# Elements of Biochemistry and Molecular Biology

Robert Schleif

Biology and Biophysics Departments

Johns Hopkins University

September 14, 2023

# Preface

This book is written for advanced undergraduates or first year graduate students in Biology, Biochemistry, Chemistry, Physics, and Biophysics who desire a deeper understanding of some of the physical laws and principles that underlie much of Biology. The objective of the material is to help students develop a deeper understanding and an intuitive understanding of the principles involved. Ideal preparation for appreciating what the book offers would be completion of college level introductory courses in Calculus, Physics, Chemistry, and either Biology or Biochemistry. The book uses no mathematics beyond elementary calculus. Students' appreciation of and understanding of some of the principles and phenomena will be assisted by their simulating phenomena on spreadsheets as developed and described in the text. These examples are described in sufficient detail that the spreadsheets may be constructed and utilized by novices. Throughout, interconnections between topics discussed in the book are noted as are connections to material not covered by the book. Approaches and strategies used in scientific research are both mentioned in passing and are inherent in the construction of much of the presentation.

# Contents

# 1. Populations

**Unrestrained Growth is Exponential**

Suppose that a population of cells has been growing for multiple doublings with no effective change in the growth conditions like limitation of a required component in the growth medium or a temperature change. During such a period of balanced growth, the average composition of the cells does not change. Under these conditions, the rate of growth of the numbers of cells or of any quantity like the amount of total protein in the culture or of any particular enzyme will be proportional to the number of cells present. That is

$$\frac{dN(t)}{dt} \propto N(t)$$

Call the constant of proportionality the growth rate $\mu$ so that

$$\frac{dN(t)}{dt} = \mu \times N(t)$$

Since we want $N(t)$, solve this simple differential equation to eliminate $dN(t)/dt$. The initial conditions will be that there were $N_0$ cells at zero time, that is, $N(0) = N_o$. The equation can be solved by rearranging and integrating.

$$\int_{N_0}^{N(t)} \frac{dN(t)}{N(t)} = \int_0^t \mu dt$$

Giving

$$\ln\big(N(t)\big) - \ln(N_0) = \mu t$$

that can be rewritten as

$$N(t) = N_0 e^{\mu t}$$

The result is that freely growing populations increase exponentially. For purposes of experimental measurements, such populations are highly desirable because the properties of the population do not change with time and should be readily reproducible.

Sometimes it is convenient to express cell numbers in terms of doubling times $T_d$ instead of the growth rate $\mu$. The relationship between $\mu$ and $T_d$ can easily be obtained in the idealized situation in which cells divide only when they reach the age $a_0$. In this case, of course, $T_d = a_0$. Since

$$N(t) = N_0 e^{\mu t}$$

$$N(T_d) = N_0 e^{\mu T_d}$$

but

$$N(T_d) = 2N_0$$

so

$$2N_0 = N_0 e^{\mu T_d}$$

and dividing by $N_0$ and taking the natural log gives

$$ln2 = \mu T_d$$

or

$$\mu = \frac{ln2}{T_d}$$

**Protein Accumulation and Dilution Rates**

Suppose that a population has been in balanced exponential growth for a long period of time before zero time. Let P represent the amount of any particular cellular protein and let N represent the number of cells at zero time as shown in Table 1-1.

Table 1-1

| Doublings | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Total N Present | N | 2N | 4N | 8N | 16N |
| Total P Present | P | 2P | 4P | 8P | 16P |
| P Synthesized in Doubling | | P | 2P | 4P | 8P |

As the cells double in number, P also doubles, so that the average amount of P per cells remains constant. Note however, that in the time to grow from N to 2N cells, the amount of P that is synthesized is only P. Similarly, in the time required to grow from 2N to 4N, an amount 2P is synthesized.

Two extremes in the induction and behavior of P are of interest. In the first, P has been synthesized at its normal rate since the very early time, but at zero time, its synthesis stops. Thereafter, the amount of P remains constant while cell numbers continue to double. The specific activity of the protein is defined as the amount of the protein per cell. Table 1-2 shows that with each doubling of the cells, the specific activity is halved.

Table 1-2

| Doublings | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Total N Present | N | 2N | 4N | 8N | 16N |
| Total P Present | P | P | P | P | P |
| Specific Activity of P | P/N | P/2N | P/4N | P/8N | P/16N |

Suppose, instead, that the protein's synthesis was induced only at zero time. Then, as shown in Table 1-3, an amount P will be synthesized as the cells double from N to 2N. Thus, the specific activity of P after one doubling is half of the value it would have had if it had been induced much earlier, and after two doublings it is three quarters of that value. These examples show that following induction, the level per cell of an induced protein increases until its rate of synthesis equals the rate of dilution caused by cell growth. These examples also illustrate why cells must be in balanced exponential for a significant number of doublings to ensure that the levels of all important cell constituents are present at their steady state levels.

Table 1-3

| Doublings | 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| Total N Present | N | 2N | 4N | 8N | 16N |
| P Synthesized in Doubling | | P | 2P | 4P | 8P |
| Total P Present | 0 | P | 3P | 7P | 15P |
| Specific Activity of P | 0 | P/2N | 3P/4N | 7P/8N | 15P/16N |

**Age Distributions**

A culture of cells in balanced exponential growth contains a greater number of young cells than old cells. To see why this is so, consider the possibility that the age distribution is uniform as shown in Figure 1-1. In this figure, it is assumed that all the cells divide exactly when they reach

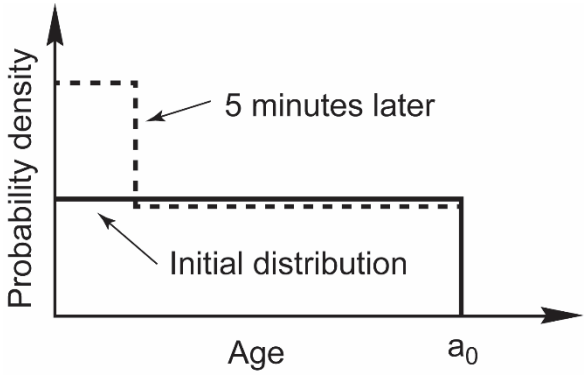



Figure 1-1 A uniform distribution of cell ages and the distribution of ages five minutes later.

age $a_0$. Now consider the distribution of ages five minutes later. All those cells that were within five minutes of $a_0$ will have reached an age of $a_0$ and divided into zero age cells. The important point is that each of these dividing cells divides into two cells. Therefore, the distribution of ages is no longer uniform and the assumption that cells in balanced exponential growth have a uniform age distribution is incorrect. Further, whatever the shape of the age distribution is, it will be higher for young cells than for old cells.

To determine the shape of the age distribution for cells in balanced exponential growth with growth rate $\mu$, let $N(a,t)$ represent the number of cells of age $a$ at time $t$. As above, consider the idealized case where all cells divide when they reach an age $a_0$. Since the cells are in balanced exponential growth and no cell divides until it reaches an age $a_0$, all cells of age $a$ were of zero age $a$ minutes earlier, that is, $N(a,t) = N(0,t-a)$. Since the population is in balanced exponential growth, numbers of cells increase as $e^{\mu t}$. Therefore, $N(0,t-a) = N(0,t)e^{-\mu a}$ which is equivalent to $N(a) = N(0)e^{-\mu a}$. As a check on the manipulations, substituting the value of $\mu$ in terms of the doubling time as determined in the previous section, gives the expected $N(a_0) = N(0)/2$ Between zero age and the age at which division occurs, the number of cells decreases exponentially at a rate equal to the growth rate of the population.

**Nonincreasing Populations**

The previous sections dealt with balanced exponential growth. Some natural populations don't spend much time in such a growth phase. Instead, their populations seem relatively static, for example, the year-to-year level of rabbits on a farm or the year-to-year level of a species of bird or insect. Generally, these populations are self-limiting. As the population increases, competition for food, or another critical resource or the level of predators reduces, limits, or entirely prevents further growth, and alternatively, when the population decreases, the reproduction rate increases. This is a form of negative feedback. If, without feedback, the population level in the nth generation is proportional to $N_n$ where the maximum of $N_n$ is 1 and the reproduction rate is r, then

$$N_{n+1} = r \times N_n$$

Negative feedback can be introduced in multiple forms. One of the simplest is

$$N_{n+1} = r \times (1 - r) \times N_n$$

As the population increases and becomes close to 1, the net reproduction rate falls.

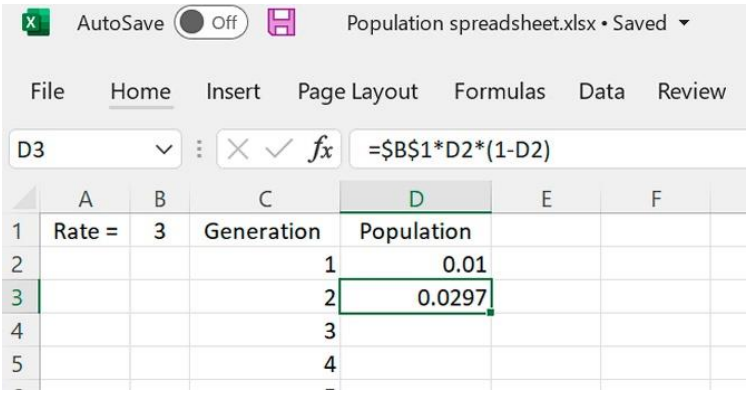
Positive and negative feedback appear in many situations. One is in electronics in analog circuits. Positive feedback is used to increase the gain of amplifiers and circuits, although at the cost of decreased stability, and negative feedback is used to stabilize amplifiers and circuits and improve frequency response, although at the cost of decreased gain.

Surprisingly, systems that are described by the equation above do not always possess the simple behavior expected of them. A useful and simple way to explore the behavior of such systems is to numerically simulate them on a spreadsheet. A spreadsheet program permits rapid simulation and the immediate plotting of results so that effects changes in the system can rapidly be explored. Throughout the book spreadsheet analysis will be described and encouraged on the part of the reader.

To study the population equation, the spreadsheet needs three parts. First, it must contain the reproductive rate $r$. It is best to enter this into a single cell of the spreadsheet and to have all cells that use the rate refer to this one cell to obtain this rate. Next, the spreadsheet needs to calculate the level of the population in successive generations using the previous population level and the reproduction rate. Some prior experimenting shows that fully seeing the behavior of this system requires at least several hundred generations, and five hundred will be used here. Finally, some of the interesting behavior of the system occurs in the first 50 generations. In addition, to see the final, steady state behavior, the population levels in the final 50 generations can also be observed. Therefore, the third part of the spreadsheet is plots of $N_n$ as a function of the generation for the first and last 50 generations.

Figure 1-2 shows a spreadsheet under construction. Cell A1 contains the name of the variable input rate that is entered in the adjacent cell, B1. It is this reproduction rate that will be used to calculate the population level after the first generation. The initial value of the population level is entered in cell D2. The key to the calculation is the formula for the population in the second generation that is in cell D3. As this cell is selected, its formula is displayed in the formula box near the top. It says to use the value of cells B1 and D2 and compute $B$1*D2(1-D2). To complete construction of the spreadsheet, cell D3 is copied and pasted into the next 500 cells in column D. In the pasting process, in each successive row, the index of D is increased by one

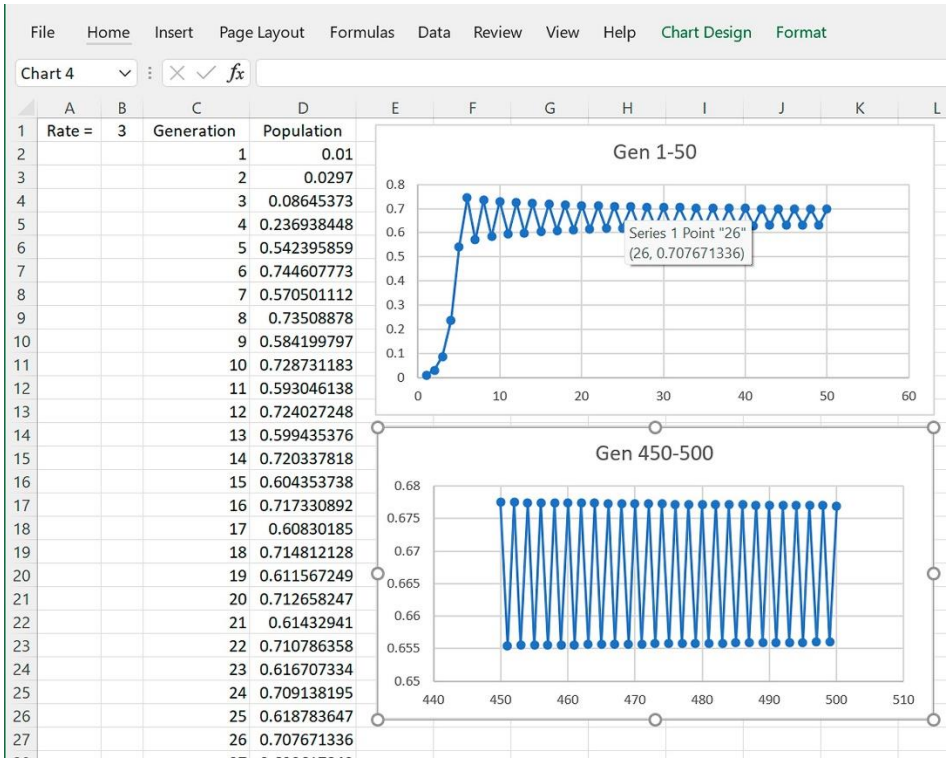
so that each cell takes its input value from the cell above. The dollar signs the equation instruct the spreadsheet not to index either the row or the column this variable, but always to take the value from cell B1. Thus, the formula that is place in cell D4 by the copy operation is $B$1*D3*(1-D3).




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Rate = | 3 | Generation | Population | | |
| 2 | | | 1 | 0.01 | | |
| 3 | | | 2 | 0.0297 | | |
| 4 | | | 3 | | | |
| 5 | | | 4 | | | |

Figure 1-2 Construction of the population spreadsheet.

After completing the construction of column D, the first 50 generations of population, columns C and D are selected and used to make a scatter plot of population as a function of generation. The same is done for the last 50 generations of data, Figure 1-3.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | Rate = | 3 | Generation | Population |
| 2 | | | 1 | 0.01 |
| 3 | | | 2 | 0.0297 |
| 4 | | | 3 | 0.08645373 |
| 5 | | | 4 | 0.236938448 |
| 6 | | | 5 | 0.542395859 |
| 7 | | | 6 | 0.744607773 |
| 8 | | | 7 | 0.570501112 |
| 9 | | | 8 | 0.73508878 |
| 10 | | | 9 | 0.584199797 |
| 11 | | | 10 | 0.728731183 |
| 12 | | | 11 | 0.593046138 |
| 13 | | | 12 | 0.724027248 |
| 14 | | | 13 | 0.599435376 |
| 15 | | | 14 | 0.720337818 |
| 16 | | | 15 | 0.604353738 |
| 17 | | | 16 | 0.717330892 |
| 18 | | | 17 | 0.60830185 |
| 19 | | | 18 | 0.714812128 |
| 20 | | | 19 | 0.611567249 |
| 21 | | | 20 | 0.712658247 |
| 22 | | | 21 | 0.61432941 |
| 23 | | | 22 | 0.710786358 |
| 24 | | | 23 | 0.616707334 |
| 25 | | | 24 | 0.709138195 |
| 26 | | | 25 | 0.618783647 |
| 27 | | | 26 | 0.707671336 |

Figure 1-3 Part of the spreadsheet calculating the population for 500 generations and plots of the population for the first and last 50 generations.

Unexpectedly, for a reproductive rate of three, the population increases from its initial value of 0.01 to around 0.60 and then stabilizes, flipping between two values. The spreadsheet design allows easily altering the reproduction rate by typing a new value into cell B1. Almost immediately the plots are updated.

At values of $r$ below 2.93 the population quickly stabilizes at a single value and at about 3.4 the two final values split and four final values are observed. At still higher values of r, eight and sixteen final values are observed and at still higher values, the long-term behavior becomes chaotic.

A number of dynamic systems show such chaotic behavior. The equations involved with weather prediction are among them. This chaotic behavior prevents forecasting weather beyond about ten days into the future. In chaotic systems very small changes in the initial conditions produce very large changes in the later system behavior.

**Poisson Distributions**

Often in working with cells, a very small fraction possesses or acquires some special property. It is not just in working with cells that one encounters a large number of possibilities in which some event has a low probability of occurring. An important question in these situations is determining the probability of $n$ events in a sample. The question could be how many cells survive a UV irradiation, how many radioactive decays occur in a certain time interval, how many times is the sequence from a particular gene represented in a massive sequencing project where short sequences from an organism's entire genome are determined?

