1 Solving Ligand-Receptor Multiple Ways

(a) In class, and then in the homework, we solved for the dynamics of mRNA production and degradation using the dynamics protocol. In this problem, we are going to use that analysis as a jumping off point for thinking about ligand-receptor binding problems. Imagine a situation in which we have a receptor fixed at some point in space as shown in the top right panel of Figure 1. Write a rate equation for the concentration of ligand-receptor pairs in terms of the concentration of ligands and receptors. Now, assume steady state and, given that equation, derive an expression for the dissociation constant

\[ K_d = \frac{[L][R]}{[LR]} \]  

in terms of the on and off rates. Make sure you explain the dimensions of your on and off rates and hence, the dimensions of \( K_d \).

(b) A second route to considering ligand-receptor interactions is to think of binding probabilistically with the probability that the receptor is occupied given by

\[ p_{\text{bound}} = \frac{[LR]}{[R] + [LR]} . \]  

Given the definition of the dissociation constant introduced in the previous part of the problem, find a simple expression for \( p_{\text{bound}}([L]) \) that is only a function of the concentration of ligand. (NOTE: for now, we are ignoring the subtlety that the amount of total ligand and free ligand are not actually the same, though in the case considered here with a single receptor we have somewhat finessed that point.) Make a plot of \( p_{\text{bound}}([L]) \) as a function of \([L]\) and comment on where \( K_d \) belongs on the axes. Later on in the course, we will solve this problem in yet another way, by using statistical mechanics.
(a) The rate equation for the concentration of ligand–receptor pairs is given by

$$\frac{d[L][R]}{dt} = k_{on}[L][R] - k_{off}[LR].$$

(3)

Before proceeding to the steady state solution, let us explore the dimensionality of the different terms. All concentrations have units of (1/volume). Therefore, the units of the left-hand side are 1/(volume × time). The two terms on the right-hand side should have matching units. From this, we can deduce the dimensions of $k_{on}$ and $k_{off}$, namely,

$$\text{dimension of } k_{on} = \frac{\text{volume}^{-1}\text{time}^{-1}}{(\text{volume}^{-1})^2} = \frac{\text{volume}}{\text{time}} = \frac{1}{\text{time} \times \text{concentration}},$$

(4)

$$\text{dimension of } k_{off} = \frac{\text{volume}^{-1}\text{time}^{-1}}{\text{volume}^{-1}} = \frac{1}{\text{time}}.$$  

(5)

At steady state, the time derivative of $[LR]$ must vanish. Equating the right-hand side of Eq. 3 to zero, we obtain

$$k_{off}[LR] = k_{on}[L][R],$$

$$k_{off} = \frac{[L][R]}{[LR]}.$$  

(6)

From the result above, we identify the dissociation constant as the ratio of the off- and on-rates, i.e. $K_d = k_{off}/k_{on}$, which has units of 1/volume or concentration.

(b) Using the expression for $K_d$ obtained above, we can use Eq. 6 to express the concentration of the ligand-receptor complex as $[LR] = [L][R]/K_d$. Substituting this form into the definition of $p_{\text{bound}}$ provided in the problem, we find

$$p_{\text{bound}} = \frac{[LR]}{[R] + [LR]} = \frac{[L][R]/K_d}{[R] + [L][R]/K_d} = \frac{[L]/K_d}{1 + [L]/K_d}.$$  

(7)

As can be seen from Figure 1, $K_d$ is the ligand concentration for which half of the receptors are bound. This result also follows directly by evaluating Eq. 7 at $[L] = K_d$. 

2
2 Cytoskeletal polymerization

In this problem, we explore how the same master equation approach presented in class to describe the distribution of mRNA molecules can be used to understand how biological polymers such as actin or tubulin are assembled. Specifically, we think of a polymer that can incorporate or lose monomers at one end, as shown in Figure 4. The rate of incorporation is $r$ and the rate with which monomers are lost is $\gamma$. Note that, unlike the mRNA case, the number of monomers that leave the filament in a time interval $\Delta t$ is given by $\gamma \Delta t$ and is independent of the filament length.

(a) The master equation of the probability of a filament having length $n$ at time $t$, $p(n,t)$, is given by

$$p(n, t + \Delta t) = p(n, t) + r\Delta tp(n-1, t) - r\Delta tp(n, t) - \gamma\Delta tp(n, t) + \gamma\Delta tp(n+1, t).$$

Justify why this is the master equation describing the length of a filament by carefully explaining the significance of each term in the master equation.

