotted in Figure 8.6.

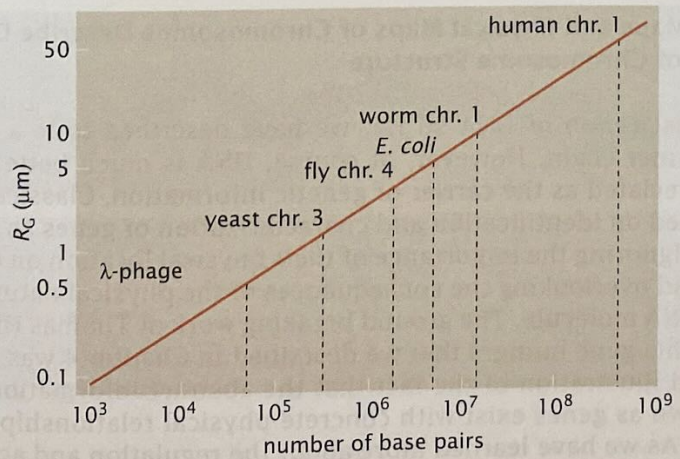


**Figure 8.6:** Size of genomic DNA in solution. Plot of the average size of a DNA molecule in solution as a function of the number of base pairs using the random walk model. The labels correspond to particular chromosomes from viruses, bacteria, yeast, flies, worms, and humans.

**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 Figure 1.16 (p. 28) 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 Figure 8.5, where it is seen that the DNA adopts a configuration in solution that 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 Equation 8.32 to formulate an estimate of the size of the DNA. Consider Figure 1.16, 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 $\xi_p \approx 50\,\text{nm}$, Equation 8.33 tells us that the mean size of the DNA seen in Figure 1.16 is $\sqrt{\langle R_G^2 \rangle} = (1/3)\sqrt{150 \times 10^3 \times 50}\,\text{nm} \approx 0.9\,\mu\text{m}$. This result is comparable to though larger than the length scale of the exploded DNA seen in Figure 1.16. 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. 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 Equation 8.33 tells us that the mean size of the DNA seen in Figure 8.5 is $\sqrt{\langle R_G^2 \rangle} \approx 5\,\mu\text{m}$. As with the phage calculation, the random walk calculation should be seen as an overestimate, since the bacterial genome is circular and the DNA is clearly forced to return to the bacterium repeatedly, inhibiting the structure from adopting a fully expanded configuration.

ESTIMATE

### 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 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 that we described in Chapter 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. During 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, $\nu$, which specifies the number of base pairs of DNA in a nanometer of chromatin fiber. For the 30 nm fiber, shown in Figure 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 $\nu$ can be made based on the micrograph in Figure 8.7(B), which shows individual nucleosomes along the 10 nm fiber. We see that there are on average two 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), and therefore $\nu \approx 2 \times 200\,\text{bp}/50\,\text{nm} = 8\,\text{bp/nm}$. For comparison, for metaphase chromosomes, $\nu \approx 30{,}000$ bp/nm.

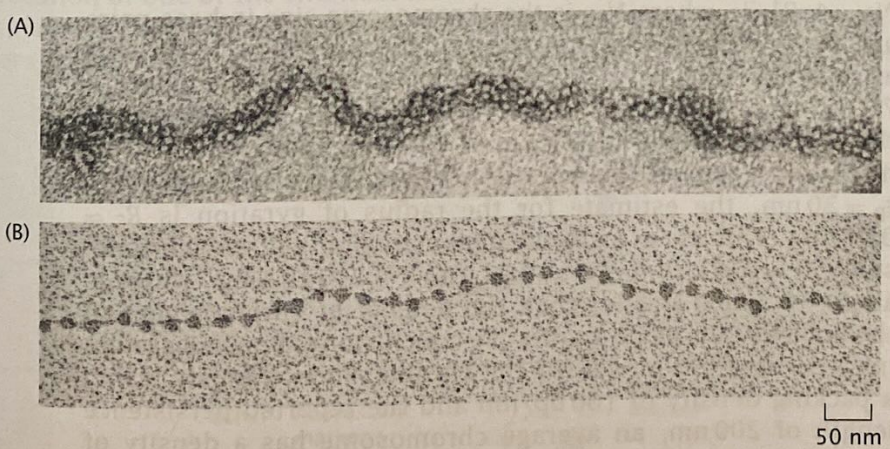


**Figure 8.7:** Electron microscopy images of chromatin. (A) Chromatin extracted from an interphase nucleus appears as a 30 nm thick fiber. (B) Stretching out a part of the chromatin reveals the "beads-on-a-string" structure of the 10 nm fiber, where each bead is an individual nucleosome. (A, courtesy of Barbara A. Hamkalo; B, courtesy of Victoria Foe.)