First, a traditional derivation of the Poisson distribution will be presented, and then, for deeper understanding, as well as preparation for the next section of the chapter, its study with a spreadsheet will be described. For convenience, assume that the events in question occur in time, and let $x$ be the very low probability that an event occurs in a short time interval. The probability that the event does not occur is $1 - x$. In $N$ time intervals, the probability that the event occurs in each of the first $n$ intervals and does not occur in the remaining $N - n$ intervals is $x^n(1-x)^{N-n}$. Since we do not care about the particular order of these two types of intervals, any order suffices. Therefore, we need to total number of different arrangements of these two types, that is, the total number of possible rearrangements of $n$ and $N - n$ items. This number is

$$\frac{N!}{n!\,(N-n)!}$$

so that the probability of $n$ events out of $N$ intervals, $P(n)$ is

$$P(n) = \frac{N!}{n!\,(N-n)!}x^n(1-x)^{N-n}$$

When $N$ is large and $x$ is very small,

$$\frac{N!}{(N-n)!} \approx N^n$$

and

$$(1-x)^{N-n} \approx (1-x)^N$$

$$\approx e^{-xN}$$

so that $P(n)$ becomes

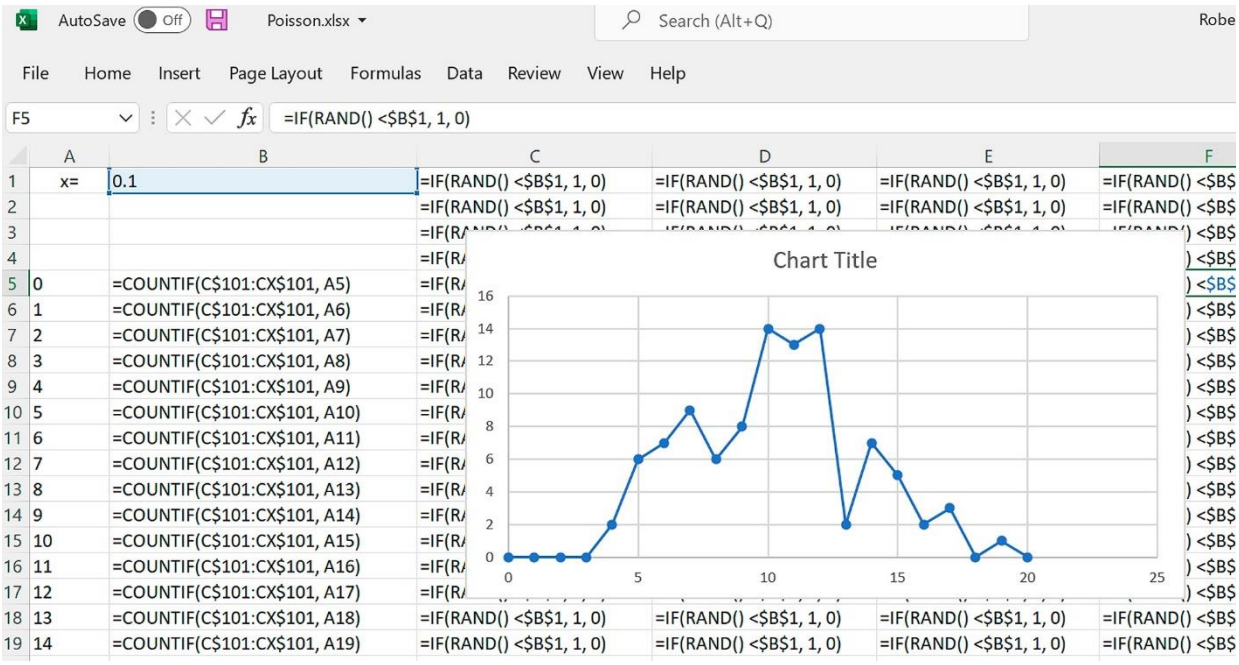
$$P(n) \approx \frac{N^n}{n!} x^n e^{-xn}$$

Since $x$ is the probability per time interval and there are $N$ intervals, $xN$ is the average number $m$ of events overall. Thus, the probability of $n$ events occurring when the average number is $m$ is

$$P(n,m) = \frac{m^n}{n!} e^{-m}.$$

A spreadsheet can be used to simulate situations that generate Poisson distributions. As in the previous spreadsheet, place the important variable $x$ in cell B1. Let each of the independent time intervals in which the event can occur be represented by a separate cell in column C. If we approximate using a value of 100 for $N$, then 100 cells will be used in column C. When the spreadsheet calculates, 1 is to be placed in a cell with a probability $x$ whose value is contained in cell B1. Otherwise, a 0 is to be placed in the cell. The Rand() function and If() function together can be used to perform the operation by writing, If(Rand()<x, 1, 0). The Rand() function returns a random value between 0 and 1, and if its value is less than x, the If() function inserts 1, and if Rand()'s value is greater than x, the If function returns 0, which is inserted in the cell. After entering "=If(Rand()<$B$1, 1, 0)", into cell C1 the command is copied and pasted into the next 99 cells in column C. The number of times the event occurred is given by counting the number of 1's in the column with the command "=Sum(C1:C100)".

To see the distribution of numbers of the events in multiple experiments, column C is copied and pasted into the next 100 columns. The spreadsheet now contains the results of 100 experiments in row 101 whose exact specification is C101:CX101. We want to know the number of times that 0, 1, 2,… appeared in this row. To do this, the function Countif() will be used. It counts the number cells in a specified range that satisfy a specified condition. To use this function, enter the numbers 1-20 in cells A5 through A25 and the function Countif(C$101:CZ$101=A5) and copy it into cells A6-A25. Note the use of the dollar sign that forces the Countif() function always to refer to row 101. With this done, instruct the spreadsheet to make a scatter plot of cells B5-B25 as a function of the values in cells A5-A25. The plot shows the distribution of the numbers of events that occurred in the 100 experiments, Figure 1-4.




| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | x= | 0.1 | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$ |
| 2 | | | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$ |
| 3 | | | =IF(RA | | | ) <$B$ |
| 4 | | | =IF(RA | | | ) <$B$ |
| 5 | 0 | =COUNTIF(C$101:CX$101, A5) | =IF(RA | | | ) <$B$ |
| 6 | 1 | =COUNTIF(C$101:CX$101, A6) | =IF(RA | | | ) <$B$ |
| 7 | 2 | =COUNTIF(C$101:CX$101, A7) | =IF(RA | | | ) <$B$ |
| 8 | 3 | =COUNTIF(C$101:CX$101, A8) | =IF(RA | | | ) <$B$ |
| 9 | 4 | =COUNTIF(C$101:CX$101, A9) | =IF(RA | | | ) <$B$ |
| 10 | 5 | =COUNTIF(C$101:CX$101, A10) | =IF(RA | | | ) <$B$ |
| 11 | 6 | =COUNTIF(C$101:CX$101, A11) | =IF(RA | | | ) <$B$ |
| 12 | 7 | =COUNTIF(C$101:CX$101, A12) | =IF(RA | | | ) <$B$ |
| 13 | 8 | =COUNTIF(C$101:CX$101, A13) | =IF(RA | | | ) <$B$ |
| 14 | 9 | =COUNTIF(C$101:CX$101, A14) | =IF(RA | | | ) <$B$ |
| 15 | 10 | =COUNTIF(C$101:CX$101, A15) | =IF(RA | | | ) <$B$ |
| 16 | 11 | =COUNTIF(C$101:CX$101, A16) | =IF(RA | | | ) <$B$ |
| 17 | 12 | =COUNTIF(C$101:CX$101, A17) | =IF(RA | | | ) <$B$ |
| 18 | 13 | =COUNTIF(C$101:CX$101, A18) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$ |
| 19 | 14 | =COUNTIF(C$101:CX$101, A19) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$1, 1, 0) | =IF(RAND() <$B$ |

Figure 1-4 A plot of a simulation that generates a Poisson distribution. In this case, with a probability of 0.1 for 100 events, the average is 10. The plot is shown overlaid with part of the spreadsheet that generated it. The display of the spreadsheet has been set to display the equations in each cell rather than the numbers generated by the equations in each cell.

The value of $x$ can be altered and the plot will then show the results of another set of experiments. Reevaluating the spreadsheet without changing the value of $x$ performs a new set of experiments. The results give an idea of the sizes of fluctuations to be expected when experimental data is described by a Poisson distribution.

A good indication of the shape of the distribution is given by its mean and its width. Two measures of width are variance and standard deviation, which is the square root of variance. Variance is the average of the squared deviation of a random variable from its mean where $x_i$ is the value of the ith variable, $n$ is the total number of $x_i$ and $\overline{x}$ is the average of the $x_i$

$$Var = \sum \frac{(x_i - \overline{x})^2}{n}$$

Typically, scientific results are presented as the mean ± standard deviation because in this way the units of the variable, the mean of the variable, and the standard deviation are all the same. For example, one might write that the length of some object is $100 \pm 8$ Å which is simpler to understand than if it were written using variance as $100$ Å $\pm 64$ Å$^2$. If we add functions to the spreadsheet that calculate the mean, variance, and standard deviation of the results of the 100 simulations, that is, the numbers in the array C101:CX101, something interesting is seen. The mean and the variance are very close to one another for all values of $x$. For the Poisson distribution itself, and not a simulation of it, these are indeed equal, as can be shown mathematically.

**The Fluctuation Test**

Properties of the Poisson distribution were used by Luria and Delbrück to demonstrate in bacteria that mutations to resistance to phage existed in some cells prior to exposure to the phage rather than somehow being induced by this exposure. This was a simple and decisive experiment that showed that Lamarckian inheritance did not apply in this particular case.

The basis of the Luria-Delbrück fluctuation test is the fact that in a Poisson distribution, the variance equals the mean. Suppose that $10^7$ bacteria are spread on a petri plate and irradiated with sufficient UV light that only 100 cells survive and grow into colonies and that the cells from these colonies are as sensitive to the UV light as cells from the original culture. This is a situation that is well described by a Poisson distribution because a large number of cases are involved, $10^7$, and the event in question occurs with a low probability. If many plates of cells are similarly treated, say 200, then the numbers of colonies per plate from the surviving cells should be Poisson distributed and have both a mean and a variance of 100.

Instead of performing the experiment this way, consider the results if the sample of cells to be spread on each petri plate had been irradiated and the survivors then allowed to double in number before plating. The average number of cells per plate will now be 200. The variance however will be 400, as seen from the definition of variance

$$Var = \sum \frac{(x_i - \overline{x})^2}{n}$$

for the new variance will be

$$New\ Var = \sum \frac{(2x_i - 2\overline{x})^2}{n}$$

$$= \sum 4\frac{(x_i - \overline{x})^2}{n}$$

so that

$$New\ Var = 4 \times Var$$

and now the variance is greater than the mean. This means that if all the numbers in a Poisson distribution are multiplied by a factor greater than one, the new variance of the population will be greater than the mean.

Suppose 100 cultures are grown from single cells and are plated out in the presence of phage so that only resistant cells can grow. If the distribution of the number of resistant colonies shows a variance greater than the mean, it means that at least some of the resistant cells arose during growth of the cells before they were plated out in the presence of the phage. This is what Luria and Delbrück found.

**Problems**

1. If all cells of a freely growing population divide when they reach an age of one, what is the average age of cells in the population?
2. Conway's game of Life provides another interesting simulation of population behavior. In this simulation, squares in a rectangular array are either alive or dead. A cell has

eight neighbors, four on the sides and four on the corners. From an initial configuration of dead and alive cells, the status of cells in the next generation is given by the following rules. A live cell dies in the next generation if it has zero or one live neighbors or if it has more than three live neighbors. A live cell survives if it has two or three live neighbors. A dead cell becomes alive in the next generation if it has three live neighbors. Until it becomes uninteresting, follow the behavior of a population that begins with four live cells in the configuration:

```
 X
XXX
```

3. Suppose in a light rainstorm, 100 raindrops fall on a square foot. What is the probability that a specified square foot receives zero raindrops?
4. Determine the average number of photons hitting each pixel of a digital camera when photographing a white card if the intensities of 20 of the pixels as reported by an image editing program were: 59.5, 59.4, 59.1, 59.4, 59.4, 59.3, 59.3, 59.4, 59.4, 59.1, 59.4, 59.4, 59.3, 59.4, 59.5, 59.5, 59.5, 59.3, 59.3, 59.5. Assume that the only source of fluctuations in the final reported values is fluctuations in the numbers of photons and does not arise in the camera electronics. Hint, note that in the example in the text of a Poisson distribution that was distorted by cell growth, the final population could be transformed into a Poisson distribution where the mean equals the variance by multiplying the number of cells on each plate by a factor of 0.5.
5. What was the political significance of the Luria and Delbrück finding that it was not exposure to the phage that induced the generation of phage resistant cells?

# 2. Random Walks, Brownian Motion, and Diffusion

### Importance in Biology

Biological phenomena critically depend on molecular motion, both of an organism's own activities and of the motion inherent in matter that is at any temperature above absolute zero. The random motion of atoms and molecules as well as of very small particles suspended in liquids is called Brownian motion. Remarkably, as late as the early 1900's, belief in the existence of discrete atoms was not widespread until Einstein provided a convincing atomistic explanation of Brownian motion. We now know that the random motion of atoms and molecules is required to provide substrates to enzymes and move small cellular components relatively short distances within cells. Over larger distances, molecular motors are required to move things about and perform indispensable roles in cells in processes like DNA synthesis and transport of molecules. The activities of these motors also critically depend on Brownian motion. This chapter will first consider random walks of individual particles undergoing Brownian motion, and then will consider diffusion, which is the collective behavior of large numbers of particles undergoing Brownian motion.

### Simulating Random Walks

One objective of this material is to help develop intuitive understandings of biological phenomena. To aid in this, first, random walks will be simulated, and after developing some feel for their behavior, they will be treated analytically. Spreadsheets will be used here because they are a convenient tool for scientific computations, simulations, and the plotting of results.

Consider a random walk in two dimensions where each step is of the same length but can be in any direction. After determining a random direction by generating a random angle, the spreadsheet can then calculate the x and y components of that step. The Excel function Rand() generates a random number between zero and one. Since trigonometric functions in Excel use radians rather than degrees, the function 2*Pi()*Rand() generates random angles over a full circle. Figure 2-1 shows the formulas needed in rows three and four of the spreadsheet. Such a view of the formulas is generated by choosing Show Formulas from the Formulas menu item. The final two columns of the sheet cumulatively add the X and Y components of the successive steps.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Random Angle | X component | Y component | Cumulative X | Cumulative Y |
| 2 | | 0 | 0 | 0 | 0 |
| 3 | =2*PI()*RAND() | =COS(A3) | =SIN(A3) | =D2 +B3 | =E2 +C3 |
| 4 | =2*PI()*RAND() | =COS(A4) | =SIN(A4) | =D3 +B4 | =E3 +C4 |
| 5 | | | | | |

Figure 2-1. Initial rows of a spreadsheet for the simulation of a two-dimensional random walk starting from the origin consisting of steps of length one taken in any direction.

After entering rows one, two, and three, row three is then copied and pasted into the next 50 or 1000 rows. Row four was included in the figure to show how the addressing of the cells is

automatically adjusted when they are copied. The cumulative values in row four correctly refer to the cells immediately above and to that row's X and Y components. To visualize the simulated random walk, the fifty rows of cumulative X and Y data are selected, Insert Charts/All Charts/X-Y Scatter, and the chart type of continuous broken lines without data points is chosen. A virtue of the spreadsheet simulation is that having the spreadsheet recalculate with F9 immediately displays the path of a new random walk, two of which are shown in Figure 2-2, a 50-step walk and a 1000-step walk.

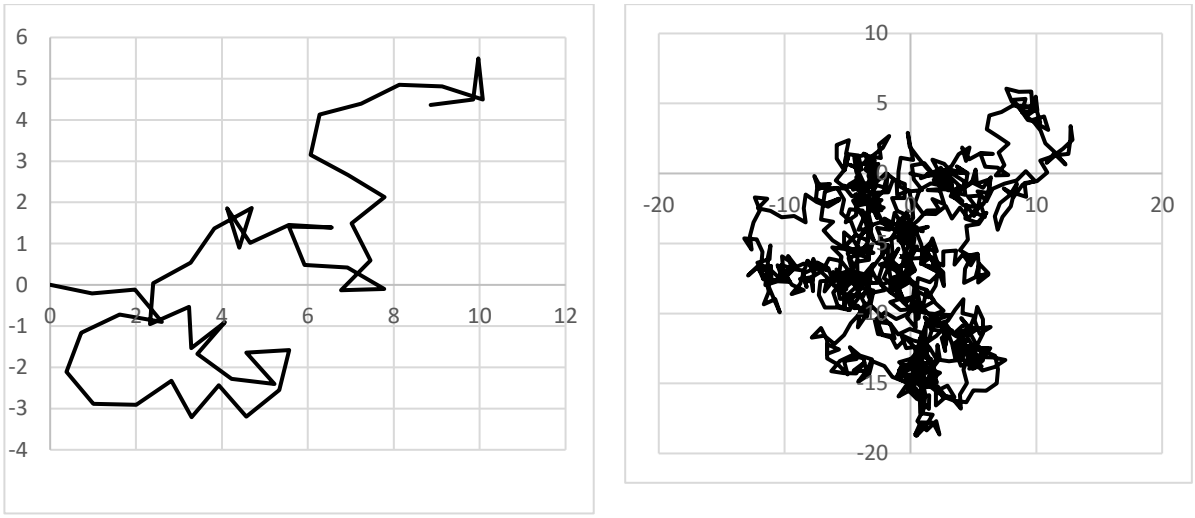



Figure 2-2 Paths of 50 and 1000 step random walks

Examining multiple plotted paths shows that the average distance from the starting point to the final point of a random walk is much less than the number of steps taken times the length of the steps. Also, comparing the 50-step and 1000-step walks shows that the average distance of the final point from the starting point does not increase in proportion to the number of steps taken. Using the spreadsheet to calculate the average of the square of the distance walked for ten or twenty walks, yields numbers close to the number of steps taken. That is, it looks like $\overline{R^2} = N$ where $R$ is the distance from the starting point to the ending point, $N$ is the number of steps taken, and the overbar indicates average.

**The Mean Squared End to End Distance in a Random Walk is Proportional to the Number of Steps Taken**

The average of the endpoint positions from a large number of random walks starting from the origin will just be the origin. Therefore, some measure other than average of the endpoints is needed to characterize the net distance achieved in a random walk. While the average of the absolute values of the endpoints would be a suitable measure, its computation is not convenient. The average of the squared position of the endpoints proves to be adequate. Mean squared means that the average of the squared distances for a large number of trial walks is computed. As in the previous section, consider a two-dimensional random walk starting from the origin where each step is of unit length, but in a random direction. As before, let $\theta_i$ be the angle of the *i-th* step with respect to the x axis. Then, for *N* steps, the coordinates of the end of the walk are

$$X = \cos\theta_1 + \cos\theta_2 + + + \cos\theta_N$$

$$Y = \sin\theta_1 + \sin\theta_2 + + + \sin\theta_N$$

Since the distance squared from the origin,

$$R^2 = X^2 + Y^2$$

$$R^2 = \left(\cos{\theta_1}^2 + \cos{\theta_2}^2 + + + \cos{\theta_N}^2 + cross\ terms\ 2\cos\theta_i cos\theta_j\ where\ i \neq j\right) +$$

$$\left(\sin{\theta_1}^2 + \sin{\theta_2}^2 + + + \sin{\theta_N}^2 + cross\ terms\ 2\sin\theta_i \sin\theta_j\ where\ i \neq j\right)$$

and since

$$\sin{\theta_k}^2 + \cos{\theta_k}^2 = 1$$

for *k=I, 2, …N*, and since the average of the sums of all the cross terms is zero because each factor in the cross terms has an equal probability of being positive or negative.

$$\overline{R^2} = N$$

If the step length had been $L$, the result would be

$$\overline{R^2} = L^2 N$$

The same result is found for random walks in one, two, three and n dimensions. This is an important result as many physical phenomena in addition to molecules undergoing Brownian motion involve random walks. The square root relationship between the number of steps taken and the net movement achieved is reminiscent of the more widely recognized fact that increasing the number of measurements $N$ of a quantity reduces the size of the standard error of its determination by $1/\sqrt{N}$. In this situation, the errors are analogous to a random walk.

The approach used here of simulating, guessing, and building up resembles the way many scientific and medical discoveries are made—incrementally, by addressing parts of the problem, and by guessing, resulting in an approach which itself often looks like a random walk. Many times, when this approach has yielded a result or finding that looks very likely of being correct, it is then possible to definitively test the correctness of the finding with a relatively simple yes-no experiment. An illustrative example would be finding the presumed two prime factors of a number that is 40 digits long. Once there are hints that a particular prime is one of the factors, it is simple to test the hypothesis. Finding such a prime factor can be very lengthy however, and this difficulty forms the basis of some cryptographic methods.

**Random Walks When Numbers Become Large**

From the behavior shown by the random walk simulations, a reasonably good guess can be made as to the behavior of the probability of reaching a distance $x$ from the origin in $N$ steps. The probability will be symmetric about the origin and since more of the endpoints occur near the origin than far from it, the probability will be peaked at the origin. Also, since the average distance reached from the origin likely increases as the $\sqrt{N}$, the probability will probably spread in width in proportion to $\sqrt{N}$. Finally, since the total probability must be unity, then $Width \times Height \cong 1$ so that the amplitude likely falls like $1/\sqrt{N}$.