(b) Using Python, solve the master equation assuming $r = 1/s$ and $\gamma = 1.1/s$ and plot the resulting distribution for different time points. Assume that at time $t = 0$ there are no filaments by setting $p(n = 0, t = 0) = 1$ and $p(n > 0, t = 0) = 0$. Be careful about the boundary conditions! In particular, remember that you cannot have filaments with a negative number of monomers. Further, you cannot have Python calculate the probability of an arbitrarily long polymer. Hence, you will have to choose a maximum
length of the polymer for your calculation. In doing so, make sure that the probability of
having a filament of length \( n \) near your chosen maximum polymer length is very close to zero.

(c) Plot the length distribution in steady state. To make sure you’re in steady state,
demonstrate that the distribution does not change as you extend the total simulation time.

(d) In class we solved the mRNA distribution of the constitutive promoter in steady state.
Follow the same approach to now solve the steady-state distribution of filament lengths. To
make this possible, explore the master equation for the specific cases of \( n = 0, n = 1, n = 2 \)
in steady state and show that the distribution can be written as \( p(n) = p(0) \left( \frac{r}{\gamma} \right)^n \), which is
a geometric distribution.

(e) If you do the math, you can infer that \( p(0) = 1 - r/\gamma \) (you don’t need to show this). Now
that you have the full distribution, plot it together with your simulation results to show that
both approaches are consistent. If you want to compare your results to real data (and figure
out more realistic values of \( r/\gamma \), you can look at Figure 15.23 of PBoC2. Note that an actin
monomer is about 2.4 nm in length.

(a) The probability of finding a filament of length \( n \) at time \( t + \Delta t \) (this is, the fraction of
length \( n \) filaments / total number of filaments) is given by \( p(n, t + \Delta t) \). This probability is
equal to the probability at the previous time point \( p(n, t) \) adjusted by processes that occur
during \( \Delta t \) that change the number of filaments with \( n \) monomers.

- \( r \Delta t p(n-1, t) \) represents filaments that in the previous time point had \( n - 1 \) monomers
  and gained a monomer during \( \Delta t \), thus adding to \( p(n, t) \)

- \( \gamma \Delta t p(n+1, t) \) represents filaments that in the previous time point had \( n + 1 \) monomers
  and lost a monomer during \( \Delta t \), thus adding to \( p(n, t) \)

- \( r \Delta t p(n, t) \) represents filaments with \( n \) monomers in the previous time point that gained
  a monomer during \( \Delta t \), thus decreasing \( p(n, t) \)

- \( \gamma \Delta t p(n, t) \) represents filaments with \( n \) monomers in the previous time point that lost
  a monomer during \( \Delta t \), thus decreasing \( p(n, t) \)

(b)(c) The plot of the filament length distribution at different time points is shown
in Figure 2. Since the distribution does not change significantly for three consequent time
points, we can assume that it has reached the steady state. The Python code used to generate
this plot is posted on the course website.

(d) We begin by evaluating the master equation at \( n = 0 \).

\[
p(0, t + \Delta t) = p(0, t) + r \Delta t p(0 - 1, t) - r \Delta t p(0, t) - \gamma \Delta t p(0, t) + \gamma \Delta t p(0 + 1, t)
\]  
(9)
We now note that at steady state, by definition $p(0, t + Δt) = p(0, t)$. Furthermore, $p(0 − 1, t) = 0$ (there are no −1 monomer length filaments) and $γΔtp(0, t) = 0$ (because there is no degradation from a nonexistent polymer). Additionally, because at steady state the function does not change with time, we keep $n$ as the only variable. With all of this in mind, the expression becomes:

$$0 = −rp(0) + γp(1) \quad (10)$$

and we can solve for $p(1)$

$$p(1) = \frac{r}{γ}p(0) \quad (11)$$

Let’s continue with $n = 1$. Invoking the same arguments as before, we get:

$$0 = rp(0) − rp(1) − γp(1) + γp(2) \quad (12)$$

By invoking our results that $γp(1) = rp(0)$, obtained above, we can reduce the expression to:

$$0 = −rp(1) + γp(2) \quad (13)$$

$$p(2) = \frac{r}{γ}p(1) = \left(\frac{r}{γ}\right)^2 p(0) \quad (14)$$

For $p(2)$, the master equation at steady state leads to:

$$0 = −rp(2) + γp(3) \quad (15)$$

$$p(3) = \frac{r}{γ}p(2) = \left(\frac{r}{γ}\right)^3 p(0) \quad (16)$$

Thus, the probability distribution associated with the master equation is $p(n) = p(0) \left(\frac{r}{γ}\right)^n$.