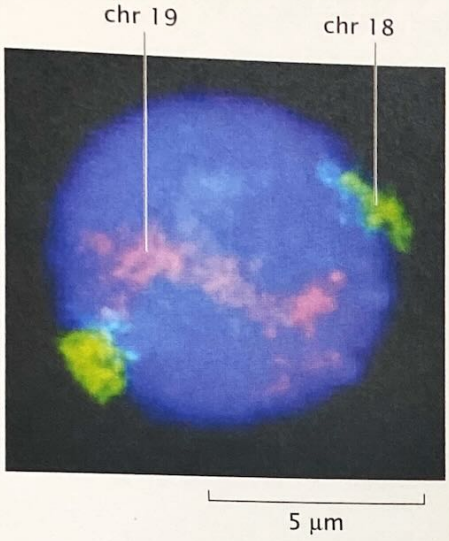


**Figure 8.8:** Fluorescently stained chromosomes 18 (green) and 19 (red) in the nucleus of a human cell. The two copies of chromosome 18 typically assume positions near the periphery of the nucleus, while the two copies of chromosome 19 are located closer to the center. (From J. A. Croft et al., *J. Cell Biol.* 145:1119, 1999.)

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

It used to be 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 Figure 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 that was 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 nonetheless observed points to the existence of mechanisms beyond polymer chain entropy and confinement that affect the spatial distribution of chromosomes. We will examine chromosome tethering as one such mechanism. Tethering scenarios posit that chromosomes have particular physical locations because they are held there by tethering molecules. Possible tethering scenarios are shown in Figure 8.9.

ESTIMATE

**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 segregated blobs not unlike meatballs. The yeast cell has 16 chromosomes in its nucleus. The diameter of the interphase nucleus is about $2\,\mu m$. The chromosome size varies between 230 kb and 1500 kb, with a total genome size of 12 Mb. This gives a mean density of $c = 12\,\text{Mb}/[(4\pi/3) \times (1\,\mu m)^3] \approx 3\,\text{Mb}/\mu m^3$. 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 $c^* = N_{bp}/(4\pi R_G^3/3)$, 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 $12\,\text{Mb}/16 = 750\,\text{kb}$ and a packing density of 8 bp/nm, the length of this polymer is $750\,\text{kb}/(8\,\text{bp/nm}) = 94\,\mu m$. Using the *in vitro* measured value of the persistence length for a 10 nm fiber, $\xi_p = 30\,\text{nm}$, the estimate for the radius of gyration is $R_G = 0.97\,\mu m$. This then leads to a density for a "free" chromosome $c^* = 750\,\text{kb}/[(4\pi/3) \times (0.97\,\mu m)^3] \approx 200\,\text{kb}/\mu m^3$, 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

$c^* \approx 500\,\mathrm{kb}/\mu\mathrm{m}^3$. This indicates that the chromosomes in the yeast nucleus should typically be found in an entangled melt-like configuration 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 Figure 8.9.

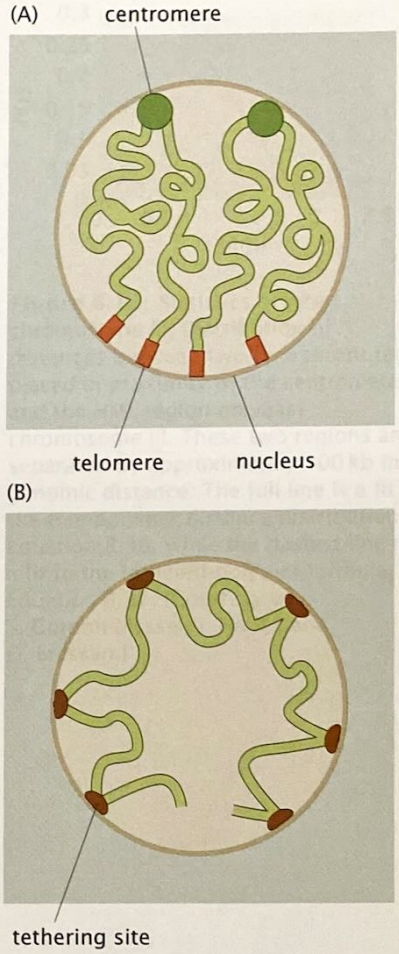


**Figure 8.9:** Cartoon representation of possible tethering scenarios of interphase chromosomes. (A) Tethering at the centromere and the two telomeres at the nuclear periphery and (B) tethering at intermediate locations. (Adapted from W. F. Marshall, *Curr. Biol.* 12:R185, 2002.)