For a one-dimensional walk with unit length steps and an equal probability for each step of going to the right or to the left, the probability of taking n steps to the right followed by taking m steps to the left is

$$P = \left(\frac{1}{2}\right)^n x \left(\frac{1}{2}\right)^m$$

$$= \left(\frac{1}{2}\right)^N where\ N = n + m$$

This is just one of many ways of taking n steps to the right and m steps to the left. The total number of distinguishable ways of taking these steps is

$$\frac{N!}{n!\,m!}$$

Therefore, the probability of ending at $x = n - m$ steps from the origin is

$$P(x) = \frac{N!}{n!\,m!}\left(\frac{1}{2}\right)^{n+m}$$

Figure 2-3 shows a plot of $P(x)$ for a total of 20 steps.

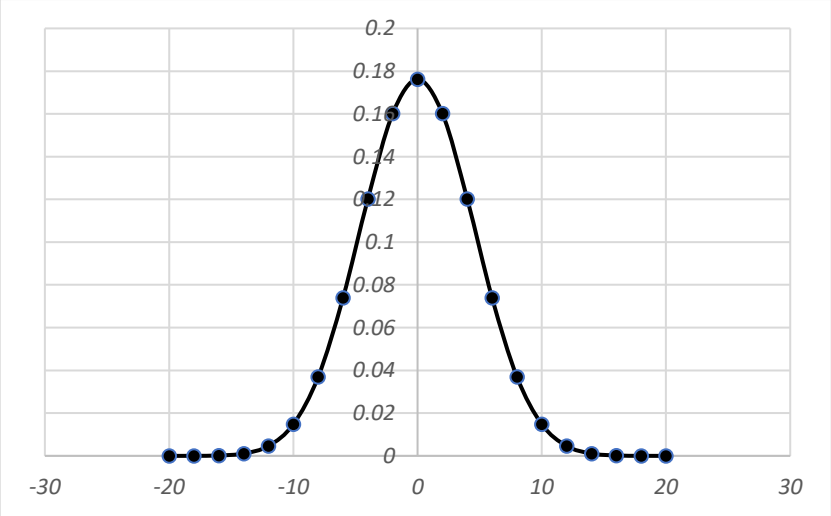



Figure 2-3 Plot of the probability of ending at position x for a random walk of 20 steps with equal probability of moving left or right with each step.

The plot looks very much like a normal distribution which is sometimes called a Gaussian curve or bell curve. Constructing a Gaussian based on the hints provided by the numerical simulations gives something like

$$P(x) = \frac{a}{\sqrt{N}} \mathrm{e}^{\frac{-bx^2}{N}}$$

Where the values of $a$ and $b$ are not yet determined. A rigorous analysis of the probability distribution for large $N$ gives

$$P(x) = \frac{2}{\sqrt{2\pi N}} \mathrm{e}^{\frac{-x^2}{2N}}$$

This result could also have been found by using the central limit theorem which says that in many cases, the averages of the sums of independent random variables tend toward a normal distribution.

The basis of diffusion is random walks. In addition to diffusion, searching strategies sometimes are based on random walks, both in the macro world and the microbial world. Fluctuations of sums also show random walk behavior, as do the total amount of twist in frequently used power cords or the conformations of long linear polymers like DNA.

**Diffusion within a Cell**

Before looking at the basic equations governing diffusion, first a result from diffusion theory will be considered. In light of the properties of random walks and understanding that the basis of diffusion is just random walks, it is not surprising that the mean squared distance that a particle can diffuse in a time $t$ is proportional to $t$

$$\overline{R^2} = 6Dt$$

where $D$ is the diffusion constant characteristic of the diffusing molecule. This equation can be used to estimate the time required for a typical protein to diffuse from one end of a bacterial cell of length $10^{-4}$ cm to the other end. This estimation requires the diffusion constant of a protein in the cell cytoplasm. From tables of diffusion constants, a typical protein of molecular weight 60,000 has a diffusion constant of about 1 x $10^{-6}$ $cm^2$/sec in water, and guessing that the relevant viscosity in the cell cytoplasm is about five times that of water, the diffusion time turns out to be the astounding value of less than 1/100 second. It means that the interior of a bacterial cell is churning about at a terrific rate.

Consider a cell at the other end of the length spectrum, a nerve cell, whose length may be about a meter. In this case, the estimated time to diffuse from one end to the other is $10^{10}$ seconds, or about 30 years, also an astonishing number. Thus, while diffusion is perfectly adequate to move proteins and small molecules around within a bacterial cell, it cannot begin to supply the transport needs of some eukaryotic cells. Small wonder then that eukaryotic cells have evolved transport systems that use fibers within the cells as railroad tracks and molecular motors to move things along these tracks.

When taking values from the literature like the diffusion constant, it is important to verity that the units in the equations are all consistent. If length units are cm, it follows from the equation above that $D$ must have units of $cm^2$/sec, as was the case in the table used for the diffusion constant.

The value for the diffusion constant for a particular protein within cells is hard to estimate. The problem is that what is relevant here is the viscosity that the protein molecule experiences as it is subjected to Brownian motion. In some cases, the relevant viscosity can strongly vary with the size of the diffusing molecule. For example, a bacteriophage cannot diffuse at all through a stiff agarose or agar gel, but a smaller molecule like glucose can diffuse through such gels as though they weren't even there.

**Motors and Brownian Ratchets**

It is not hard to imagine the basic principles that underlie the action of the motors that we encounter in everyday life. In an electric motor, passing an electric current through a coil generates an electromagnet with north and south poles. Reversing the direction of the current reverses the poles. Therefore, by periodically reversing the current flow in a coil in the motor, the coil can be periodically attracted and repelled from another magnet or electromagnet. Properly arranged, this property can yield rotational or linear movement. Similarly, periodically admitting steam into a cylinder with a piston can generate a periodic force that, again, can form the basis of a motor. What about motors at the molecular level?

At the molecular level, there are no forces that can be turned on and off as they can be in the macro world. Brownian motion however, provides a solution to the molecular motor dilemma. Suppose a molecular motor or a part of it is sufficiently small that it is subject to significant Brownian motion. A motor can be realized if the Brownian motion of the motor or a part of it is allowed to move in one direction, but movement in the reverse direction is blocked. Of course, since this is a "No free lunch" universe, energy must be consumed somewhere in the process. In most cases the energy consumption occurs at the step of blocking reverse motion but not blocking the forward motion. Such mechanical devices are called ratchets. DNA polymerase provides an example of such a Brownian ratchet molecular motor. In the addition of one base to an elongating DNA strand, the nucleoside monophosphate is added to the growing chain and a pyrophosphate is released. The incorporated base blocks the backward movement of the polymerase. Brownian motion however, soon jostles the polymerase forward, leaving space for the incorporation of the next nucleotide. Thus, Brownian motion is providing the actual driving force for the motor, and the energy that is released by the cleavage of phosphodiester bond permits the ratchet to work.

**Diffusion Equations**

The objective of the next several sections is to derive the diffusion equation and then to understand how it can be applied to problems of biological interest. Since random walks and Brownian motion underlie diffusion, there must be an upper limit to the rate at which an enzyme's substrate can diffuse to it. With the diffusion equation, this rate can be determined as shown below.

Consider a solution containing molecules of species $A$ that are undergoing Brownian motion. If the concentration of $A$ is the same everywhere, there will be no net flow anywhere. If, however the concentrations are not equal everywhere, there will be net flow from regions of higher concentration to regions of lower concentration. The greater the concentration difference between regions, the greater will be the flow. A local measure of concentration difference is the rate at which the concentration changes with distance,

$$\frac{dc}{dx} = \lim_{\Delta x \to 0} \frac{c(x + \Delta x) - c(x)}{\Delta x}$$

This is called the gradient. The constant of proportionality relating flow rate to gradient is called the diffusion constant. Thus, in one dimension, the flow rate $J$ at a point is

$$J = -D \frac{\partial c(x)}{\partial x}$$

This is known as Fick's law. The flow, which is sometimes called the flux, will change the concentrations. To account for changes in the concentrations, it is necessary to account for how the flux changes the concentrations. Considering only a concentration gradient in the x direction, the net flow into a box of dimension $\Delta x \Delta y \Delta z$ equals the area of the ends of the box,

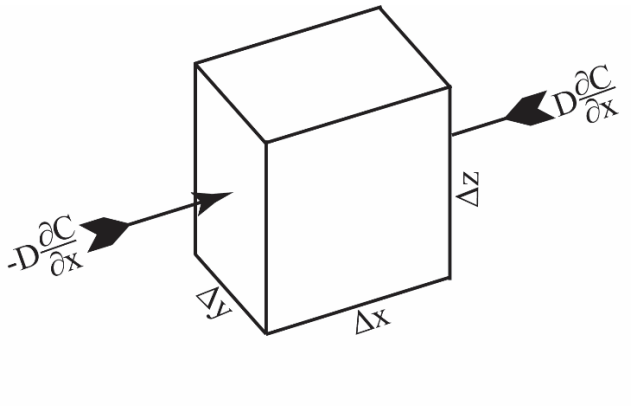



Figure 2-4 Flow into the box changes the concentration inside.

$\Delta y \Delta z$, times the flow rate in the positive direction at position $x$ plus the flow rate in the negative direction at the position $x + \Delta x$, times the period of flow, $\Delta t$, Figure 2-4.

$$\text{Total flow in} = \left( -D \frac{\partial c(x)}{\partial x} \Delta y \Delta z + D \frac{\partial c(x + \Delta x)}{\partial x} \Delta y \Delta z \right) \Delta t$$

$$= D \left( \frac{\partial c(x + \Delta x)}{\partial x} - \frac{\partial x(x)}{\partial x} \right) \Delta y \Delta z \Delta t$$

$$= D \left( \frac{\frac{\partial c(x + \Delta x)}{\partial x} - \frac{\partial c(x)}{\partial x}}{\Delta x} \right) \Delta x \Delta y \Delta z \Delta t$$

The total flow in divided by the volume of the box equals the change in the concentration in the time $\Delta t$, or

$$D \left( \frac{\frac{\partial c(x + \Delta x)}{\partial x} - \frac{\partial c(x)}{\partial x}}{\Delta x} \right) \Delta t = \Delta c(x)$$

Dividing by $\Delta t$, taking the limit as $\Delta x, \Delta y, \Delta z, \Delta t$ approach zero, the equation becomes

$$D \frac{\partial^2 c(x,t)}{\partial x^2} = \frac{\partial c(x,t)}{\partial t}$$

This is the diffusion equation in one dimension. In three dimensions it is

$$D \left( \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) c(x.t) = \frac{\partial c(x,t)}{\partial t}$$

This equation is the same as the equation governing heat flow and it and simple variants of it show up in many scientific and engineering situations. Interestingly, the Schrodinger equation of the wave function in quantum mechanics is a diffusion equation in imaginary time.

**Diffusion to a Black Hole (An Enzyme's Active Site)**

Since an enzyme cannot operate more rapidly than substrate can diffuse to its active site, it is of interest to see how the diffusion limited rate of an enzyme compares to its actual rate *in vivo*. The elongation rate of DNA polymerase in prokaryotes seems like a situation where a high turnover number could be of great importance. To make such an estimation, the diffusion equation can be used to calculate the rate of diffusion to a sphere with a radius of about the same radius as the dimensions of the active site of the polymerase. Once the next nucleotide has been incorporated into the growing chain, it is no longer a freely diffusing nucleotide and, as far as nucleotides diffusing to the enzyme, it disappears from consideration. Thus, the active site is a black hole with respect to nucleotides.

Before carefully deriving an equation governing the diffusion in this case, it is useful to guess what the important variables are and how they might enter the final result. Such an analysis helps develop one's intuition and also provides a check on the final, carefully derived result. The rate of diffusion to the active site will have the dimensions of $moles/sec$. Therefore, the final expression for the rate should also have these dimensions. The maximum diffusion rate ought to be proportional to the concentration of substrate at great distances from the enzyme $c_0$. Closer to the active site, the concentration should be lower, and right at the boundary of the active site, the concentration should be zero.

The units of the concentration of substrate, $c$, will be $moles/cm^3$. The rate will be proportional to the diffusion constant $D$ of the substrate and have dimensions $cm^2/sec$. Finally, it seems likely that the rate will involve either the radius of the active site, $r_0$, with dimensions $cm$, or be proportional to its area,$r_0^2$ with dimensions $cm^2$. One combination of the variables having the required dimensions is $Dc_0r_0$. This type of analysis in which combinations of the relevant variables are sought that possess the required dimensions is called dimensional analysis, and is often helpful in developing theories to explain physical phenomena.

If the active site is placed at the origin of spherical coordinates, then the only coordinate variable relevant to concentration will be the distance *r* from the origin. Also, by considering the situation after equilibrium has been achieved, $\partial c/\partial t = 0$. Then, the diffusion equation in spherical coordinates becomes

$$\left(\frac{\partial}{\partial\theta}\right) + \left(\frac{\partial}{\partial\varphi}\right) + \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial}{\partial r}\right)Dc(r,\theta,\varphi,t) = 0$$

Since $c$ is a function only of $r$, the equation becomes even simpler

$$\frac{2}{r}\frac{dc(r)}{dr} + \frac{d^2c(r)}{dr^2} = 0$$

There are several ways to solve a differential equation like this. In contrast to the relative ease of finding the derivative of an expression, finding the inverse, that is, solving a differential equation is not straightforward. One efficient method of solving many such equations is to use Google to find a website that solves differential equations and to hope that the site can solve

yours. Here another method is described. Sometimes the best or sometimes the only way to solve a problem is to guess candidates and then check to see if they solve the problem. Here a guess is made as to the form of a potential solution and this is then adjusted to see if it can be made to exactly solve the equation and also to satisfy the boundary conditions. In this case, the boundary conditions are that $c = c_0 \; at \; r = \infty$ and $c = 0 \; at \; r = r_0$. A plausible form for the solution is

$$c = r^n$$

Where the value of $n$ is yet to be determined. Substituting this into the diffusion equation gives

$$2nr^{n-2} + n(n-1)r^{n-2} = 0$$

This equation can be made to be true if

$$n = 0, -1$$

Therefore, a general solution to the differential equation is

$$c = a + \frac{b}{r}$$

where *a* and *b* can have any values. The solution will be acceptable if *a* and *b* can be adjusted to fit the boundary conditions. Applying them at $r = \infty$ gives

$$a = c_0$$

and at $r = r_0$ gives

$$0 = c_0 + \frac{b}{r_0}$$

$$b = -c_0 r_0$$

so that the full solution is

$$c = c_0 \left(1 - \frac{r_0}{r}\right)$$

As required, the concentration far from the active site is $c_0$ and this decreases and reaches zero as the active site is reached.

The flow rate through a sphere of radius $r$ centered at the origin is the gradient of the concentration times the diffusion constant,

$$\text{Flow} = -D\frac{dc}{dr}$$

which is

$$\frac{Dc_0 r_0}{r^2}$$

The total flow through a sphere of radius *r* is the area of the sphere times the flow rate per unit area through the sphere

$$\text{Flow} = 4\pi r^2 \times \frac{Dc_0 r_0}{r^2}$$

$$= 4\pi D c_0 r_0$$

The total flow through a sphere of any radius centered on the active site is the same. This is as it should be since molecules disappear only at the active site, and the flow through any sphere must be the same and equal to the rate the molecules disappear. The result involves the diffusion constant, the concentration far from the active site, and the radius of a sphere approximating the active site, a result expected from the dimensional analysis which was made at the beginning of this section.

### The Diffusion Limited Elongation Rate of DNA Polymerase

The result derived above can be used to estimate the maximum rate at which nucleotides could diffuse to DNA polymerase to support DNA replication. The diffusion constant for nucleoside triphosphates in water is about $5 \times 10^{-6} cm^2/sec$, and estimating the viscosity for a nucleotide in a bacterial cell to be $10 \times$ that of water, $D \approx 5 \times 10^{-7} cm^2/sec$. The concentration of deoxynucleoside triphosphates is about $2 \times 10^{-4}\ molar = 2 \times 10^{-7} moles/cm^3$. Using $10\ Å$ for $r_0$ the flow is

$$\text{Flow}\left(\frac{moles}{sec}\right) = 4\pi \times 5 \times \frac{10^{-7} cm^2}{sec} \times 2 \times \frac{10^{-7} moles}{cm^3} \times 10^{-7} cm$$

$$\approx 1.2 \times 10^{-19} moles/sec$$

$$\approx 7 \times 10^4\ nucleotides/sec$$

The experimentally measured elongation rate is between 1,000 and 2000 nucleotides/sec. Thus, DNA polymerase is not rate limited by the rate of diffusion of substrates to its active site. Several enzyme catalysts however, are close to having diffusion-limited rates of reaction. One of these is carbonic anhydrase, an enzyme mentioned in the chapter on hemoglobin.

### Diffusion Limited Rates of Reactions

The next chapter deals with kinetics, but as a prelude, here a rate question will be addressed. The rate that two types of molecules can bind to one another or interact with each other can be no faster than the diffusion limited rate of their encountering one another. What is this rate? A closely related rate was calculated in the previous section where the rate that molecules at a given concentration could diffuse to a single target was calculated. Here that formulation will be modified so that instead of a single target molecule, we have target molecules in solution at a given concentration.

Often, binding reactions between types of molecules are viewed chemically, and therefore are described using rate constants. For example, if species $A$ binds to species $B$ to form the complex $AB$, the initial forward rate of binding or interaction is given by the rate equation

$$\frac{d[AB]}{dt} = k \times [A] \times [B]$$

where $k$ is the forward rate constant and has the units $1/molar \cdot sec$. The question then is what is the upper limit to the value of $k$?

To address the question of the maximum binding rate, we will use the result from above giving the rate of diffusion to a single target. In our new situation however, we must consider the reaction as taking place in some volume, which here is taken to be one liter. Let $D$ be the effective diffusion constant of either type of molecule. Since $D$ varies slowly with radius, approximate it as twice the diffusion constant of either $A$ or $B$. Let the constant $c_0$ as used above correspond to the concentration of species $A$. Let $r_0$ correspond to the separation between the two molecules as they collide. Instead of a single target molecule, the number of target molecules is $B \times N$ where $B$ is the concentration and $N$ is Avogadro's number. Therefore, setting the rates equal,

$$k \times [A] \times [B] = 4\pi D r_0 \times 4 \times [A] \times [B] \times N,$$

Gives the diffusion limited rate constant as

$$k = 16\pi D r_0 N.$$

If the units of the diffusion constant are $cm^2/sec$ and the radius of the molecules is in $cm$, including the units, the maximum rate becomes

$$16\pi D \frac{cm^2}{sec} \times r_0 cm \times N \frac{1}{mole}.$$

Since

$$1cm^3 = \frac{1}{1000} liter$$

and

$$\frac{liter}{mole} = \frac{1}{molar}.$$

we finally get

$$k = \frac{16\pi D r N}{1000} \frac{1}{molar \cdot sec}$$

For $r = 30\text{Å}$ this evaluates to about $10^{10}/_{molar \cdot sec}$, and this number is often indiscriminately cited, but in the derivation, several issues have been glossed over or ignored altogether. First, this result is based on the rate of diffusion of molecules to a black hole after a steady state has been achieved. At this time, the concentration near the target is lower than it is further away. In the case of molecules of two species colliding, such a lowering of the nearby concentration does not occur. For the case of macromolecules like proteins where binding to form a complex requires a precise orientation of the colliding molecules and not just their random collision in any orientation, the rate must be much lower than is given by the equation above.