(e) This analytical solutions is plotted along with the numerical solution (part c) in Figure 3.
In this problem we go beyond the calculation on mRNA production we did in class, and think about how transcription and translation shape the protein-to-mRNA ratio inside cells.

(a) In class, we described the temporal evolution of the number of mRNA molecules using the equation

\[ m(t + \Delta t) = m(t) + r_m \Delta t - \gamma_m m(t) \Delta t. \]  (17)

Here, \( m(t) \) is the number of mRNA at time \( t \), \( r_m \) is the rate of mRNA production, and \( \gamma_m \) is the mRNA decay rate. Write the corresponding equation for the number of protein molecules given a rate of protein production per mRNA of \( r_p \) and a protein decay rate \( \gamma_p \). Make sure to incorporate the fact that the number of mRNA molecules present will determine how many proteins are produced in a time interval \( \Delta t \).
(b) Calculate the ratio of protein to mRNA in steady state, $p_{ss}/m_{ss}$ and show that it is given by $r_p/\gamma_p$. Find typical values for the various model parameters in E. coli and estimate the ratio of proteins to mRNA molecules. How do your numbers compare to those measured in Figure 3C of Taniguchi et al., which is provided on the course website?

We can also obtain this protein-mRNA ratio in the context of fruit flies.

(c) Using flies with different dosages of Bicoid-GFP, Petkova et al. measured the relation between the number of bicoid mRNA molecules deposited by the mother, and the resulting number of Bicoid proteins. Read their paper (available on the course website) and write a short paragraph about how their Figure 3 is generated.

(d) The paper by Drocco et al. (available on the course website) uses a photoactivatable fluorescent protein to measure the lifetime of the Bicoid protein. Read the paper (available on the course website) and explain the technique in one paragraph. You might find it useful to draw a schematic plot such as shown in Figure 1f of the paper.

(e) Assuming that Bicoid-GFP is in steady state, use what you learned about $r_p$ and $\gamma_p$ for the Bicoid protein in order to calculate its protein-mRNA ratio $r_p/\gamma_p$.

(a) The temporal evolution of the number of protein molecules is described by the equation

$$p(t + \Delta t) = p(t) + r_p m(t) \Delta t - \gamma_p p(t) \Delta t.$$ 

Here, $p(t)$ is the number of protein molecules at time $t$, $r_p$ is the rate of protein production per mRNA molecule, and $\gamma_p$ is the protein degradation rate. Note that the amount of protein produced in time $\Delta t$ depends on the number of mRNA molecules present, while the amount of protein molecules that are degraded in time $\Delta t$ depends on the amount of proteins present.

We can manipulate the discrete version of the equation describing the temporal evolution of the number of proteins to arrive at a continuous time version as follows:

$$p(t + \Delta t) - p(t) = r_p m(t) \Delta t - \gamma_p p(t) \Delta t \quad \Rightarrow \quad \frac{p(t + \Delta t) - p(t)}{\Delta t} = r_p m(t) - \gamma_p p(t) \quad \Rightarrow \quad \frac{dp}{dt} = \lim_{\Delta t \to 0} \frac{p(t + \Delta t) - p(t)}{\Delta t} = r_p m(t) - \gamma_p p(t),$$

(b) In steady state, $p(t) = p_{ss}$ is constant and $\frac{dp}{dt} = 0$ such that

$$\frac{dp}{dt} = r_p m_{ss} - \gamma_p p_{ss} = 0 \quad \Rightarrow \quad \frac{p_{ss}}{m_{ss}} = \frac{r_p}{\gamma_p}.$$ 

7
Note that we have used the fact that \( m(t) = m_{ss} \) must also be constant for the number of proteins to be constant.

The average translation initiation rate in \( E. \ coli \) is 5 min\(^{-1}\) mRNA\(^{-1}\) (BNID 112001) which corresponds to an average production rate of 5 proteins/min/mRNA. To find the average protein degradation rate for \( E. \ coli \), we can use the fact that 1\% of the bulk protein pool is degraded in an hour (BNID 109924). Because the rate of protein degradation is small, we can infer that \( \gamma_p \approx 1/100 = .01\) h\(^{-1}\) from the fact that 1 in 100 proteins in the bulk protein pool is degraded in an hour.