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

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

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 Figure 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

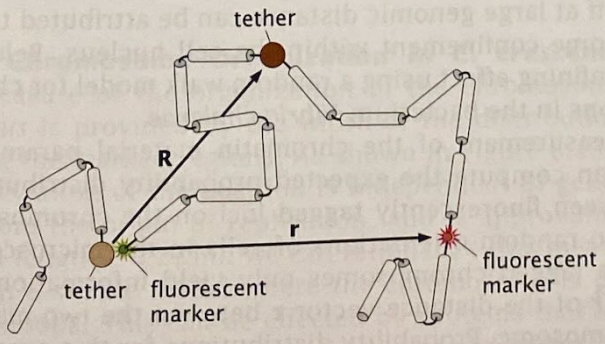


**Figure 8.10:** Simple configuration of a tethered chromosome. The two tethers are at fixed locations in space, and the second tether is at position $\mathbf{R}$ with respect to the first. The distribution of distances between the two fluorescent markers, one being at the same position on the chromosome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

$\mathbf{r}$ between the markers,

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

where $N'$ is now the number of Kuhn segments between the second tether and the second marker. This formula follows simply from Equation 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 Equations 8.34 and 8.35 is that chromosome 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 section, 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, distributions of distances between monomers are described by random walk statistics.

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 $Na = N_{\text{bp}}/\nu$, where $\nu$ is the linear packing density of DNA in chromatin. For example, two genomic loci $N_{\text{bp}} = 100\,\text{kb}$ apart would be separated by a 30 nm fiber, which is $100\,\text{kb}/(100\,\text{bp/nm}) = 1\,\mu\text{m}$ 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 10 times as large given the 10 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 $Na = N_{\text{bp}}/\nu$, the mean end-to-end distance squared can also be written as $\langle R^2 \rangle = N_{\text{bp}} a/\nu$. 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 4 are shown in Figure 8.11. The fit to the data yields an estimate of $a/\nu = 2\,\text{nm}^2/\text{bp}$, which is nothing but the initial slope of the linear portion of the data. The fact that the data level 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 *Vibrio cholerae*.

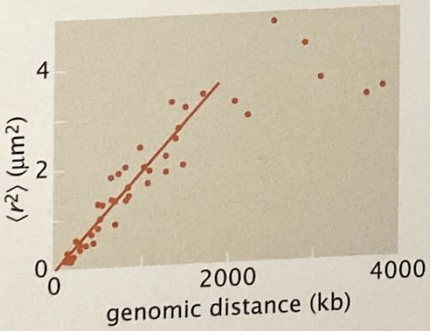


**Figure 8.11:** Physical distance between two fluorescently labeled loci on human chromosome 4 as a function of the genomic distance. The physical distance is measured in terms of the average squared distance between the two labels (dots). The curve corresponds to a linear fit as discussed in the text. (Adapted from G. van den Engh et al., *Science* 257:1410, 1992.)

With a measurement of the chromatin material parameter $a/\nu$ 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 Equations 8.34 and 8.35 by integrating out the angular variables $\theta$ and $\phi$ associated with the vector $\mathbf{r}$. This procedure yields

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

for the free-polymer case, and

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

when the polymer is tethered. Note 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*. In Figure 8.12, we show the distance distribution measured between two fluorescent tags, one placed near the so-called 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, Equation 8.36, while the tethered-polymer formula, Equation 8.37 does the job well.

The fit to the tethered-polymer distribution yields two quantities that characterize the model, namely, the mean-squared distance between the tether and the fluorescent marker at HML, $N'a^2 = 0.5\,\mu m^2$, and $R \approx 0.9\,\mu m$, 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 $a/\nu$ that characterizes chromatin structure. For that, measurements like those leading to Figure 8.11 for human chromosome 4 are needed.