**Problems**

1. Simulate a random walk of 100 steps in one dimension in which all steps are to the right, but the step length has a uniform probability of being anything between 0 and 1. Comment.
2. Construct a spreadsheet for simulating a random walk in one dimension making use of the following equation for cell A2 of a spreadsheet program," =A1 + If(Rand() < 0.5, -1, 1)".

3. The diffusion constant of a protein likely varies as some power of the radius. From a table of diffusion constants for different proteins and a suitable plot, determine what that power is.
4. Bacteria undergoing chemotaxis towards an attractant execute a variant of a random walk. What is it?
5. Consider a circular Brownian ratchet machine in which two Brownian ratchets are set up so that each serves as the block of the ratchet of the other. Presumably then, the pair would go round and round without an external source of energy. Beyond the fact that perpetual motion machines are impossible, what is wrong with this scheme?
6. Many books and web sites state that it was pollen that was observed by Brown in his observation of what is now called Brownian motion. This is not quite correct. What actually was observed?
7. Show that the mean squared distance from the starting point achieved in a random walk in three dimensions increases as the number of steps taken, just as it does in random walks in one or two dimensions.

# 3. Kinetics

### Biology is All About Rates and Binding

This chapter concerns rates and binding. The chemical reactions that occur in cells are highly important, but of equal, if not greater importance and interest are binding and dissociation reactions that do not directly involve covalent changes. Proteins bind to each other to form complexes that regulate and perform many of the important cellular tasks like DNA, RNA, and protein synthesis, and proteins bind to DNA and RNA for control of replication, recombination, transcription, and protein synthesis.

### Rate and Equilibrium Constants

Consider the reaction of $A$ plus $B$ forming the complex $A \cdot B$.

$$A + B \rightarrow A \cdot B$$

The rate of the reaction, $dAB/dt$ is proportional to the concentrations of $A$ and $B$. This familiar result follows because forming the complex $AB$ requires a collision between a molecule of $A$ and a molecule of $B$. If the concentration of $A$ is doubled, the rate of the productive collisions between $A$ and $B$ doubles. The same is true when $B$ doubles. Hence the fundamental rate equation for the rate of formation of the $AB$ complex follows.

$$\frac{d[A \cdot B]}{dt} = k_1[A] \times [B]$$

A word about nomenclature. Usually rate constants are written in lower case and equilibrium constants in upper case. Subscripts can indicate what reaction the constant refers to, or the direction of the reaction. In this case, $k_1$ indicates the forward reaction, and $k_{-1}$ would indicate the reverse reaction.

If the $AB$ complex has no memory, for instance no delay after the formation of the complex until it has settled into a stable steady state conformation, then the rate of decay of the complex will be proportional to the amount of the complex present.

$$\frac{d[A \cdot B]}{dt} = -k_{-1}[AB]$$

Initially after mixing $A$ and $B$, the net rate of formation of the complex will be high, but as the concentration of $AB$ increases, the rate of the back reaction will increase. Eventually the rates of formation and of dissolution of the $AB$ complex will be sufficiently close to each other that it is reasonable to write

$$k_1[A] \times [B] = k_{-1}[A \cdot B]$$

This can be rearranged to yield

$$\frac{[A] \times [B]}{[A \cdot B]} = \frac{k_{-1}}{k_1}$$

Since the two rates, $k_1$ and $k_{-1}$ are constants and independent of the concentrations of $A$ and $B$, their ratio $k_{-1}/k_1$ is a constant as well. As written above,

$$K_d = \frac{[A] \times [B]}{[A \cdot B]}$$

that is, when equilibrium has been reached, the equilibrium dissociation constant equals the concentration of the dissociated species divided by the concentration of complex. Thus, the resulting constant is called a dissociation constant, and often will be written as $K_d$. Inverses of the preceeding equations can also be written, in which case, the resulting equilibrium constant, is usually called an association constant and

$$K_a = \frac{[A \cdot B]}{[A] \times [B]}.$$

**How Complex Concentration Varies with Ligand Concentration**

Most generally, as a reaction between molecules of species $A$ and $B$ proceeds or when the reaction has reached equilibrium, the initial concentrations of $A$ and $B$ are reduced by the incorporation of some of these molecules into the $AB$ complex. Frequently however, in reactions of interest to biologists and biochemists, the concentration of one of the reactants is much greater than the other. Typically, this is the case of a protein binding a small molecule ligand. The ligand may be present at concentrations greater than $10^{-4}$ molar while the protein is present at concentrations below $10^{-5}$ molar. Even if the protein is fully bound with ligand, the reduction in the concentration of ligand is negligible and can be ignored. Thus, in the reaction,

$$P + L \rightleftharpoons P \cdot L$$

let $P_{tot}$ represent the total concentration of the protein. This equals the amount of free $P$ plus the amount of $P$ with bound ligand, $P \cdot L$. The system is then described by the two equations in which $P_{tot}$ and $L$ are known $P$ and $PL$ are variables that we wish to know. The two equations involving the known and unknown parameters can be written,

$$\frac{[P] \times [L]}{[P \cdot L]} = K_d$$

$$[P] + [P \cdot L] = [P_{tot}]$$

As is standard in solving such sets of equations, one of the equations is solved for one of the unknowns in terms of the other factors, and then this is substituted in the other equation. Solving the first equation for $[P \cdot L]$ gives

$$[P \cdot L] = \frac{[P] \times [L]}{K_d}.$$

And substituting this in the second equation and simplifying gives

$$[P]\left(1 + \frac{[L]}{K_d}\right) = [P_{tot}]$$

$$\frac{[P]}{[P_{tot}]} = \frac{1}{1 + \frac{[L]}{K_d}}$$

This can then be further simplified to give

$$\frac{[P]}{[P_{tot}]} = \frac{K_d}{K_d + [L]}$$

and substituting in the first equation gives

$$\frac{[P \cdot L]}{[P_{tot}]} = \frac{[L]}{K_d + [L]}.$$

This result is known as the Michaelis-Menten equation and sometimes as the Langmuir isotherm. Figure 3-1 shows a plot of this function.

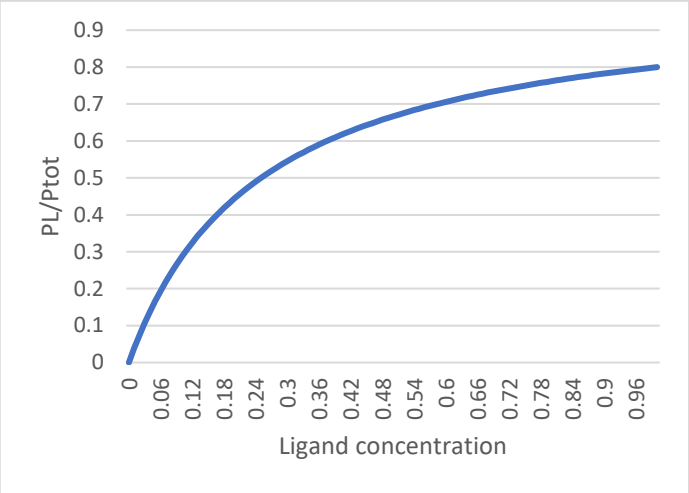



Figure 3-1 Plot of $P \cdot L/P_{tot}$ as a function of $L$ for $K_d = 0.25$.

Note that at values of $L$ well below the $K_d$ the concentration of the $PL$ complex increases in proportion to $L$. When the concentration of $L$ equals $K_d$ half of $P$ remains free and half is in the $PL$ complex, and when $L$ becomes very large compared to $k_d$, $P \to 0$ and $PL \to P_{tot}$. The Michaelis-Menten equation is often used to describe the relationship between the rate of an enzyme reaction, $V$ and the concentration of the enzyme's substrate where the rate of the reaction is approximated as being proportional to the concentration of the enzyme-substrate complex. The maximum rate would occur when all the enzyme is complexed with substrate so that if $V_{max}$ is the maximum possible rate,

$$\frac{V}{V_{max}} = \frac{[P \cdot L]}{[P_{tot}]} = \frac{[L]}{K_d + [L]}$$

or, as it is commonly written where $P$ and $L$ become $E$ and $S$ for enzyme and substrate

$$V = \frac{V_{max} \times S}{K_d + S}.$$

**The Exponential Relationship Between Free Energy Differences and Equilibrium Constants**

Thermodynamics proves to be a valuable tool in describing and predicting extents of biological reactions and processes. This section and the next present a simple way to think about the

meanings of the thermodynamic quantities free energy, enthalpy, and entropy, and the plausibility of the Boltzmann distribution. Then several examples will be presented of the use of these thermodynamic quantities.

Consider a collection of molecules, each of which can be in either of two states, $A$ or $B$ and that these states are of different energy as indicated in the energy diagram, Figure 3-2.

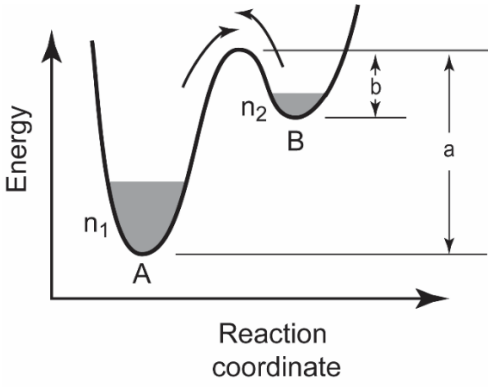



Figure 3-2 An energy level diagram showing the existence of two stable states, A, and B that are occupied at equilibrium by $n_1$ and $n_2$ molecules. All the molecules in state A are approximated as having the same potential energy, and similarly for the molecules in state B.

Suppose that for a molecule to go from state A to state B, a noncovalent bond must be broken. That makes state B be of higher energy than state A. The energy difference between the two states makes it more likely that a molecule will be found in state $A$ than in state $B$. It also means that the number of molecules in state $A$, $n_1$, will be greater than the number in state $B$, $n_2$. At equilibrium however, the rate of transfer from state A to state B must equal the reverse rate. Each of these rates will be proportional to the numbers of molecules in the two states and will also be some function $f$ of the energy required to overcome an energy barrier between the two states. Hence,

$$n_1 f(a) = n_2 f(b)$$

If the barrier between the states were increased, Figure 3-3, the rates of transfer between the states will change, but the distribution $n_1, n_2$, cannot change.

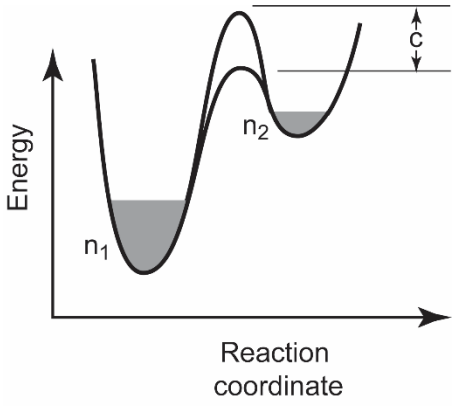



Figure 3-3 Raising the energy barrier separating the two states $A$ and $B$.

The reason it cannot change is that if the height of an energy barrier between the two states were to change the relative occupancy of the two states, then the system could be a source of energy. This follows because raising and lowering the barrier would allow the cyclic overpopulation of one state that could drive the equivalent of a turbine and hence be a source of energy. This is impossible so that it can be concluded that the relative occupancy of the two states cannot be altered by the height of an energy barrier between the states. Hence, it must also be true that

$$n_1 f(a+c) = n_2 f(b+c)$$

Dividing the first expression by the second gives

$$\frac{f(a)}{f(a+c)} = \frac{f(b)}{f(b+c)}$$

There is only one function that can satisfy this relationship for all $a, b,$ and $c$. That is

$$f(x) = \mathrm{e}^{-\beta x}$$

for any $\beta$. The equilibrium constant between states A and B then must have the form

$$K = \frac{n_2}{n_1} = \mathrm{e}^{-\beta(b-a)}$$

**Connecting the Exponential Relationship to Thermodynamics and a Way of Thinking About Entropy**

Suppose the state $B$ that was used in the previous section were to become degenerate, with an additional way that a molecule could have the same energy. This will be called a microstate. With $n_1$ molecules in state $A$, there could now be twice as many molecules in state $B$ as before, or

$$n_2 \rightarrow 2n_2$$

and $K \rightarrow K'$

$$K' = \frac{2n_2}{n_1}$$

$$= 2K$$

$$= 2\mathrm{e}^{-\beta(b-a)}$$

and making use of the definition of natural logs,

$$x = \mathrm{e}^{ln(x)}$$

$$K' = \mathrm{e}^{ln(2)}\mathrm{e}^{-\beta(b-a)}$$

$$= \mathrm{e}^{-\beta(b-a)+ln(2)}$$

Generalizing to any number, m, of microstates in $B$ gives

$$K' = \mathrm{e}^{-\beta(b-a)+ln(m)}$$

which shows that the natural log of the number of states is an important variable.

It is possible to carry this analysis a bit further before directly connecting it to the standard thermodynamic variables.

Suppose that the temperature of the $A - B$ system were to become very high. Then the thermal motion would become so vigorous that the energy difference between the states $A$ and $B$ would become irrelevant. On the other hand, the contribution to the equilibrium resulting from the number of microstates in $B$ is not sensitive to energy or thermal motion and therefore this term's contribution should be independent of temperature. One way to incorporate these two expected relationships between bond energy, number of microstates in B, and temperature would be for

$$K' = \mathrm{e}^{\frac{-\beta(b-a)}{T}+ln(m)}$$

A careful and lengthy analysis of the thermodynamics gives the result, which has the expected form,

$$K = \mathrm{e}^{-(\Delta H - T\Delta S)/RT}$$

and $\Delta G = \Delta H - T\Delta S$ where $G, H, S, T, R$ are free energy, enthalpy, entropy, absolute temperature, and the universal gas constant, $8.3 J/degree\ mole$.

Instead of considering concentrations of molecules in a system, sometimes a problem is best analyzed considering each molecule. Then, the probability, $P_i$, of a molecule being in state $i$ with an energy $e_i$ is given by

$$P_i = \mathrm{e}^{\epsilon_i/kT}$$

where $k$ is the Boltzmann constant, 1.38 x $10^{-16}$ ergs/degree and $e_i$ is in ergs. This last equation gives the relationship between the energy of a state and the probability that the system occupies this state. It is known as the Boltzmann distribution.

Thus, thermodynamics fits expectations from a simple analysis and identifies entropy as being proportional to the natural log of the number of indistinguishable states.

A brief aside: It may be distressing that the equation above for $K$ as a function of free energy is dimensionless and the equation defining $K$ in a reaction like $A + B \rightleftharpoons AB$ gives it a dimension. In this case, $molar^{-1}$. The explanation involves the definition of the thermodynamic standard state and will not be provided here. It suffices merely to ensure that one is working in the same system of units in which the standard state is defined, which usually is molar. In this case the appropriate units can be added to the dimensionless thermodynamic result.

**Thermodynamics in Protein Denaturation**

Experimentally, many proteins are observed to be stable and function well up to a critical temperature and to lose significant activity at temperatures not much higher than this. Some of these proteins regain full activity upon lowering the temperature. Physical studies show that for such proteins, the loss of activity parallels unfolding to a denatured state and that the folded native state of those proteins is in equilibrium with the denatured state. When the systems can be in equilibrium like this, they can be described thermodynamically. Microcalorimetry experiments have determined the unfolding enthalpy and entropy of the enzyme carbonic anhydrase, an enzyme mentioned several times in the book, to be, $\Delta H = 536\ kJ/mol$, $\Delta S = 1.6\ kJ/mol$.

The temperature at which the carbonic anhydrase is half denatured occurs when the equilibrium constant between the denatured and native states is one. Since $e^0 = 1$, this point occurs when $\Delta H = T\Delta S$, or T=335 degrees Kelvin, which is 62 degrees centigrade. The sharpness of the denaturation curve is illustrated by calculating the temperature at which the equilibrium constant between the denatured and native states is 10. From the equations above

$$RTlnK = -\Delta H + T\Delta S$$

Since $K = 10$, $lnK = 2.3$, $R = 8.3 \times 10^{-3}\ kJ/moledegree$, and with the thermodynamic parameters given,

$$8.3 \times 10^{-3} \times T \times 2.3 = -536 + 1.6T$$

$$0.019T = -536 + 1.6T$$

$$T = 339 \text{ degrees Kelvin or } 66 \text{ degrees Centigrade}$$

The protein loses nearly all of its activity upon a temperature increase of merely four degrees above the temperature at which it has lost half of its activity. Should both the enthalpy of unfolding and the entropy of unfolding each increase by a factor of ten, then a protein that is half denatured at 62 degrees would retain only 10% of its original activity at 62.4 degrees. This example illustrates how the enthalpy of unfolding, that is, the energetic cost of breaking the bonds that hold the protein in a specific three-dimensional conformation, competes against the entropy of unfolding. In most proteins the balance between the forces for remaining folded just barely overcome the forces for unfolding. When the temperature rises too much, the protein unfolds.

The strengths of the most important bond types in biology and biochemistry as shown in Table 3-1, should be compared to the average kinetic energy of an atom or molecule due to thermal motion. Thermal motion is capable of breaking one or two hydrogen bonds, but not covalent bonds.

Table 3-1 Energies

| Bond or Kinetic Energy | Enthalpy of Bond Formation or Motion |
|---|---|
| Covalent | 150 kJ/mole |
| H bond | 1-6 kJ/mole |
| Van der Waals | 4 kJ/mol |
| Thermal motion at room temperature | 2.5 kJ/mole |

**The Chelate Effect with Biomolecules**

Consider a dimeric DNA binding protein that binds to a DNA binding site with a dissociation constant of $10^{-12}$ molar. From the relationship between a dissociation constant and the free energy change in the binding reaction,

$$K_d = \mathrm{e}^{-\Delta G/RT}$$

The free energy of binding at room temperature is given by the following.

$$10^{-12} = \mathrm{e}^{-\Delta G/1.98\ 10^{-3} kcal\ K^{-1} mol^{-1} \times 293 K}$$

giving

$$\Delta G = -16 \frac{kcal}{mol}.$$

In this binding reaction it is very tempting to think that each subunit of the protein contributes equally to the $16\,kcal/mol$ of the free energy. If this were true, then each subunit would contribute $-8\,kcal/mol$, and we would expect that a monomer of the protein would then have a dissociation constant of $10^{-6} molar$. In reality, the binding of a monomer of such a protein would have to be much weaker than this. That is, its dissociation constant would be much larger than $10^{-6}$. The discrepancy is explained by the chelate effect.