**Note:** For a more general approach with \( \gamma_p \) of arbitrary size, we convert this to a degradation rate as follows: Assuming all of the bulk protein pool decays via simple first-order degradation and that no protein is produced, the amount of protein at time \( t \) is given by

\[
p(t) = p_0 e^{-\gamma_p t}.
\]

We can mathematically express the fact that 1\% of the bulk protein pool is degraded in 1 hour as

\[
p(t = 1 \text{ h}) = .99 \times p_0.
\]

Plugging into our expression for \( p(t) \) in the case where there is no production and there are \( p_0 \) proteins at \( t = 0 \), we find

\[
p(t = 1 \text{ h}) = .99 \times p_0 = p_0 e^{-\gamma_p \cdot 1 \text{ h}}
\]

\[
\Rightarrow .99 = e^{-\gamma_p \cdot 1 \text{ h}}
\]

\[
\Rightarrow \ln (.99) = -\gamma_p \cdot 1 \text{ h}
\]

\[
\Rightarrow \gamma_p = -\frac{\ln (.99)}{1 \text{ h}} \approx .01005 \text{ hour}^{-1} \approx .01 \text{ hour}^{-1}
\]

Note that both approaches gave us the same value for \( \gamma_p \) in this case.

The average protein half-life in \( E. \ coli \) is very long such that the effect of protein degradation is dominated by the dilution rate resulting from cell division. When a cell divides the concentrations of its proteins are cut in half. The effect of protein degradation is therefore negligible when compared to dilution. Assuming a doubling time of \( \sim 30 \) minutes (BNID 109057), we can find \( \gamma_p \):

\[
\frac{1}{2} = e^{-\gamma_p \cdot 30 \text{ min}} \Rightarrow \ln \left( \frac{1}{2} \right) = -\gamma_p \cdot 30 \text{ min} \Rightarrow \gamma_p = \frac{1}{30 \text{ min}} \ln 2 \approx .02 \text{ min}^{-1} = 1.4 \text{ hour}^{-1},
\]

where we have used the same approach as described in detail in the box above. Plugging these values for \( r_p \) and \( \gamma_p \) into our expression for the ratio of proteins to mRNA molecules gives:

\[
p_{ss}/m_{ss} = \frac{r_p}{\gamma_p} = \frac{5 \text{ proteins} \cdot \text{min}^{-1} \cdot \text{mRNA}^{-1}}{.02 \text{ min}^{-1}} = 250 \text{ proteins/mRNA}.
\]
Note that if we had used $\gamma_p = 0.01\ \text{hour}^{-1} \approx 2 \times 10^{-4}\ \text{min}^{-1}$, we would have found

$$p_{ss}/m_{ss} = \frac{r_p}{\gamma_p} = \frac{5\ \text{proteins} \cdot \text{min}^{-1}\text{mRNA}^{-1}}{2 \times 10^{-4}\ \text{min}^{-1}} = 25000\ \text{proteins}/\text{mRNA},$$

which is 100-fold larger than our estimate of the ratio using the dilution rate.

Figure 3C from Taniguchi et al. shows a scatter plot of mRNA copy number against protein copy number. Notice that each dot is a gene and that the measurements were done in bulk, meaning they represent the average in a population of cells. There is a clear trend in the data that is consistent with $p_{ss}/m_{ss} = r_p/\gamma_p \approx 10^3$, which is about 4 times higher than our estimate.

(c) To get the relationship between protein and mRNA level, Petkova, et al. have engineered transgenic flies that express Bcd-GFP at different levels. Then, they quantified differences in protein concentration by comparing average nuclear Bcd-GFP intensities across AP positions between a given transgenic test line and a reference line. Meanwhile, the maternally deposited bicoid-gfp mRNA expression is measured by qPCR.

(d) Bcd-Dronpa has bistable bright and dark states, conversions between which are rapid and inducible in living and fixed material. As shown in Figure 5(A), Bcd-Dronpa can be converted between dark and bright states by illuminating the embryo with either 496nm light or 405nm light. Using a method that employs repeated photoswitching of the Bcd-Dronpa fusion (Figure 5(C)), they can measure and modify the lifetime of Dronpa-Bcd in living Drosophila embryos.

(e) We know that the translation rate of Bcd protein is about 2 proteins/mRNA·min (Petkova et al.) and the degradation rate is approximately 0.028 protein/min (Drocco et al.). Based on the equation we derived, we can calculate the protein-mRNA ratio of Bcd protein

$$p_{ss}/m_{ss} = \frac{r_p}{\gamma_p} = \frac{2\ \text{proteins} \cdot \text{min}^{-1}\text{mRNA}^{-1}}{0.028\ \text{min}^{-1}} \approx 70\ \text{proteins}/\text{mRNA}.$$

This is similar to the measured ratio from Petkova et al. ($p_{ss}/m_{ss} \approx 100$).