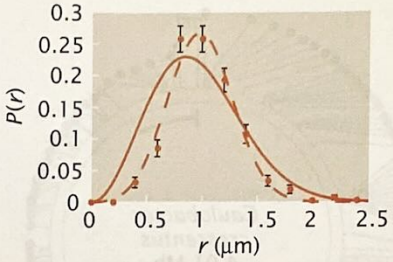


**Figure 8.12:** Statistics of yeast chromosome III. Distribution of distances between two fluorescent tags placed in proximity of the centromere and the HML region on yeast chromosome III. These two regions are separated by approximately 100 kb in genomic distance. The full line is a fit to the free-polymer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

### Chromosome Territories Have Been Observed in Bacterial Cells

Bacterial chromosomes used to be thought of as unstructured and random. This view has been seriously challenged by experiments that utilize fluorescent markers placed at different genomic locations, as shown in Figure 8.13. In this experiment, 112 different mutants of *Caulobacter crescentus* 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 Mb circular chromosome as a polymer loop confined to the cell.

**Estimate: Chromosome Organization in *C. crescentus***
Another measure of the organization of the chromosome in *C. crescentus* is provided by the width of the distribution of positions of the marked regions. As shown in Figure 8.13, the standard deviation of the position is independent of genomic distance from the origin of replication, and is approximately $0.2\,\mu m$ (cell length $L \approx 2\,\mu m$). We can rationalize this measurement within a simple model where the chromosome is partitioned into loops. This can be effected by proteins that make contact between different locations on the chromosome (H-NS

for its thermodynamic conjugate, the average polymer length,

$$\langle L\rangle = -\frac{\partial G}{\partial f} = Na\left[\coth\left(\frac{fa}{k_B T}\right) - \frac{k_B T}{fa}\right]. \tag{8.72}$$

The small-force limit, $fa/k_B T \ll 1$, in this case gives the same Hookean expression, $f = k\langle L\rangle$, as the one-dimensional freely jointed chain, except the effective spring constant is 3 times as large, $k = 3k_B T/L_{tot}a$. The same result follows from Equation 8.69. Not surprisingly, the two-dimensional version of the model, whether it be defined on a lattice or not, gives $k = 2k_B T/L_{tot}a$.

At large forces when the polymer approaches full extension, the force–extension formula, Equation 8.72, 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 Chapter 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 the function of a protein follows from its structure. So far, we have shown how the random walk model can be applied to nucleic acids. Proteins are polymers comprising 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 that 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 Hamiltonian walks), which are self-avoiding random walks that visit every site of the lattice, usually taken to be cubic, as depicted in Figure 8.26. 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.

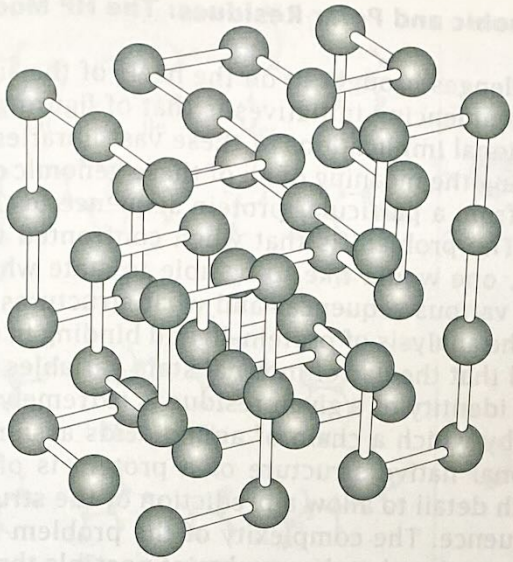

**Figure 8.26:** Compact polymer configuration on a $4 \times 4 \times 3$ cubic lattice. Each ball represents an amino acid. (Adapted from P. D. Thomas and K. A. Dill, *Proc. Natl Acad. Sci. USA* 93:11628, 2006.)

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 PDB reveals a systematic dependence of the protein's size on its mass. In particular, for globular proteins, the radius of gyration scales roughly with the cube root of the mass. The relation between the physical size of proteins and their sequence size is shown in Figure 8.27.

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 Figure 8.26), 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 that is even more expanded than that of a simple random walk.

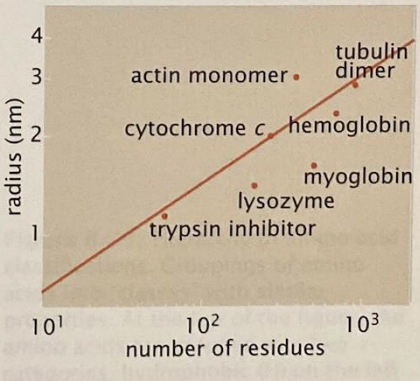


**Figure 8.27:** Scaling of protein size as a function of the number of amino acid residues. The line has a slope of 1/3, corresponding to a space-filling packing.