Consider the binding of the dimeric protein to consist of two steps, Figure 3-4. First, one subunit of the dimer binds to half of the binding site, and then in the second step, the second subunit

Figure 3-4 The multiple steps in the binding of a dimeric protein to two half-sites on DNA.

binds to the other half of the binding site. In the first step, bonds between the first subunit and the half-site contribute, $\Delta H$ to the binding free energy. This binding energy is opposed by the entropic cost of immobilizing the dimer on the DNA. Formerly, it was free in solution. That is,

the system contains a great many microstates with the protein free in solution, and only a few with the protein immobilized with the first subunit bound to one half site. Therefore, in binding, a substantial entropic cost is extracted from the binding enthalpy. As the second subunit binds to the other DNA half site, it, too, contributes $\Delta H$ to the binding of the dimer. However, because this second subunit is already positioned very close to its DNA binding site, its binding pays only a small entropic cost, and virtually all the binding enthalpy can go into holding the dimer on the DNA. Consequently, adding a second subunit to a monomer and doubling the size of the DNA binding site greatly increases the binding affinity.

**DNA Binding by Two Loosely Connected DNA Binding Domains**

This section concerns an example that illustrates the chelate effect. Suppose two DNA binding domains are tightly connected by a very short but flexible linker to form a functional dimer. If the linker allows both subunits of the dimer to contact a pair of DNA sites at the same time and without significant strain, then the chelate effect as described in the previous section should apply. This would allow the dimer to bind with very high affinity. As the length of the linker increases, the magnitude of the chelate effect that gives the high binding affinity will decrease because of the increasing entropic cost of immobilizing the second subunit to its binding half site. Eventually, when the linker is very long, the two DNA binding domains will bind almost independently of one another.

Three states of the protein must be considered as shown in Figure 3-4. These are the protein free in solution with neither subunit bound to DNA, the protein with one subunit bound,and one unbound, and finally the state with both subunits bound to the DNA. Let $[P_0]$ be the concentration of free dimer, $[P_1 \cdot D]$ the concentration of DNA with one of the two subunits bound, $[P_2 \cdot D]$ be the concentration of DNA with both subunits bound, $K$ be the dissociation constant between a single subunit and its binding site on the DNA and $C$ be the effective concentration in the $P_1 D$ complex of the unbound subunit in the vicinity of its DNA binding site. The first binding step is described by the equation

$$\frac{2[P_0] \times 2[D]}{[P_1 \cdot D]} = K$$

Which upon rearrangement becomes

$$\frac{[P_1 \cdot D]}{[D]} = \frac{4[P_0]}{K}.$$

The factor of four comes from the presence of two DNA binding domains in the protein and the two DNA binding sites on the DNA.

In the second step in the full binding reaction, one DNA binding domain is already bound and the other is held near its DNA binding site where it can bind or not depending on its effective concentration, $C$, and the dissociation constant. By analogy to the equation for the binding of the first subunit, the binding of the second subunit is described by

$$\frac{[P_2 \cdot D]}{[P_1 \cdot D]} = \frac{C}{K}$$

Substituting for $[P_1 \cdot D]$ from the first binding reaction and inverting gives

$$\frac{[P_0] \times [D]}{[P_2 \cdot D]} = \frac{K^2}{4C}$$

Therefore, the effective dissociation constant for the binding of the two DNA binding domains tethered together and binding to DNA is

$$K_{eff} = \frac{K^2}{4C}.$$

Approximating the linker to be perfectly flexible and of length $r$ Angstroms, the unbound subunit is free to occupy any position within the sphere of radius $r$. One molecule per such a sphere would have an effective concentration at the DNA binding site of whatever molar concentration in solution gives an average volume per molecule equal to the volume of the sphere. This can be calculated as the moles of equivalent sphere volumes that fit into a volume of one liter. This is

$$C = No.\,spheres\,per\,liter \times \frac{1}{Avogrado's\,no.}$$
$$= \frac{Volume\,of\,liter}{Volume\,of\,sphere} \times \frac{1}{Avogrado's\,no.}$$
$$= \frac{\left(10cm \times 10^8 \text{Å}/cm\right)^3}{\frac{4}{3}\pi r^3 \times 6.02 \times 10^{23}}.$$

Table 3.2 shows for a set of tether lengths=radii, the resulting effective concentration, and the effective dissociation constant if a monomer has a dissociation constant of $10^{-4} molar$.

Table 3-2

| r, Angstroms | $C_{eff}$, Molar | $K_{eff}$, Molar |
|---|---|---|
| 5 | 3.2 | $7.8 \times 10^{-10}$ |
| 15 | 0.119 | $2.11 \times 10^{-8}$ |
| 50 | 0.0032 | $7.81 \times 10^{-7}$ |
| 150 | 0.000119 | $2.11 \times 10^{-5}$ |
| 250 | 0.0000256 | $9.77 \times 10^{-5}$ |

For a linker length of 5 Angstroms, a substantial chelate effect is present and the dimer has a dissociation constant smaller than the square of the monomer dissociation constant. As the linker becomes longer, the binding becomes weaker, and when it reaches a length of 250 Angstroms, the dimer binds only a little more tightly than a monomer.

At the other end of the length range, a minor paradox appears. Why don't at least some dimeric DNA binding proteins bind with near infinite affinity? If the first DNA binding domain to bind were to perfectly position the second DNA binding domain, why wouldn't the combined binding affinity be infinite? The reason this cannot occur is that the combination of thermal motions and flexibility in both the domains and in the DNA binding site prevents perfectly positioning and orienting the second DNA binding domain.

**The Equipartition Theorem**

The average energy of an atom, molecule, or system equals the sum over all possible states of the system, of the energy of a state times the probability of being in that state.

With increasing temperature, higher and higher energy states of systems are increasingly occupied. Equivalently, the average energy of a system increases with temperature, but what sort of function of temperature is this increase? Remarkably, for many systems, the increase is directly proportional to the temperature and independent of the details of the system. For each degree of freedom, the average energy is $1/2\,KT$, where $K$ is the Boltzmann constant and $T$ is the absolute temperature. For example, a monatomic gas molecule can move in three dimensions and has three degrees of freedom, and therefore the average energy of a molecule of such a gas is $3/2\,KT$ and is independent of the mass of the molecule. This fact is utilized later in the chapter on hemoglobin.

The finding described in the previous paragraph is known as the equipartition theorem because each degree of freedom of a system carries the same amount of energy. The theorem applies to systems in which the energy is proportional to the square of some variable. For example, in the case of a gas molecule, the kinetic energy is $1/2\,mv^2$. In the case of an atom constrained in a potential well, the energy of stretching the constraining bond is proportional to the square of the displacement from the equilibrium position. Many systems possess this property since in many systems the force for restoring a system to the equilibrium position is proportional to the displacement from equilibrium. Since the work required to generate a displacement, or equivalently, the energy in the displacement, is the integral of force times distance,

$$W = \int F(x)dx$$

$$= \int -cxdx$$

$$= -\frac{1}{2}cx^2$$

such a system satisfies the requirements of the equipartition theorem.

The two critical facts about the equipartition theorem are that the average energy is proportional to temperature and that the average energy is independent of the constant of proportionality between the energy and the displacement. For example, in the case of gas molecules, the energy is independent of the mass of the molecules, and in the case of an atom bound by a restoring potential, is independent of the strength of the restoring force.

Proof of the equipartition theorem is straightforward. If the energy $E(x)$ of a system is proportional to the square of the variable $x$, the relative probability of having the value of x is given by the Boltzmann distribution

$$p(x) = \mathrm{e}^{-x^2/KT}$$

Since the sum of the probabilities for all $x$ must be one,

$$A\int_0^\infty \mathrm{e}^{-x^2/KT}\,dx = 1$$

or

$$A = \frac{1}{\int_0^\infty \mathrm{e}^{-x^2/KT}\, dx}$$

This provides the value of A. Next, since the average energy of the system equals the integral over all $x$ of the energy times the probability of possessing that value of x, the average energy, $\bar{E}$ is

$$\bar{E} = A \int_0^\infty x^2\, \mathrm{e}^{-x^2/KT} dx$$

Evaluating the integrals gives

$$\bar{E} = \frac{KT}{2}.$$

Problems

1. Earlier we saw that if the rate at which a bacterial culture grows is proportional to the number of bacteria present, the number increases exponentially. What about dying? Simulate on a spreadsheet how the number of viable bacteria change with time if the rate of death of bacteria in a culture is proportional to the number still alive.
2. At what temperatures does the average kinetic energy of an atom or molecule equal the bond energy of Van der Waals, hydrogen, and covalent bonds?
3. The text gave an example where the amount of activity of a protein remaining as a function of temperature was calculated from the $\Delta H$ and $\Delta S$ of denaturation. Explain how the reverse can be determined, that is, how the $\Delta H$ and $\Delta S$ of denaturation can be determined by measuring the amount of folded protein remaining as a function of temperature.
4. Look up the definite integrals of the last section and verify that they indeed lead to
$$\bar{E} = \frac{KT}{2}$$
5. If a protein binds to DNA at the diffusion limit, and protein is present in a reaction at $10^{-10}M$ and DNA is present at $10^{-11}M$, about how long after mixing the two would 10% of the DNA be bound with the protein if dissociation of the protein from DNA can be neglected at these early times in the reaction?
6. Why is it meaningless to speak of the thermodynamic parameters associated with frying an egg?

# 4. Protein Synthesis, Measuring its Elongation Rate

**Why Measure the Elongation Rate?**

Sensible design of experiments to learn more about nature, sensible approaches to solving a problem in bioengineering, and sensible research or treatment of many medical problems all require a thorough understanding of the biology involved. As this chapter shows, even asking seemingly simple experimental questions requires careful consideration of a large number of biological variables. In this chapter, the polypeptide elongation rate and the considerations required for its meaningful measurement will be considered.

Because the fundamental physical, chemical, and biochemical principles apply to all cells, the elongation rate found in bacteria ought to be similar to the rates found in other cells. Properly measuring this elongation rate involves consideration of multiple factors. Many of these or similar factors are common to measurements made for other purposes on other biological systems and ignoring any of them can make the most elegant and otherwise careful measurements, design process, or treatment entirely worthless.

**Design of an Elongation Rate Measurement**

A conceptually simple way to measure protein elongation rate is simply to start the synthesis of a particular protein and then measure the time required for the first appearance of its newly completed polypeptide chains. Figure 4-1 shows a timeline for the execution of such a measurement. Multiple factors must be considered even for an experiment this simple. One

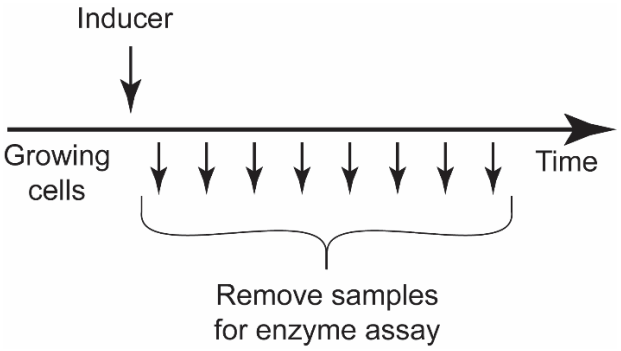



Figure 4-1 A timeline of an experiment to measure the time required after addition of inducer to a culture until newly synthesized enzyme appears.

important consideration in the experimental design is the choice of the protein and how it will be detected. Most physical detection methods like staining of proteins after electrophoretic separation or mass spectrographic separation require large numbers of the protein molecule and may also be cumbersome or expensive. On the other hand, measurement of enzyme levels can be made both sensitive and convenient by measuring the activity of an enzyme in generating its product. The turnover numbers of different enzymes in producing product varies from lower than one per second to over a million per second, with numbers in the range of $10^3$-$10^5$ per second being typical. With these rates, even very low enzyme levels, occasionally down

to a single molecule, can be measured by allowing the sufficient reaction time in the enzyme's assay for the accumulation of sizeable numbers of product molecules.

The following factors are important in measuring elongation rate using an enzyme. What cells should be used, and are they growing normally? Next, as indicated in Figure 4-1, the synthesis of the enzyme must be induced, typically this is done by adding an inducer molecule to the culture medium. How rapidly does the inducer enter the cells? After entering the cells, the inducer must lead to initiating synthesis of the messenger RNA for the enzyme and then translation of the messenger must begin. During the period after initiating enzyme synthesis, samples will be withdrawn from the culture medium for assay of the enzyme. In order that the samples reflect enzyme levels at their time of withdrawal, it is necessary to stop the elongation process in the sample as soon as it is taken. Next, while the full polypeptide of the enzyme might have been completed at the time of taking a sample, the peptide may require additional time to fold to an enzymatically active structure. It is also necessary that a polypeptide only acquire enzymatic activity after it has been fully synthesized. An additional factor to consider is whether the enzyme remains stable and active throughout the enzyme assay reaction period. A final consideration is how the data from the enzyme measurements will be analyzed to then yield a value for the elongation rate. These factors will be considered in the sections that follow.

**State of the Cells**

On one hand cells do all the work. They require a few basic molecule types in their growth medium and they synthesize internally the hundreds of other small molecules and the macromolecules that they need for growth and division. On the other hand, many factors can still interfere with proper cell growth and prevent experimental results from having any value. Growth medium can easily be formulated incorrectly, or the standard salts added to make the growth medium may be deficient in a required trace element, the growth temperature may not be held steady at the correct value, the culture may be contaminated with foreign cells or a virus, or cell growth rate may be limited by insufficient nutrients or oxygen limitation. Finally, the cells may not be in balanced exponential growth. Such a state of growth is necessary for reproducibility in experiments, both within a lab and between labs. As cells grow into senescence or stationary phase and cease net growth, a few proteins are induced to assist survival of the cells and the synthesis of a few proteins begins to fall even before there is a noticeable change in growth rate. These changes can dramatically alter some properties of the cells like the inducibility of some enzymes.

To avoid problems caused by cells that have not yet fully adapted to balanced, exponential growth, the cells must be grown for many generations without reaching cell densities high enough to initiate the stationary phase response. In some cases, this means growing cells for at least ten doublings, that is by more than a factor of 1,000 from the time that they have been diluted from a culture that had reached stationary phase. Often this requires several dilutions during this phase of the growth. Of course, the cells must be diluted into prewarmed growth medium to avoid generating a heat shock or cold shock response. Monitoring cell density during the growth period can indicate whether or not the cells are growing properly. As seen in the earlier chapter, the growth should be exponential, which will yield a straight line on a log plot, Figure 4-2. Even a small deviation from a straight line can indicate serious problems with the cells. Following the growth in this way also allows prediction of the time at which the culture will be at a density suitable for beginning the experiment.

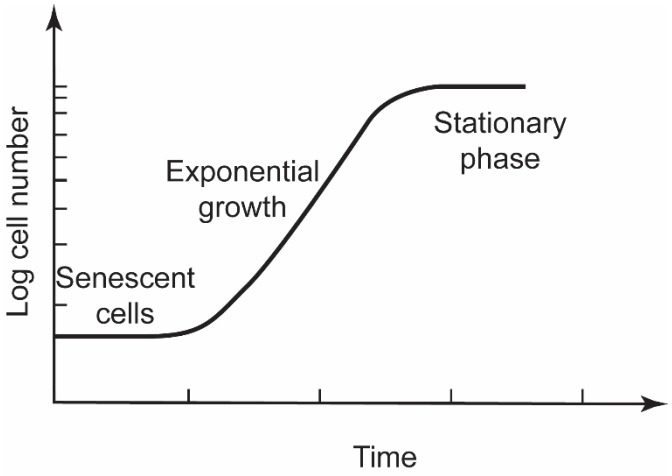


Figure 4-2 Growth phases of a bacterial culture plotted on a semi log graph.

**How Rapidly Does Inducer Enter Cells?**

If inducer enters cells slowly, induction will be delayed until intracellular concentrations reach inducing concentrations. This delay will add to the apparent time required to synthesize the protein. Therefore, the smaller the limits that can be placed on the delay, the more accurate and useful will be the final determination of the rate of protein elongation. If the delay cannot be estimated, then the delay could have any value between instantaneous entry to a delay extending until just before the time at which induced enzyme appears.

Inducer could enter cells via passive diffusion or via active transport mechanisms. While many active transport systems are inducible, since we are concerned with the initial seconds of the induction of the enzyme, changing activities of an induced active transport system will not be a confounding variable. Therefore, over the time scales and inducer concentrations important in the experiment, the majority of inducer likely will enter the cells via passive diffusion. In this case, the rate of inducer entry into the cells will be proportional to the concentration difference across the cell membrane, $a(C_{out} - C_{in})$. This is just like Fick's law that we saw with respect to diffusion, but in this case, a membrane separates the solutions at two concentrations. The rate of change of the internal concentration inducer will be

$$\frac{dC_{in}}{dt} = Net\ Flow\ rate\ in/Volume$$
$$= \frac{a(C_{out} - C_{in})}{V}$$

This is solved by separating the variables $C_{in}$ and $t$, and integrating

$$\int_0^{C_{in}} \frac{dC_{in}}{(C_{out} - C_{in})} = \int_0^t \frac{adt}{V}$$

Giving

$$-ln(C_{out} - C_{in)} - ln\ (C_{out)} = \frac{at}{V}$$

Rearranging gives

$$C_{in} = C_{out}\left(1 - e^{-at/V}\right)$$

Figure 4-3 shows a plot of this as a function of $at/V$. This is the form of the result for any quantity that approaches its final value at a rate proportional to the difference between its present value and the final value. For example, it applies to the voltage across a capacitor that is being charged through a resistor or the accumulated decays of a radioactive sample.

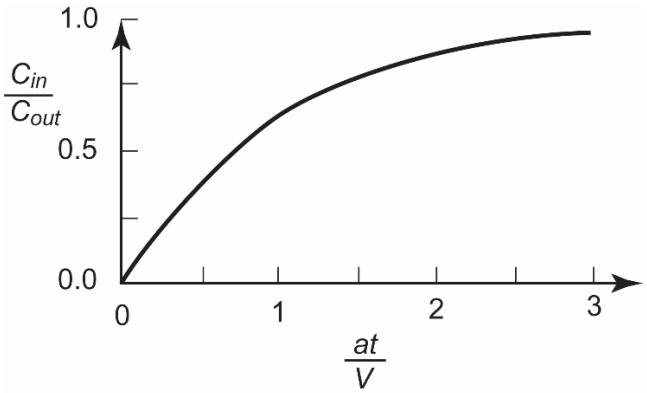



Figure 4-3 Plot of $C_{in}/C_{out}$ as a function of $at/V$.