4 Phase diagram for the logistic equation

In Homework 3, we solved the logistic equation numerically. This equation can describe the saturation of a bacterial culture by accounting for a limited food supply

$$\frac{dN}{dt} = rN \left(1 - \frac{N}{K}\right), \quad (18)$$
Figure 5: Description of Dronpa-Bcd construct and degradation measurement. (A) Confocal images of in vivo Dronpa-Bcd-expressing embryos with Dronpa-Bcd predominantly in the bright or dark state. (B) Confocal image of a fixed Dronpa-Bcd embryo with Dronpa predominantly converted to the dark state except an inscription produced by a targeted 405-nm reactivation pulse. (C) Schematic of degradation measurement, beginning with a Dronpa-Bcd population converted to the dark state

where \( N \) is the number of cells, \( r \) is the growth rate, and is the carrying capacity or maximum population size. Note that this equation can also be written as

\[
\frac{dN}{dt} = rN - \frac{rN^2}{K}.
\] (19)

Here, we can identify a “cell production” term and “cell destruction” term.

Like we did for the case of the constitutive promoter, plot a phase diagram where both production and destruction terms are plotted (either by hand or in Python). Use this plot to graphically show that there are two stable points at which the production and destruction are balanced out.

Here we have two rates. Both “cell production” and “cell destruction” rates are functions of the number of cells, \( N \). The production term \( f(N) = rN \) should be a straight line with a slope of \( r \) and the destruction term, \( g(N) = \frac{rN^2}{K} \), should be a quadratic function of \( N \). Note that when \( N = 0 \) and \( N = K \), the production and destruction terms are equal which indicates that these are steady state solutions of the logistic growth equation.

Assuming \( r = 10 \) cells/min and \( K = 1000 \), we can plot production and destruction rates as shown in Figure 4. We can see that there are two intersections between production and destruction term which indicates that there are two stable points. The Python code used to generate the figure is posted on the course website.
5 Random walks and biological polymers

Physicists know how to solve just a handful of problems. Fortunately, many dissimilar phenomena in physics and biology alike can be mapped onto such problems for which we know a solution. Here, we explore the mathematical connection between diffusion and the spatial arrangement of polymers such as DNA, actin, and microtubules.

(a) Read the introduction to Section 8.2 of PBoC (“Random Walk Models of Macromolecules View Them as Rigid Segments Connected by Hinges”) to learn more about how polymers can be thought of as chains of connected rigid segments. Pay close attention to Figures 8.1 and 8.2. Here, the Kuhn length $a$ is defined as the length of the segments. Look up the Kuhn length for DNA, actin, and microtubules in order to get a feeling for these polymers. Note that you might find reference to the persistence length $\xi_p = a/2$ instead of the Kuhn length.

(b) Now, think of a polymer chain of $N$ segments in 1D. As shown in Figure 8.3 of PBoC each segment can either be pointing to the right of to the left. Given $n_R$ and $n_L$ segments pointing to the right and left, respectively, the position of the end of the chain is given by $L = (n_R - n_L) a$. Map this problem onto the diffusion problem we solved in class where we calculated the $\langle x \rangle$ and $\langle x^2 \rangle$ of a random walker that start at the origin shown in Figure 3. To make this possible, note that each segment can be randomly pointing to the left or right. In particular, calculate $\langle n_R - n_L \rangle$ and $\langle (n_R - n_L)^2 \rangle$ and show that the size of the polymer is given by

$$\text{size} \approx \sqrt{\langle L^2 \rangle} = a\sqrt{N} \tag{20}$$

by repeating graphical the derivation we did in class.

(c) Think of the size of the polymer you derived in (b) as the linear dimension of the blob the polymer will make on a surface such as shown in the figures below. Use the derived formula
to estimate the genome length (in µm and bp) of the bacteriophage T2 shown in Figure 1.16 of PBoC and of the E. coli in Figure 8.5 of PBoC. How well did your estimate do?

All relevant figures from PBoC can also be found in Figures 4 and 5 below.

(a) DNA: $\xi_p \approx 50$ nm $\implies a = 100$ nm (BNID 103112)
actin: $\xi_p \approx 10$ µm $\implies a = 20$ µm (BNID 106830)
microtubules: $\xi_p \approx 1.4$ mm $\implies a = 2.8$ mm (BNID 105534)

(b) Here we use the same way of thinking we employed to derive a formula for the diffusion of particles in 1D. By following the same logic, we hope to be able to find a similar law for the size of a polymer chain. We start by imagining the growth of a polymer. Each time a segment is added it “flips a coin” to decide whether it grows to the left or to the right.

If polymer growth can be captured by the same law as diffusion, i.e. if the growing polymer behaves like a random walk, then we expect:

$$\langle n_R - n_L \rangle = 0$$
$$\langle (n_R - n_L)^2 \rangle = N$$

We can check all possible combinations of “coin flips” ($N$), for cases where we have $N = 1, 2,$ and 3 to confirm that the expected expressions for $\langle n_R - n_L \rangle$ and $\langle (n_R - n_L)^2 \rangle$ are correct. All possible individual paths are enumerated for $N = 1, 2, 3$ in Tables 1, 2, and 3, respectively:

<table>
<thead>
<tr>
<th>$L = a(n_R - n_L)$</th>
<th>$n_R - n_L$</th>
<th>$n_R$</th>
<th>$n_L$</th>
<th>paths</th>
<th># of paths</th>
<th>probability per path</th>
<th>total probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>right</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$-a$</td>
<td>-1</td>
<td>0</td>
<td>1</td>
<td>left</td>
<td>1</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
</tr>
</tbody>
</table>

Table 1: Probabilities for all possible values of $\Delta n$ given $N = n_R + n_L = 1$.

<table>
<thead>
<tr>
<th>$L = a(n_R - n_L)$</th>
<th>$n_R - n_L$</th>
<th>$n_R$</th>
<th>$n_L$</th>
<th>paths</th>
<th># of paths</th>
<th>probability per path</th>
<th>total probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>right-right</td>
<td>1</td>
<td>$(\frac{1}{2})^2 = \frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>right-left</td>
<td>2</td>
<td>$(\frac{1}{2})^2 = \frac{1}{4}$</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>-2a</td>
<td>-2</td>
<td>0</td>
<td>2</td>
<td>left-left</td>
<td>1</td>
<td>$(\frac{1}{2})^2 = \frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

Table 2: Probabilities for all possible values of $\Delta n$ given $N = n_R + n_L = 2$.

A summary of this information is given in the first three rows of Table 4, from which we can deduce a general formulas for $\langle n_R - n_L \rangle$ and $\langle (n_R - n_L)^2 \rangle$ for a polymer with $N$ segments (row 4 of Table 4).
\[ L = a(n_R - n_L) \]

\[ (n_R - n_L) = \sum_{n_R=0}^{N} (n_R - n_L) \cdot P_N(n_R) \]

\[ \langle (n_R - n_L)^2 \rangle = \sum_{n_R=0}^{N} (n_R - n_L)^2 \cdot P_N(n_R) \]

\[ \langle |L|^2 \rangle = \frac{1}{a^2} \langle L^2 \rangle \]

\[ N = \frac{1}{a^2} \langle L^2 \rangle \]

\[ \langle L^2 \rangle = a^2 N \]
Thus the size of the polymer (or how far is its end from its beginning in 1D) can be approximated by:

\[ L \approx \sqrt{\langle L^2 \rangle} = a\sqrt{N} \]  

(24)

(c) From figures, we can estimate that \( L_{\text{phage}} \approx 1 \mu\text{m} \). The length of the E. coli genome is \( L_{\text{E. coli}} \approx 10 \mu\text{m} \). The Kuhn length of DNA is \( a_{\text{DNA}} = 100 \text{ nm} \).

Using the formula we derived above we can solve for the number of segments \( N \) in the T2 bacteriophage genome:

\[ N_{\text{phage}} = \left( \frac{L_{\text{phage}}}{a_{\text{DNA}}} \right)^2 = \left( \frac{1 \mu\text{m}}{100 \text{ nm}} \right)^2 = \left( \frac{10^3 \text{ nm}}{100 \text{ nm}} \right)^2 = 100 \text{ segments} \].

The length of the fully-stretched genome is simply the number of segments multiplied by the length of each segment, which we can get from the Kuhn length of DNA:

Length of phage genome (\( \mu\text{m} \)) \( \approx \) 100 segments \( \times 100 \frac{\text{nm}}{\text{segment}} = 10^4 \text{ nm} = 10 \mu\text{m} \).

To convert this to base pairs, we simply divide this number by the average height for a base pair, which is .34 nm/bp = 3.4 \( \times \) 10\(^{-4} \) \( \mu\text{m} \):

Length of phage genome (bp) \( = 10 \mu\text{m} \times \frac{1\text{bp}}{3.4 \times 10^{-4} \mu\text{m}} \approx 3 \times 10^4 \text{ bp} \).