For small $t$, using the approximation for small $x$, $e^x \approx 1 + x$, $C_{in}$ reduces to the linear relationship,

$$C_{in} = C_{out}\frac{at}{V}$$

If significant induction only occurs after the internal inducer concentration has exceeded a certain value and if the concentrations involved lie in the range where the response is linear, the apparent delay in induction after addition of inducer will be roughly

$$Delay = \frac{C_{crit}}{C_{out}}\frac{V}{a}$$

That is, the time after inducer is added until induction occurs is proportional to $1/C_{out}$. The assumptions that have been made up to this point can be checked by examining the delay as a function of the concentration of inducer. A plot of the delay as a function of $1/C_{out}$ should be close to linear, Figure 4-4.

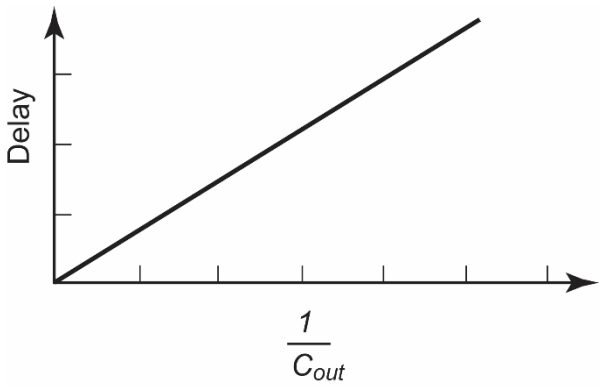


Figure 4-4 An idealized plot of the apparent delay of induction as a function of $1/C_{out}$.

In such a case, the time required for inducer to enter the cells can be made very small by using a sufficiently high concentration of inducer. If a high concentration cannot be achieved, a value for the effective delay can still be determined from such a graph by extrapolation.

**How to Take Samples and Stop Synthesis**

When a sample is taken from the induced culture, further protein synthesis must be immediately stopped. This requires blocking additional protein elongation without denaturing or inactivating any of the completed polypeptide chains of the induced enzyme. Ideally, energy production in the cells should not be blocked in case any steps in protein folding require a source of energy like ATP. Addition of chloramphenicol is the most convenient and safest way to do this. This drug blocks protein synthesis at the elongation step. Its rate of blocking protein synthesis can be tested in a side experiment in which a radioactive amino acid and chloramphenicol are simultaneously added to cells. After several minutes, a sample is taken, and the amount of radioactivity in cellular protein is measured. As it is found to be negligible when the drug is used at sufficiently high concentrations, the chloramphenicol must have acted promptly to block further protein synthesis.

Not yet answered is the question of whether or not the radioactive amino acid is quickly taken up and incorporated into growing polypeptide chains. Another separate side experiment can be used to show this. In this test, the radioactive amino acid is added, and at intervals thereafter, samples are taken into a strong acid that quickly and completely kills cells and denatures all their proteins. Then, the amount of radioactivity in protein is measured. This experiment shows that following its addition, some of the amino acids are incorporated into protein with only a few seconds' delay. This means that the intracellular pool sizes of those amino acids are very small and contain only sufficient amounts of amino acid to support several seconds worth of protein synthesis.

**How to Measure Enzyme Levels**

Once a sample has been taken from the induced culture, it can be added to a tube containing chloramphenicol that is maintained at the growth temperature for at least a few minutes. This allows time for completed polypeptide chains of the enzyme to fold into their active conformation. Only then can the measurement of the levels of enzyme in the multiple samples be begun. For an enzyme like β-galactosidase, measurement of the enzyme level is particularly

simple. The colorless lactose analog, ortho-nitro-phenyl-galactoside, ONPG, is cleaved by β-galactosidase to nitro-phenol, which is strongly colored yellow. Hence, by measuring the time of incubation with ONPG and by measuring with optical density the amount of nitro-phenol that has been produced, the amount of enzyme present in the sample can be determined.

**Expected Induction Kinetics and a Problem**

The addition of inducer to a growing culture leads to the initiation of transcription of the messenger RNA for the enzyme. Translation does not begin until ribosomes bind to the messenger and begin translation. While this could take some time, in bacteria the delay between initiation of transcription of a messenger RNA molecule and initiation of translation of that messenger is on the order of seconds. This is known from examining electron micrographs of gently ruptured cells that show DNA, RNA polymerase molecules transcribing from the DNA, mRNA molecules extending from the RNA polymerase molecules, and ribosomes attached to the mRNA. Newly extruded mRNA molecules are seen to contain bound ribosomes, thus transcription and translation occur coordinately, as sketched in Figure 4-5.

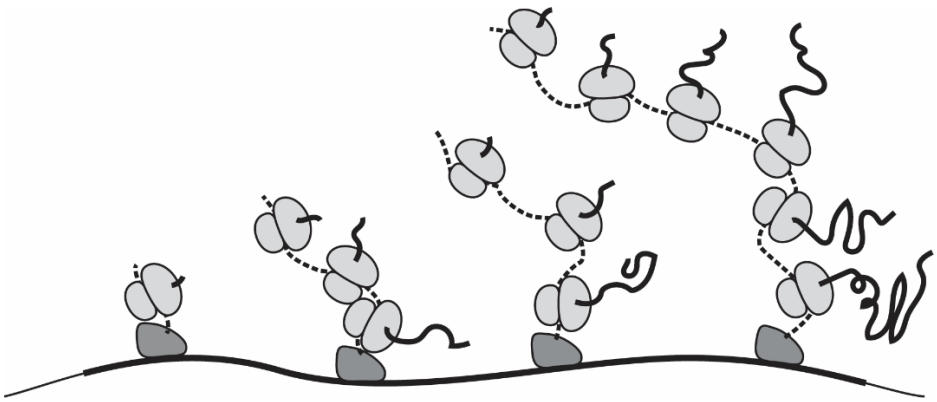


Figure 4-5 Schematic showing the cotranscription-translation of a gene. The gene is transcribed from left to right. On the far left, is an RNA polymerase molecule that just initiated transcription and a ribosome has begun translation on the short transcript projecting from the polymerase. The translation product is just emerging from the ribosome. To the right of this first polymerase is one that initiated transcription somewhat earlier. Its longer transcript has been engaged by three ribosomes and protein products of different lengths have emerged from these ribosomes. Still further to the right are polymerases, ribosomes, and translated protein that are more advanced. On the far right, the first ribosome that bound to the polymerase that first began translation is about to complete translation of the gene product. Its protein is already partially folded.

Following the addition of inducer, the initiation of transcription of messenger RNA molecules begins at a constant rate. If the RNA polymerase molecules require a time $\Delta t$ to traverse the entire length of the gene, then after a delay of a little longer than $\Delta t,$ completed messenger will

begin to accumulate at a constant rate. That is, $Completed\ messenger \propto t - \Delta t$. Since ribosomes follow right along with the RNA polymerase, they will complete translation of the messenger at a rate proportional to the amount of completed messenger, and since the amount of enzyme will be the integral of its rate of synthesis, the amount of completed enzyme will be proportional to $c(t - \Delta t)^2$, Figure 4-6.

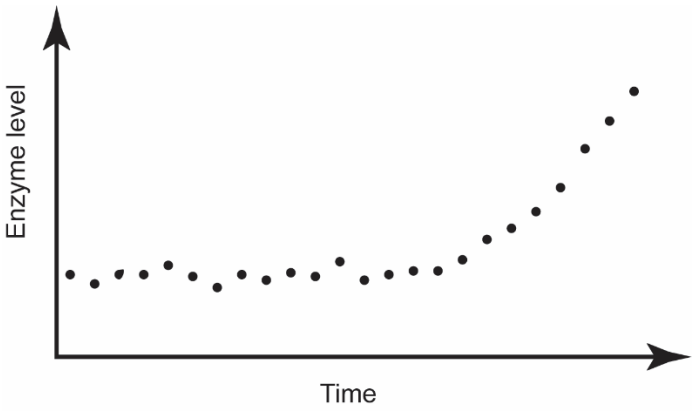



Figure 4-6 The levels of an induced enzyme following induction at time zero. The scatter in the points represents typical fluctuations in experimental data.

To determine the elongation rate, we need $\Delta t$, and, of course, the number of amino acids in the polypeptide of the enzyme. The total amount of enzyme present at any time equals the basal, or uninduced level of the enzyme plus the newly synthesized enzyme. The basal level can be subtracted from the measurements of total enzyme to yield the amounts of newly synthesized enzyme. Determining exactly where a plot of enzyme level as a function of time first increases above this basal level poses a minor problem in addition to the problem generated by minor inaccuracies in the determination of enzyme levels. The problem is that the initial appearance of enzyme is flat. Seeing precisely where it begins to rise above the basal level can be somewhat uncertain.

**Classical Method for Analyzing the Data**

One way to identify the time at which induced enzyme first appears is to note that if

$$E(t)_{induced} \propto (t - \Delta t)^2$$

Then

$$\sqrt{E(t)_{induced}} \propto (t - \Delta t)$$

and determining $\Delta t$ from a plot of $\sqrt{E(t)}$ is straightforward, Figure 4-7.

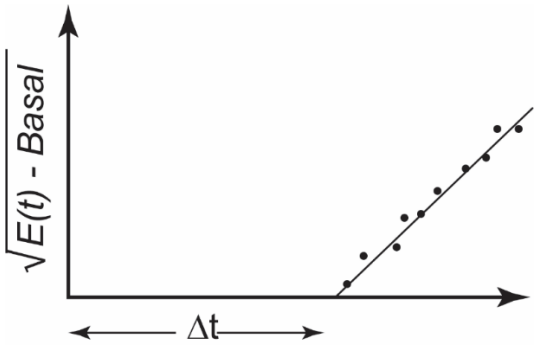



Figure 4-7 Plot of the square root of the enzyme level above the uninduced basal level. The intercept of the plot with the time axis is $\Delta t$.

An additional virtue of this approach is that all the measurements of the enzyme level contribute to identifying $\Delta t$. Before computers and calculators were widely available, it was common to transform an equation giving the expected functional response to a form in which the data would fall on a straight line whose slope or intercepts with the axes would provide a desired quantity. With the wide availability of computers, better methods for analyzing the data have become widespread. When all the precautions mentioned above have been taken, measurement of the protein elongation rate in the bacterium *Escherichia coli* growing at 37° C yields a polypeptide elongation rate of about 15 amino acids per second.

**Curve Fitting via Computer and an Example of Manual Least Square Fitting**

These days, the best way of extracting important parameters from experimental data utilizes computer programs to find the best fit between experimental data and an equation that ought to describe the data. The standard criterion for fitting an equation to experimental data is known as least square fitting. This means that the variables of the equation are adjusted to give the lowest value to the sum of the squares of the differences between the data and the values predicted by the equation chosen. In our case of the induced enzyme, we would provide the experimental data of the enzyme levels as a function of time. Then, the program would be instructed to fit to the following equation by finding values for $a1, a2,$ and $a3$ that give the best least squares fit to the data.

$$y = a_1 + a_2(t - a_3)^2$$

Although procedure that is used by the computer program sounds complex and obscure, it is not. As an example, the following shows the steps of finding a least squares fit of the equation

$$y = a_1 + a_2 x$$

to the three data points, $x = 1, y = 1;\ x = 2, y = 3;\ x = 3, y = 4$. Table 4-1 shows the details of the calculation of the sum of the squares of the differences between the data and the equation.

Table 4-1

| $x$ | $y$ $Eqn$ | $y$ $Data$ | $\Delta$ $Eqn - Data$ | $\Delta^2$ |
|---|---|---|---|---|
| 1 | $a_1 + a_2$ | 1 | $a_1 + a_2 - 1$ | $a_1^2 + a_2^2 + 1 + 2a_1a_2 - 2a_1 - 2a_2$ |
| 2 | $a_1 + 2a_2$ | 3 | $a_1 + 2a_2 - 3$ | $a_1^2 + 4a_2^2 + 9 + 4a_1a_2 - 6a_1 - 12a_2$ |
| 3 | $a_1 + 3a_2$ | 4 | $a_1 + 3a_2 - 4$ | $a_1^2 + 9a_2^2 + 16 + 6a_1a_2 - 8a_1 - 24a_2$ |

$$Sum\ \Delta^2 = 3a_1^2 + 14a_2^2 + 26 + 12a_1a_2 + 16a_1 + 38a_2$$

This is an equation in the two unknowns, $a_1, a_2$ and we wish to minimize the value of the equation. Often complicated equations in multiple unknowns can only minimized numerically, that is, substitute values for the unknowns, see what the result is, then adjust the values to improve the result. Usually, this approach is greatly aided by making decent guesses for the initial trials. Our case is sufficiently simple that the minimum can be found by locating the values for $a_1$ and $a_2$ such that

$$\frac{\partial((Sum\ \Delta)^2)}{\partial a_1} = 0$$

and

$$\frac{\partial((Sum\ \Delta)^2)}{\partial a_2} = 0$$

which gives

$$6a_1 + 12a_2 + 16 = 0$$

and

$$28a_2 + 12a_1 + 38 = 0$$

This pair of linear equations is easily solved to yield $a_1 = -1/3, a_2 = 1.5$ so that the least squares best fit to the data is $y = -.333 + 1.5x$, Figure 4-8.

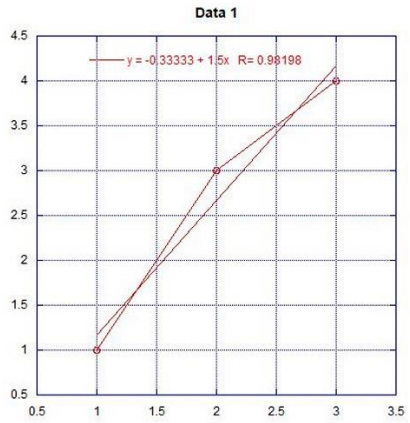



Figure 4-8 Plot showing the original data points and the least squares fit of a straight line to the data.

Problems

1. Suppose that an amino acid is present in the cell's growth medium allows the cells to grow normally without synthesizing any of the amino acid themselves. Without trying to directly measure free amino acids, how could you determine the size of the internal pool of amino acid not yet incorporated into protein?
2. Instead of fitting the data given in the text to $y = a_1 + a_2 x$ "fit" to $y = a_1 + a_2 x + a_3 x^2$.
3. Suppose an average bacterial cell contains 20,000 ribosomes, 2 x $10^{-13}$ gm protein, and has a doubling time of 45 minutes. What is the average protein elongation rate per ribosome?
4. Provide an argument that was not provided in the text that most polypeptide chains should not acquire biological activity until they are virtually complete.

# 5. Multisubunit Proteins and Allostery

**Why Nature Uses Multisubunit Proteins**

More than a third of the different proteins in most cells are found in oligomers rather than monomers. What sort of evolutionary pressures might have led to oligomer formation? One factor is protein stability. Larger proteins are more stable. Because the amino acid constituents of proteins are of finite size, the larger a protein, the greater the fraction of amino acid residues that will be interior. These can form more bonds to other amino acid residues and stabilize the protein rather than being on the surface and being unable to form as many stabilizing bonds per residue. That is, with increasing size, an increasing fraction of the bonds that can be made can go into maintaining the protein's structure whereas the entropic opposition to remaining folded increases with molecular weight at a constant rate.

Why not then, simply encode a very long polypeptide that contains multiple enzyme active sites? To gain the stabilizing effect that results from a large size, the regions responsible for the different activities would need to fold into a very large domain, and this appears to be difficult challenge for evolution to handle. Another reason is accuracy. The overall process of protein synthesis cannot be perfectly accurate. Translation mistakes would be costly if a single mistake were to inactivate an entire multiunit polypeptide. On the other hand, if multiple correct copies of a single unit peptide can assemble into an oligomer, an occasional fatal translation error would inactivate only that polypeptide unit. It would therefore not be incorporated into the multimer and not impair the activities of the other units. Those copies that had been synthesized without errors could still oligomerize and be functional.

Multisubunit proteins also allow for special properties resulting from the communication from one subunit to another. One such example is hemoglobin which is discussed later in the book. Another example would be a hetero oligomeric enzyme whose product is at the start of a multistep metabolic pathway leading to a product P. Such enzymes often have special regulatory properties. Generally, they contain two types of subunits, one which binds substrate(s) and catalyzes a chemical reaction, and one which binds the end product of the pathway. The binding of the product to the enzyme changes the structure of the enzyme to reduce its activity. This would be an example of feedback inhibition and it can prevent the unnecessary and potentially harmful overproduction and accumulation of the product P.

Multisubunit proteins are often called allosteric. Sometimes the term means other shape as the oligomer binds two or more ligands of different shape, and the binding of one affects the binding of the other. Other times the term means "other site", as something that happens at one site affects what happens at a different site. Finally, sometimes the term is applied to proteins in which a conformational change is known to be important. Probably it is best to expect a protein that is called "allosteric" to exhibit some sort of complex behavior. The following three sections address some of the simpler properties unique to multisubunit proteins.

**Homodimer, Dissociation Constants--the Protein or a Subunit?**

The previous chapter considered the binding of one ligand molecule to a protein. This section and the next two address two of the complications that can be encountered when more than one molecule can bind to a protein. Almost always when more than one molecule of a ligand binds, the binding is to separate subunits, and such will be assumed here. It is possible for the binding

of the first molecule of the ligand to affect the binding of the second. This binding can increase the binding affinity of the second molecule—called either cooperative or positive cooperativity, or it can decrease the binding affinity of the second molecule—called anticooperative or negative cooperativity.

The binding equations for oligomers can be written in terms of the concentrations of the oligomers, in which case the equilibrium constants are called macro equilibrium constants, or in terms of the concentrations of the individual subunits, in which case the constants are called micro equilibrium constants. Both work fine and make the same experimental predictions, but sometimes the use of one form is more appropriate than the other.