This is in good agreement with the size of the bacteriophage T7 genome, which is \( \approx 4 \times 10^4 \text{ bp} \) (BNID 106939), and the size of the bacteriophage lambda genome, which is \( \approx 5 \times 10^4 \text{ bp} \) (BNID 105770). We can assume these genomes are fairly similar in size to that of the T2 bacteriophage.

Taking a similar approach for \( E.\text{coli} \):

\[ L_{\text{E. coli}} \approx \sqrt{\langle L_{\text{E. coli}}^2 \rangle} = a_{\text{DNA}}\sqrt{N_{\text{E. coli}}} \]

\[ N_{\text{E. coli}} = \left( \frac{L_{\text{E. coli}}}{a_{\text{DNA}}} \right)^2 = \left( \frac{10 \mu\text{m}}{100 \text{ nm}} \right)^2 = \left( \frac{10^4 \text{ nm}}{100 \text{ nm}} \right)^2 = 10^4 \text{ segments} \].

Length of \( E.\text{ coli} \) genome (\( \mu\text{m} \)) \( \approx 10^4 \text{ segments} \times 100 \frac{\text{nm}}{\text{segment}} = 10^6 \text{ nm} = 10^3 \mu\text{m} \).

Length of \( E.\text{ coli} \) genome (bp) \( = 10^3 \mu\text{m} \times \frac{1\text{bp}}{3.4 \times 10^{-4} \mu\text{m}} \approx 3 \times 10^6 \text{ bp} \).

This is good agreement with the measured size of the \( E.\text{ coli} \) genome, which is \( \approx 5 \times 10^6 \text{ bp} \) (BNID 100269).
Returning to 4(b): Formal proofs showing that \( \langle n_R - n_L \rangle = 0 \) and = \( \langle (n_R - n_L)^2 \rangle = N \)

Note that \( n_R + n_L = N \) as every one of the \( N \) segments must either point to the right or to the left. We can define a probability distribution \( P(n_R) \) for \( n_R = 0, \ldots, N \):

\[
P(n_R) = \left( \frac{1}{2} \right)^n_R \left( \frac{1}{2} \right)^{n_L} \frac{N!}{n_R! \ n_L!} = \left( \frac{1}{2} \right)^n_R \left( \frac{1}{2} \right)^{N-n_R} \frac{N!}{n_R! \ (N-n_R)!} \quad \text{(because } n_L = N - n_R) \]

\[
\implies P(n_R) = \left( \frac{1}{2} \right)^N \frac{N!}{n_R! \ (N-n_R)!}.
\]

The sum of the probabilities \( P(n_R) \) for \( n_R = 0, \ldots, N \) must be equal to 1:

\[
\sum_{n_R=0}^{N} P(n_R) = \sum_{n_R=0}^{N} \left( \frac{1}{2} \right)^n_R \frac{N!}{n_R! \ (N-n_R)!} = \left( \frac{1}{2} \right)^N \sum_{n_R=0}^{N} \frac{N!}{n_R! \ (N-n_R)!} = 1,
\]

\[
\implies \sum_{n_R=0}^{N} \frac{N!}{n_R! \ (N-n_R)!} = 2^N.
\]

Since \( n_L = N - n_R \), then

\[
\langle n_R - n_L \rangle = \langle n_R - (N - n_R) \rangle = \langle 2n_R - N \rangle = 2\langle n_R \rangle - N.
\]
We now calculate \( \langle n_R \rangle \):

\[
\langle n_R \rangle = \sum_{n_R=0}^{N} n_R P(n_R)
\]

\[
= \left( \frac{1}{2} \right)^N \frac{N!}{n_R! (N - n_R)!} \sum_{n_R=0}^{N} n_R (n_R - 1)! (N - n_R)!
\]

\[
= \left( \frac{1}{2} \right)^N \sum_{n_R=1}^{N} \frac{N!}{n_R(1)!(N - n_R)!} (n_R - 1)!
\]

\[
= \left( \frac{1}{2} \right)^N \sum_{m=0}^{N-1} \frac{N!}{m!(N-(m+1))!} (N-m)! (N-1)! (N-(N-1)-m)!
\]

\[
= N \left( \frac{1}{2} \right)^N \sum_{m=0}^{N-1} \frac{N!}{m!(N-1)!(N-1)-m)!}
\]

\[
= N \left( \frac{1}{2} \right)^N 2^{N-1}
\]

\[
= \frac{N}{2}
\]

Plugging this into our expression for \( \langle n_R - n_L \rangle \):

\[
\langle n_R - n_L \rangle = 2\langle n_R \rangle - N = 2\frac{N}{2} - N = N - N = 0
\]

Note that we could have arrived at this relationship without doing any calculations by noticing that it must be the case that \( \langle n_R \rangle = \langle n_L \rangle \). The probability distribution \( P(n_R) \) could be written equivalently as both a function of \( n_R \) and \( n_L \) as follows:

\[
P(n_R) = P'(n_R, n_L) = \left( \frac{1}{2} \right)^{n_R} \left( \frac{1}{2} \right)^{n_L} \frac{N!}{n_R! n_L!}
\]

\[
= \left( \frac{1}{2} \right)^{n_R+n_L} \frac{N!}{n_R! n_L!}
\]

This probability function \( P'(n_R, n_L) \) is symmetric, which means that it is unchanged by swapping \( n_R \) and \( n_L \), i.e. \( P'(n_R, n_L) = P'(n_L, n_R) \). It follows that, for any function \( f(n) \),

\[
\langle f(n_R) \rangle = \langle f(n_L) \rangle
\]

for the probability distribution \( P'(n_R, n_L) = P(n_R) \). Now we can move on to calculating \( \langle (n_R - n_L)^2 \rangle \). First we can manipulate this expression so it is written entirely as a function of \( n_R \) and constants by plugging in \( n_L = N - n_R \) and expanding:

\[
\langle (n_R - n_L)^2 \rangle = \langle (n_R - (N - n_R))^2 \rangle = \langle (2n_R - N)^2 \rangle = 4\langle n_R^2 \rangle - 4N \langle n_R \rangle + N^2 = 4\langle n_R^2 \rangle - N^2
\]
We now calculate $\langle n_R^2 \rangle$:

$$\langle n_R^2 \rangle = \sum_{n_R=0}^{N} n_R^2 P(n_R) = \sum_{n_R=0}^{N} n_R^2 \left( \frac{1}{2} \right)^N \frac{N!}{n_R! (N-n_R)!}$$

$$= \left( \frac{1}{2} \right)^N \sum_{n_R=0}^{N} n_R(n_R - 1 + 1) \frac{N!}{n_R! (N-n_R)!}$$

$$= \left( \frac{1}{2} \right)^N \sum_{n_R=0}^{N} [n_R(n_R - 1) + n_R] \frac{N!}{n_R! (N-n_R)!}$$

$$= \left[ \left( \frac{1}{2} \right)^N \sum_{n_R=0}^{N} n_R(n_R - 1) \frac{N!}{n_R! (N-n_R)!} \right] + \left[ \left( \frac{1}{2} \right)^N \sum_{n_R=0}^{N} n_R \frac{N!}{n_R! (N-n_R)!} \right]$$

$$= \left[ \left( \frac{1}{2} \right)^N \sum_{n_R=2}^{N} n_R(n_R - 1) \frac{N!}{n_R! (N-n_R)!} \right] + \langle n_R \rangle$$

$$= \left[ \left( \frac{1}{2} \right)^N \sum_{n_R=2}^{N} \frac{N!}{(n_R - 2)! (N-n_R)!} \right] + \frac{N}{2}$$

$$= \left[ \left( \frac{1}{2} \right)^N \sum_{m=0}^{N-2} \frac{N!}{m! (N-(m + 2))!} \right] + \frac{N}{2}$$

$$= \left[ \left( \frac{1}{2} \right)^N \sum_{m=0}^{N-2} \frac{N(N-1)(N-2)!}{m! ((N-2) - m)!} \right] + \frac{N}{2}$$

$$= \left[ \frac{N(N-1)}{2} \right] \left( \frac{1}{2} \right)^N \sum_{m=0}^{N-2} \frac{(N-2)!}{m! ((N-2) - m)!} + \frac{N}{2}$$

$$= \left[ \frac{N(N-1)}{2} \right] \left( \frac{1}{2} \right)^N 2^{N-2} + \frac{N}{2}$$

$$= \frac{N(N-1)}{4} + \frac{N}{2}$$

$$= \frac{N^2}{4} + \frac{N}{4}.$$ 

Now we plug into our expression for $\langle (n_R - n_L)^2 \rangle$:

$$\langle (n_R - n_L)^2 \rangle = 4\langle n_R^2 \rangle - N^2 = 4 \left( \frac{N^2}{4} + \frac{N}{4} \right) - N^2 = N^2 + N - N^2 = N$$