First, for the macro formulation, let $P$ represent the concentration of a dimeric protein and $L$ the concentration of ligand. Let $K_1$ be the equilibrium dissociation constant for the binding of the first molecule of ligand to $P$ and $K_2$ be the constant for the second molecule. In terms of the dimer concentrations, the relevant equilibria are:

$$P + L \rightleftharpoons P \cdot L_1$$

$$P \cdot L_1 + L \rightleftharpoons P \cdot L_2$$

and the equilibrium equations are:

$$K_1 = \frac{[P] \times [L]}{[P \cdot L_1]}$$

$$K_2 = \frac{[P \cdot L_1] \times [L]}{[P \cdot L_2]}$$

Since total protein equals unliganded plus singly liganded plus doubly liganded,

$$[P] + [P \cdot L_1] + [P \cdot L_2] = [P_{tot}].$$

Solving these three equations for $[P]$, $[P \cdot L_1]$, and $[P \cdot L_2]$ in terms of $K_1$, $K_2$, $[P_{tot}]$, and $[L]$ gives

$$[P] = \frac{K_1 K_2 [P_{tot}]}{K_1 K_2 + K_2 [L] + [L]^2}$$

$$[P \cdot L_1] = \frac{K_2 [L][P_{tot}]}{K_1 K_2 + K_2 [L] + [L]^2}$$

$$[P \cdot L_2] = \frac{[L]^2][P_{tot}]}{K_1 K_2 + K_2 [L] + [L]^2}.$$

These three quantities can be used to obtain the equations for various quantities of interest. For example, the amount of subunits with bound ligand will be $P \cdot L_1 + 2P \cdot L_2$, which is

$$\frac{(K_2 [L] + 2[L]^2)[P_{tot}]}{K_1 K_2 + K_2 [L] + [L]^2}.$$

For the formulation in terms of micro equilibrium constants, let $P$ and $L$, as before, represent the concentrations of the dimer and of the ligand, but let $K'_1$ and $K'_2$ be the equilibrium dissociation constants in terms of the concentrations of monomers of the ligand binding subunit. Then, the dissociation constant for the binding of the first molecule of ligand is

$$K'_1 = \frac{2[P] \times [L]}{[P \cdot L_1]}.$$

The factor of two in the numerator results from the fact that for each dimer represented by $P$, there are two subunits capable of binding ligand. Similarly, in the case of ligand binding to the $P \cdot L$ complex there is only one subunit capable of binding ligand, but there are two subunits in $P \cdot L_2$ from which $L$ can dissociate, thus giving

$$K'_2 = \frac{[P \cdot L_1] \times [L]}{2[P \cdot L_2]}.$$

Solving as before gives

$$[P] = \frac{K'_1 K'_2 [P_{tot}]}{K'_1 K'_2 + 2K'_2[L] + [L]^2}$$

$$[P \cdot L_1] = \frac{2K'_2[L]}{K'_1 K'_2 + 2K'_2[L] + [L]^2}$$

$$[P \cdot L_2] = \frac{[L]^2}{K'_1 K'_2 + 2K'_2[L] + [L]^2}$$

and with a few more steps, remembering that the number of subunits with bound ligand in $P \cdot L_2$ is two times $P \cdot L_2$, the amount of subunits with bound ligand is, found to be,

$$\frac{(2K'_2[L] + 2[L]^2)[P_{tot}]}{K'_1 K'_2 + 2K'_2[L] + [L]^2}.$$

Comparing these two formulations shows that $K'_1 = 2K_1$ and $K'_2 = K_2/2$.

Cooperativity in binding occurs when the dissociation constant for the second ligand molecule to bind is less than for the first, that is, when $K'_2 < K'_1$. In terms of the macro dissociation constants, cooperativity occurs when $K_2 < 4K_1$.

To check the equations, the total free energy of the binding must be independent of whether macro or micro dissociation constants are used. The total free energy change of binding is the free energy of the binding of the first ligand molecule plus the free energy of the binding of the second ligand molecule. In terms of the macro dissociation constants using the relationship discussed earlier between free energy and the equilibrium constant,

$$\Delta G = -RTln(K_1) - RTln(K_2)$$

and in terms of micro dissociation constants,

$$\Delta G = -RTln(K'_1) - RTln(K'_2)$$

but since $K'_1 = 2K_1$ and $K'_2 = K_2/2$ the second equation becomes

$$\Delta G = -RTln(2 \times K_1) - RTln\left(\frac{1}{2} \times K_2\right)$$

$$= -RTln(2) - RTln(K_1) - RTln\left(\frac{1}{2}\right) - RT\ln K_2$$

$$= -RTlnK_1 - RTlnK_2.$$

As required, the free energy change of the binding is the same when computed with micro or macro dissociation constants.

**Analysis of Binding to a Dimeric Protein**

To gain a deeper feel for cooperative binding, consider ligand binding to dimers. For this analysis, it is more convenient to use micro-dissociation constants. Binding is cooperative whenever the binding of the second ligand molecule to bind to a dimer is tighter than the binding of the first. As noted above, in terms of the dissociation constants, cooperativity exists when $K'_2 < K'_1$ .From the previous section, the fraction of saturation of subunits is:

$$\text{Fraction of saturation} = \frac{1}{2} \times \frac{2K'_2[L] + 2[L]^2}{K'_1K'_2 + 2K'_2[L] + [L]^2}$$

This equation for dimers is too complex to understand merely by inspection. One way to help understand it and cooperative binding is by plotting the fraction of saturation of binding as a function of ligand concentration for various values of the dissociation constants $K'_1$ and $K'_2$.

The calculations and plotting are conveniently done on a spreadsheet program. The symbols $K'_1$and $K'_2$ can be placed in cells A1 and A2, and the numerical values for these parameters can be put in cells B1 and B2. Values for the concentration of ligand starting from 0.00 to 2.00 and incrementing by 0.01 can be placed in column C. Column D will contain the values for fraction of saturation. Cell D1 should contain the equation "=(B$2*C1 +C1^2)/(B$1*B$2 +2*B$2*C1 +C1^2)". Cell D1 is then copied and pasted into cells C2-C200. The values in column D are then plotted as a function of the values in column C. Figure 5-1 shows three plots of the fraction of saturation as a function of ligand concentration for three combinations of $K'_1$ and $K'_2$. The curve for $K'_1 = 2, K'_2 = 0.02,$ which indicates highly cooperative binding since $K_2 \ll K_1$, initially curves upward before then curving downward. This shape of curve is called sigmoidal. When the cooperativity is reduced, but still present by shifting $K'_2$ to 0.2, the sigmoidal character can just barely be seen.

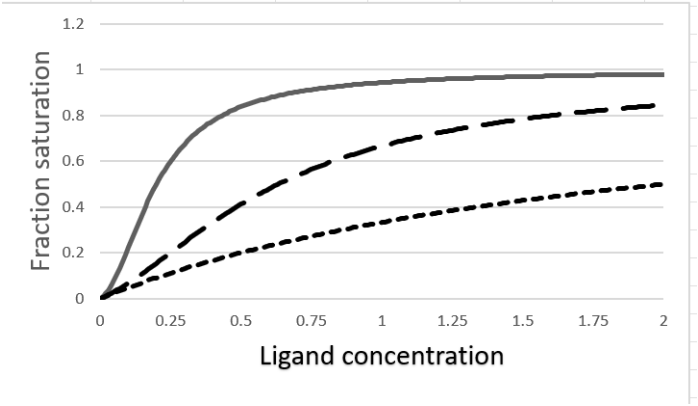



Figure 5-1. Plot of fraction of saturation of a homodimer as a function of ligand concentration. Solid, K1'=2, K2'=0.02; Long dash, K1'=2, K2'=0.2; Short dash, K1'=2, K2'=2.

A sigmoidal character of an oligomer's binding curve shows the presence of cooperative binding. It is reasonable to inquire about the converse, as seemingly innocent questions like this occasionally lead to important discoveries, or at least deeper understandings. Rarely are the beautiful and elegant experiments or demonstrations that are described in textbooks actually the way the discoveries were made. Generally, curiosity or weak hints may suggest a possibility. Then, quick, and simple experiments or numerical tests are applied to the question to see whether the proposition may possibly be true or easily shown to be false. Eventually it may appear worth substantial effort to obtain a definitive demonstration, although this often is far from elegant. Finally, sometimes long after the initial finding, a beautiful and elegant demonstration is found, and this is what often appears in textbooks. The following somewhat mimics the typical scientific approach.

Note that having a sigmoidal character means curving upward. Therefore, the property of curving upward can be used to address the question of whether or not binding to a protein can be cooperative without the ligand binding curve showing any sigmoidal character. Mathematically, curving upward means that the second derivative of binding with respect to ligand concentration is positive for some values of ligand concentration. Taking the second derivative of the binding equation for a dimer that was derived above and then determining whether it is ever positive when cooperativity is present, that is, when $K'_2 < K'_1$, looks like a laborious and unpleasant task. Therefore, instead, start with simple numerical tests utilizing the spreadsheet described above that calculates and plots the amount of binding for a range of ligand concentrations.

From the definition of derivative as $\lim_{\Delta x \to 0} (f(x + \Delta x) - f(x))/\Delta x$, it follows that in the spreadsheet, the differences between successive binding values are approximately proportional to the first derivative of the binding, and the differences between successive values of these are approximately proportional to the second derivative. Two columns that calculate these differences are easily added to the spreadsheet, as well as a plot of the values of the second added column. By varying the values of $K'_1$ and $K'_2$ it soon becomes clear that the second derivative is never positive when $K'_2 > K'_1/2$. However, the binding is cooperative by definition whenever $K'_2 < K'_1$. Therefore, it looks like the dimer can show cooperative binding but not

have a sigmoidal binding curve when $K'_1/2 < K'_2 < K'_1$. Figure 5-2 shows that if $K'_2$ is just a little less than $K'_1/2$ the second derivative is slightly positive when $L$ is close to zero and hence, the curve is very slightly sigmoidal. If, instead of numerically computing the second derivative, it is computed analytically, the same is found. The easiest way to do this is first to make use of the graphical results that showed the values of $L$ relevant to our question are at the origin. Therefore, it suffices to expand the binding equation in powers of $L$ about the origin and then calculate the second derivative of this at the origin. The result is found not to be positive, that is not sigmoidal, just for the values of the dissociation constants given above.

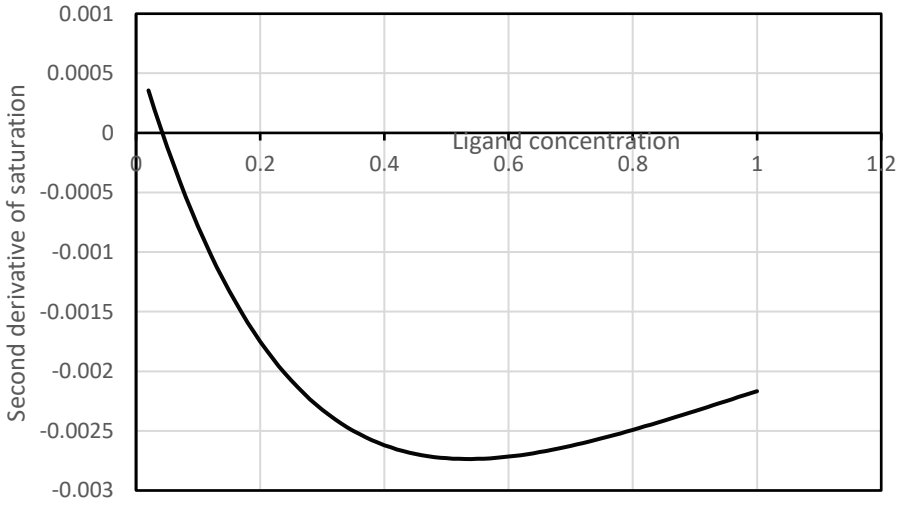



Figure 5-2. Plot of the degree of upward curvature of the amount of binding of a ligand to homodimers with the dissociation constants $K_1 = 2, K_2 = 0.95$.

**With Two Ligands, Order of Binding Can Matter, Thermodynamic Linkage**

Frequently, gene regulatory proteins sense some intracellular condition by binding one ligand, and as a result, their binding to another ligand, often DNA, is changed, ultimately leading to a change in a gene's expression. A specific and well studied example is the AraC protein from *E. coli*. The binding of arabinose to AraC increases its affinity for binding to a DNA site overlapping the *araBAD* promoter by a factor of 30, thereby inducing expression of the *ara* genes. Intuitively it seems that if the binding of arabinose changes the DNA binding affinity of AraC, then binding to DNA should change the arabinose binding affinity of AraC. This proves to be correct, as shown below.

AraC can be free of ligands, can bind arabinose, can bind DNA, or can bind both arabinose and DNA. These four different ligand-bound states of the AraC, A, B, C, and D are shown in Figure 5-3 along with the equilibrium association or dissociation constants. The product of the four

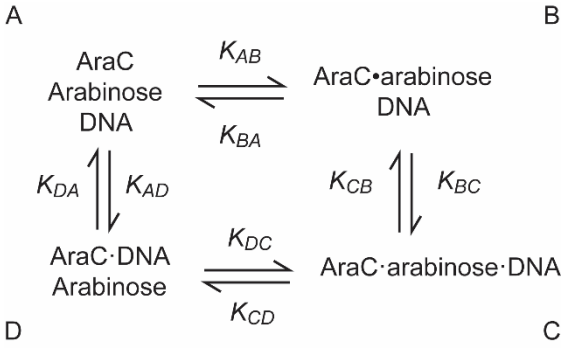



Figure 5-3 Linkage diagram showing the four states and four binding and dissociation reactions that are possible when two different ligands can bind to a protein.

equilibrium constants written for the reactions going clockwise from A to B to C to D and back to A equals one,

$$K_{AB} \times K_{BC} \times K_{CD} \times K_{DA} = 1.$$

This follows by writing them out, or more simply by noting that each combination of reactants in the four states appears once in the numerator of the four factors, and once in the denominator. Another way to see that the product must be one is to note that the total free energy change in completing a full circuit of the four states has to be zero and $K = 1$ when $\Delta G = 0$. Rearranging slightly gives

$$K_{AB} \times K_{BC} = \frac{1}{K_{CD} \times K_{DA}}.$$

Since the equilibrium constants for reactions in opposite directions are reciprocals, $K_{CD} = 1/K_{DC}$ and $K_{DA} = 1/K_{AD}$. The equation becomes

$$K_{AB} \times K_{BC} = K_{AD} \times K_{DC}.$$

This says that the product of the equilibrium constants in the path A to B to C equals the product of the equilibrium constants in taking the other path to C, that is, A to D to C.

Now, the magic. If the binding of arabinose increases the affinity of AraC for DNA by a factor of 30, $K_{BC} = 30K_{AD}$. then to satisfy the equation, $K_{DC}$ must equal $30K_{AB}$. That is, if the binding of arabinose to AraC increases its affinity for DNA by a factor of 30, then its binding to DNA increases its affinity for arabinose also by a factor of 30.

The linkage relationships like that just considered prove to be valuable experimentally. Often, testing a mechanistic hypothesis for a protein with two ligands requires knowing all four dissociation constants, but sometimes, determining the value of one of them is difficult. By virtue of the linkage relationships, its value can be determined by measuring the other three. Also, of importance experimentally is the fact that failure to consider the linkage relationships can lead to experimental errors.

### General Mechanisms for Allostery

Designing an allosteric protein *de novo* or meaningfully reengineering a naturally occurring allosteric protein requires understanding the underlying allosteric mechanism. Such mechanisms prove to be exceptionally difficult to study however because multiple similar structures and multiple small movements often are involved. Later, the book considers protein structures and their determination and it also considers how the dynamics of protein structures can be studied. However, the application of these tools to allosteric systems rarely provides sufficient information to unambiguously demonstrate precisely what is happening and why it happens as its various ligands bind.

Even using the word "structure" when applied to an allosteric protein can be an oversimplification or misleading. Instead of possessing a single structure, proteins possess multiple structures as portions of them vibrate and move in motions that involve anything from an atom to major portions of the protein. If different structures in the ensemble of structures possess different activities, it may be hard or impossible to determine exactly what is happening in the protein as ligand molecules bind and change the relative energies of different structures and possibly also the number of different structures involved. That is, the activities of proteins in general are determined by free energy differences and the number of available states. The enthalpic contributions to free energy are relatively easy to visualize in terms of bond strengths calculated from specific structures. The entropic contributions to free energy are much harder to visualize because they cannot be calculated from any single molecular structure.

Because of the extraordinary complexity possible in allosteric proteins, meaningful progress can best be made using simple models. The usefulness of such models then critically depends on two factors: first, how closely does the action of the model resemble the true situation, and second, is the model sufficiently simple that it can be understood? A bonus from a good model is that some of the parameters in the model apply to the true situation and can be experimentally measured. Early in the history of the study of allostery three models for allosteric systems were developed. More recently more complicated, but more realistic models have also been developed. The three classical models will be discussed in the next section.

### Classical Models of Allostery

The Hill model is the simplest, and also the most widely used approximate model for ligand binding to an oligomeric, allosteric protein. In it, $n$ molecules of a ligand $L$ are modeled as binding simultaneously to the protein $P$ in a reaction that can be written as $P + nL \rightleftharpoons P \cdot L_n$. Intermediate states of ligand binding are approximated as not ever present. If the binding were not cooperative, $n$ would be $1$. At the other extreme, if the binding were extremely cooperative, $n$ would be the number of subunits or the number of ligand molecules that could bind to the oligomer. Consequently, $n$ can never exceed the number of subunits. In the Hill approximation, $n$ is allowed to take on nonintegral values, with increasing cooperativity giving increased values of $n$. Systems displaying negative cooperativity are approximated with values of $n$ less than one. The binding curve given by the Hill approximation is derived just like the Michaelis-Menten binding curve. To maintain consistency in the dimensions, the dissociation constant is usually entered as $K_d^n$, giving

$$K_d^n = \frac{[P] \times [L]^n}{[P \cdot L_n]}$$

The same manipulations as used in the derivation of the Michaelis-Menten equation give

$$\frac{[P \cdot L_n]}{[P_{tot}]} = \frac{[L^n]}{K_d^n + [L]^n}$$

Values for $n$ and $K_d$ can be determined from experimental values for $[P \cdot L_n]/[P_{tot}]$ by curve fitting. Often, just the degree of cooperativity given by $n$ is sufficient. This can be determined by using binding data for values of $L$ near zero. In this region,

$$\frac{[P \cdot L_n]}{[P_{tot}]} \cong [L]^n$$

or

$$log\left(\frac{[P \cdot L_n]}{[P_{tot}]}\right) = nlog[L].$$

This shows that $n$ is the slope of the graph when the log of the response or binding is plotted as a function of the log of the substrate concentration for values of the substrate near zero.

The Monod-Changeux-Wyman model is more complicated than the Hill model. With more adjustable parameters, it should generally be able to fit experimental data more closely than the Hill model. In this simplification of reality, all the subunits of each oligomer of a protein with $n$ subunits are taken to be in either a low affinity conformation or a high affinity conformation for the binding of a molecule of the ligand. Further, when no ligand has bound, the two conformational states are in thermal equilibrium with each other. Finally, whenever a molecule of ligand binds, it locks all the subunits of that oligomer in their high affinity conformation. Whereas the Hill model approximates cooperativity by approximating the binding of multiple molecules of ligand to be simultaneous, the M-C-W approximates cooperativity by simultaneously locking of all the subunits in their high affinity state whenever a molecule of ligand binds to any subunit.

The Koshland-Némethy-Filmer model is another phenomenological model for oligomeric allosteric proteins. In the K-N-F model, the subunits of an oligomer are allowed three possible states. When no molecules of ligand have bound, all the subunits are in their ground state. The binding of a ligand molecule to a subunit shifts that subunit to a new state. This state can have higher affinity for the ligand in which case, the allosteric protein will display positive cooperativity. When such a subunit has shifted states, it decreases the energetic cost of another subunit or subunits of the oligomer for binding the ligand and shifting to the high affinity state. This also generates positive cooperativity. If the bound state has lower affinity for the ligand than the ground state, the binding of ligand has increased the energetic cost of binding a molecule of the ligand, and the protein will display anticooperative binding. This model differs from the M-W-C model in that the different states of the subunits are not considered to be in thermal equilibrium and the subunits shift to the ligand bound state in a more sequential manner than all together.

### An Allosteric Mechanism for Regulating DNA Binding Affinity

In a few simple cases, definitive experiments can be done to answer specific mechanistic questions about an allosteric system. One such case is that of a DNA binding protein whose

DNA binding affinity is regulated by the binding of a small molecule ligand. There are two basic ways that the binding of a small molecule ligand to a DNA binding protein can affect the protein's binding ability or affinity. Either access of DNA to the DNA-binding region of the protein can be controlled, or the affinity of the protein's binding site for DNA can be controlled. An additional regulation mechanism becomes available to oligomeric DNA binding proteins that utilize two or more subunits to contact multiple copies of their DNA target sequences, Figure 5-4.

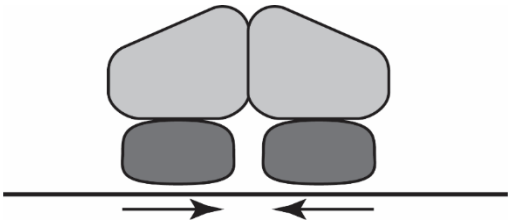


Figure 5-4 Schematic of a dimeric DNA binding protein with distinct dimerizing and DNA binding domains binding to a pair of inverted repeat half-sites.

The chelate effect, that was described earlier shows that the simultaneous utilization of multiple binding interactions can give rise to tight binding. It leads to a particularly flexible mechanism for regulating DNA binding affinity. This is simply controlling the relative position and orientation of DNA binding domains of a protein so that they either closely correspond to the relative positions of the DNA binding sites, or they do not. In the first case, the DNA binding affinity can be high, and in the second case, it will be lower. Therefore, the question for many oligomeric DNA binding proteins is whether ligand binding changes the basic structure of the DNA binding sites on the subunits—intrinsic regulation, or whether ligand binding changes the relative positioning and orientation of the DNA binding domains so that only one domain at a time can contact the DNA—extrinsic regulation.

A conceptually simple method can be used to resolve the question of whether a protein utilizes an intrinsic or extrinsic mechanism to regulate its DNA binding affinity. In an intrinsic mechanism, the intrinsic DNA binding affinity of individual DNA binding domains is regulated, and in an extrinsic mechanism, the DNA binding affinity of the individual DNA binding domains is constant, but their ability to bind together at the same time is regulated. Consider a dimeric DNA binding regulatory protein whose normal binding site on DNA consists of two nearly identical nucleotide sequences. Suppose the affinity of the protein for this DNA increases by substantially, say 50-fold, in the presence of a particular small molecule ligand. If, instead of being normally connected by four base pairs of double stranded DNA, suppose the two DNA binding sites are connected by a length of 20 or 30 bases of single-stranded DNA, Figure 5-5.

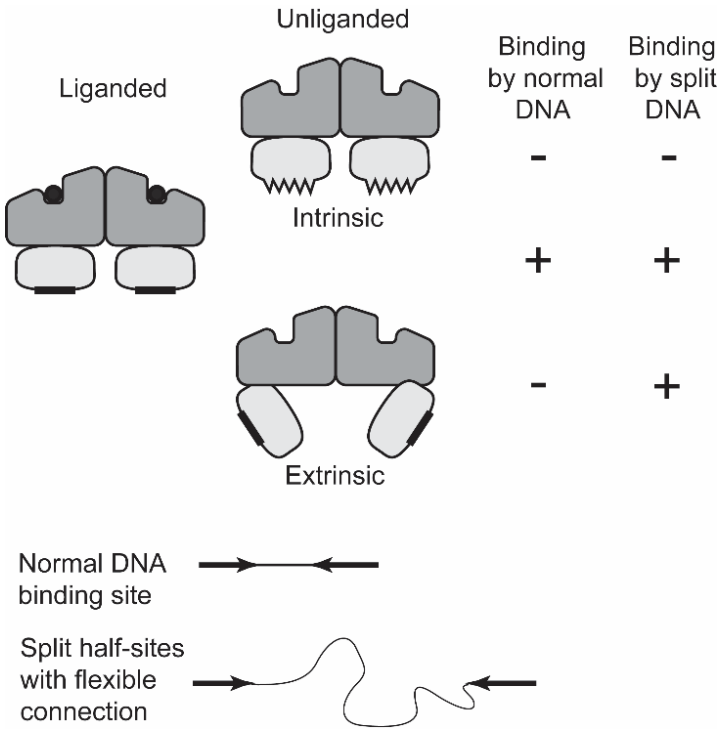



Figure 5.5 The differences between an intrinsic and extrinsic mechanism for modulating DNA binding affinity using a normal DNA binding site and split DNA with half-sites separated by flexible, single-stranded DNA.

As will be discussed later, single stranded DNA is highly flexible. This special DNA should accommodate most ligand-induced changes in the positions of the DNA binding domains. Therefore, if the protein utilizes an intrinsic mechanism for regulating its DNA binding affinity, its affinity for this split site DNA should show a largely normal change in affinity in response to ligand binding. On the other hand, if the protein utilizes an extrinsic mechanism, then its affinity for the split site DNA should be independent of ligand binding.

**Problems**

1. How would you determine experimentally using gel exclusion chromatography or ultracentrifugation and SDS gel electrophoresis the fraction of intracellular proteins in any cell type that are found in oligomers?

2. For noncooperative ligand binding to a dimer show that the binding equations reduce to the Michaelis-Menten binding equation.

3. Instead of using the split-site DNA to distinguish between intrinsic and extrinsic mechanisms for regulating DNA binding affinity, in principle, DNA containing only a single binding site could be used. Why might it be difficult experimentally to use this idea?

4. The binding of a ligand molecule to a protein is frequently found to stabilize the protein against denaturation. Making use of this fact, how would you design a single polypeptide allosteric protein?

5. Why is the addition of subunits to an oligomeric protein cooperative?

6. In terms of macro dissociation constants for a dimeric protein, what is the relationship between the dissociation constants that will yield positive cooperativity in ligand binding, but not generate a sigmoidal binding curve?

# 6. Measuring Binding

### Much of Biology is Binding

Complicated and intricate series of binding reactions accompany many of the important cellular processes. For example, prokaryotic DNA replication involves more than ten different proteins that comprise a DNA polymerase and the several other complexes that interact with the DNA or with the polymerase. This does not include the multitude of additional proteins that form complexes that control the initiation of DNA replication, participate in error correction, recombination, or termination of replication. Processes in eukaryotic cells involve even larger complexes. For example, the initiation of transcription at a eukaryotic promoter involves over 90 proteins that interact in several multiprotein complexes. Small molecules are also involved in important binding reactions. Enzymes bind their substrates, products, cofactors, and allosteric regulators. Finally, many different proteins individually bind to DNA or RNA.

Overall, a wide diversity is seen in the types of molecules that are involved in some sort of binding reaction. Similarly, a wide diversity is seen in the conditions in which binding is to be detected and studied, and finally, there are a wide variety of physical consequences of binding that enable binding to be detected and quantitated. Therefore, a huge number of methods have been employed to study binding reactions of interest in biology. Only a few will be discussed and described in this chapter.

### Precipitation

Sometimes the binding between two species of molecule, $A$ and $B$ is so tight that their dissociation rate will be sufficiently slow that binding can be measured simply by physically separating the bound species $AB$ from $A$ or $B$ and measuring its amount. Often, antibodies against $A$ or $B$ are used and the assay is called a pull-down assay. Also, filtration or ultracentrifugation can sometimes be used for the required separation.

### Equilibrium Dialysis

If one of the binding species readily penetrates a dialysis membrane, and the other does not, equilibrium dialysis can sometimes be used to detect and quantitate binding, Figure 6-1.

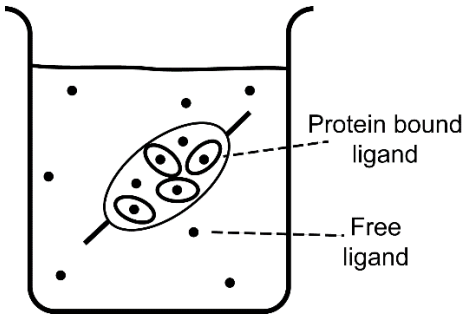



Figure 6-1. Equilibrium dialysis after equilibrium has been established.

If the total concentration of the nondialyzable protein that binds the small molecule ligand is $P_t$ and the concentration of the dialyzable ligand is $L$, then, when binding equilibrium has been reached in an equilibrium dialysis measurement, the concentration of free ligand inside the sack will equal its concentration outside the sack. Inside the sack, in addition to the free ligand that is not bound to the protein, ligand bound to the protein will also be present. Thus, using the Michaelis-Menten equation derived and discussed earlier, the concentration of ligand inside the sack will be.

$$[L_{in\ sack}] = [L_{free}] + \frac{[L_{free}] \times [P_t]}{K_d + [L_{free}]}$$

Typically, the ratio of the concentrations of ligand in a volume of the outside buffer and from inside the sack are measured and their ratio is computed, giving

$$1 + \frac{[P_t]}{K_d + [L_{free}]}$$

for the signal from the protein. From this, by varying $[L]$, the $K_d$ and $P_t$ can be determined.

**Altered Fluorescence**

In measuring the binding of a small molecule ligand to a protein, often one can measure binding-induced changes in the ligand or in the protein. One of the more common changes that occur in a protein involves the amino acid tryptophan. Being an aromatic amino acid, tryptophan absorbs light in the UV region. On the order of $10^{-9}$ seconds after absorbing an exciting photon, most of the energy of the exciting photon is emitted as a somewhat longer wavelength photon, Figure 6-2.

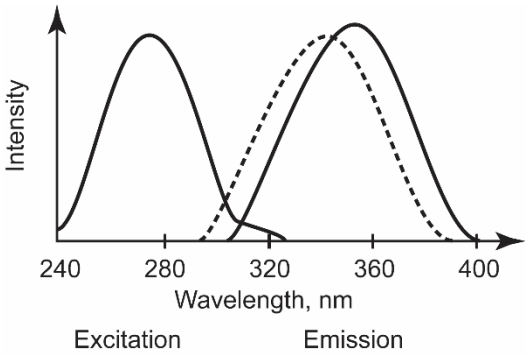



Figure 6-2. Representations of the excitation and emission spectra of tryptophan, solid lines, and an emission spectrum shifted by ligand binding, dashed line.

This is fluorescence, and it can be used to measure ligand binding if a tryptophan residue in the protein lies sufficiently close to the bound ligand. Such fluorescence is called intrinsic fluorescence because it is intrinsic to the protein. This is in contrast to fluorescence that depends on the post synthesis addition of a fluorescent molecule to a protein.

The fluorescence of a tryptophan residue is sensitive to the polarity of its environment. Water molecules, which are highly polar, in a ligand binding pocket on a protein will be displaced by the binding of a ligand molecule. This can change the magnitude of the protein's tryptophan fluorescence or the wavelength distribution of the tryptophan fluorescence, or both.

Typically, in measuring ligand binding, the protein's fluorescence is excited by illumination with UV light at a wavelength near 280 nm and the position of the emission peak is determined by taking a few measurements near the peak. This is sufficient when the emission spectrum is sufficiently well defined and shifts a substantial amount in response to ligand binding. This shift in the peak can be directly used as an indication of ligand binding. The relative amount of the shift can be taken to be proportional to the fraction of the protein with bound ligand.

In cases where the peak shift is not well resolved; relative amounts of ligand binding can still be determined by using most of the entire emission spectrum rather than just the few values of the spectrum near the maximum. Information from all, or most of the emission spectrum can be simply combined by determining the average emission wavelength over a broad range that includes the emission peak. Where $\lambda_n$ is the wavelength of the n-th data point and $E(\lambda_n)$ is the emission intensity at that wavelength, then

$$Average\ emission\ wavelength = \frac{\sum_n(\lambda_n \times E(\lambda_n))}{\sum_n E(\lambda_n)}.$$

This gives a much more precisely determined value that is related to the amount of protein with bound ligand.

Using many points in the emission spectrum to determine a property of the spectrum more precisely is analogous to methods used in single molecule experiments. There, multiple measurements of position are often used to precisely determine locations of molecules or of very small latex beads to which molecules of DNA are attached. Using such methods, locations of latex beads can be determined with uncertainties as small as 10 angstroms. Such averaging methods became possible with computers and automatic methods for determining the locations and intensities of multiple points in an image. Before the use of such computer methods, peak positions of objects were determined visually, and resolution of microscopes was defined as the minimum separation between two point-like objects that yielded two image peaks with a detectable minimum between them. This so-called resolution is much larger than the precision with which a peak can be defined by utilizing many points from the image.

**Microcalorimetry**

Enthalpy is a measure of the internal heat of a system. In biological systems, changes in enthalpy result almost entirely from the formation or breaking of bonds. These processes either release or cost energy. If there is a net energy change upon the binding of a small molecule or a protein ligand to another protein, the change in binding energy can be sensitively measured in a microcalorimeter as heat that is either evolved or absorbed. This property can be used to screen for small molecule ligands that bind to a protein as in a search for drugs. It can also be used to determine the thermodynamic properties of a binding reaction.

A microcalorimeter consists of two cells, one, a reference cell, contains several milliliters buffer solution, and the other, the sample cell, contains several milliliters of the same buffer containing one of the two species of molecule whose binding is to be measured, Figure 6-3.

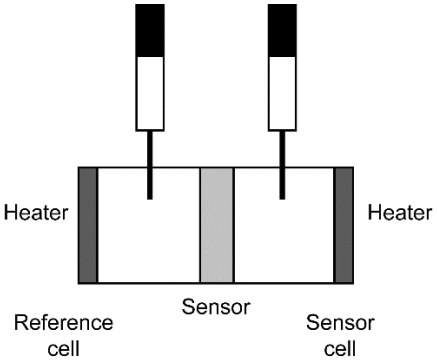



Figure 6-3. A microcalorimeter showing its micro syringes for injection of sample into the reference and sensor cells, the temperature sensor between the cells and the two heaters.

Small aliquots of buffer containing the other molecular species, the ligand, are injected into both the reference cell and the sample cell. Binding of ligand to the molecules in the sample cell either raises or lowers the temperature of the buffer in the sample cell compared to the reference cell. Multiple thermocouples in series sensitively detect very small temperature differences between the cells. Their signal controls the heating of either the reference or sample cell until the temperatures in the two cells are again equal. After temperature equilibrium has been reached, another small sample of ligand is added to the sample cell. This process is repeated until the concentration of ligand in the sample cell is high enough to saturated the binding and no further heat is either evolved or absorbed. The data from such an experiment can be used to determine the equilibrium dissociation constant of the binding reaction. The dissociation constant of the binding reaction always equals

$$K_d = \frac{\left[Lig_{free}\right] \times \left[Prot_{free}\right]}{[Lig \cdot Prot]}.$$

Therefore, a binding experiment is designed so that the three unknowns in the equation may easily be determined. This is possible by adjusting concentrations initial concentrations such that at the point halfway to saturation in the evolution or absorption of heat, the concentration of free ligand in solution in the sample cell is essentially the same as it would have been if no binding were occurring. This requires that at the halfway point, the amount of ligand that has been bound is much smaller than the total amount of ligand that has been added. If the total amount of protein that was initially added to the sample cell is known, then at the halfway point,

$$\left[Prot_{free}\right] = [Lig \cdot Prot]$$

so that

$$K_d = \left[Lig_{free}\right]$$

.

As we saw in an earlier chapter,

$$\Delta G = -RTln(K_d)$$

so that once $K_d$ is known, $\Delta G$ can be calculated. Since the heat absorbed or evolved equals the enthalpy, H, it can be calculated, and since

$$\Delta G = \frac{\Delta \mathrm{H} - \mathrm{T}\Delta \mathrm{S}}{RT},$$

$\Delta S$ it too can be calculated. Thus, microcalorimetric measurements allow determination of the equilibrium dissociation constant, $K_d$, of the reaction of ligand binding to the protein and the three thermodynamic parameters $\Delta G, \Delta H,$ and $\Delta S$ of the binding reaction.

To satisfy the requirement that at the halfway point in the titration, the concentration of the free ligand in the sample cell will approximately be the total amount of ligand added, it is sufficient that $[P_{total}] \ll K_d$.

**Principles of SPR, Surface Plasmon Resonance**

Measurements of binding may be made using the phenomenon of surface plasmon resonance, spr. Although this discussion is oriented to measurements of protein-protein interactions, this versatile and sensitive technique can also be used to measure binding between proteins and small molecule ligands or proteins and short pieces of DNA. First, spr will be described, and then the application of spr to the measurement of the equilibrium dissociation constant and association and dissociation rate constants between two proteins will be described.

The spr technology combines the phenomena of total internal reflection, and plasmon resonance. A beam of light is bent when it crosses a boundary between two media of different indices of refraction, Figure 6-4.

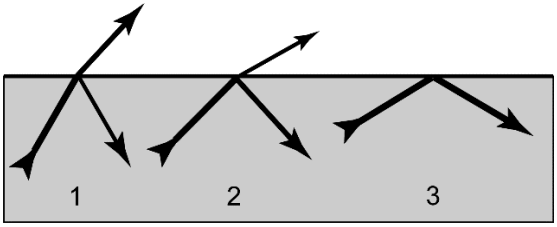



Figure 6-4. Bending of a beam of light when it passes between media with different indices of refraction. In this case, light is passing from a medium of higher index of refraction to one of lower index of refraction. Beam three is incident at the critical angle, and is totally reflected by the interface.

As shown in the figure, a light beam passing from glass to air where the index of refraction changes from about 1.3 to near 1.0, is bent towards the plane of the interface. A portion of the incident light beam is also reflected from the interface. As the angle of incidence is increased still further, the refracted beam comes closer and closer to being parallel to the interface. When the refracted beam would be parallel to the surface, there is no refracted beam. At this point

and beyond it, all the energy of the incident beam is reflected, and the light is said to be totally internally reflected.

Total internal reflection can be observed in a swimming pool. While underwater, it is possible to see objects outside the pool if one is looking nearly directly upwards. Outside this circle however, one sees only a reflection from the surface of whatever is in the pool. An artistic application of total internal reflection is found in gemology. Many of the cuts of diamonds are designed to utilize total internal reflection to make the diamond brilliant and glisten by reflecting all the incident light.

When a beam of light is reflected by total internal reflection, no energy continuously crosses the boundary that produces the reflection. However, standing waves not carrying energy anywhere are set up at the boundary. These have a strength falling exponentially with distance and penetrate only about one wavelength of the light, around 6000 Å, into the medium of lower refractive index. These are called evanescent waves and prove to be most useful for illuminating very small volumes, as is often needed in microscopy. Although the evanescent waves do not normally carry energy away from the interface, a small object or molecule in the evanescent wave area sufficiently disturbs evanescent wave that the object can absorb and then reradiate energy. This permits very sensitive detection in a microscope and tracking of individual protein or nucleic acid molecules tagged with fluorophores.

Spr utilizes total internal reflection of a laser beam at a glass-water interface. The glass surface possesses a very thin layer of gold. When the incident light of a suitable wavelength is just at the angle where the refracted light is parallel to the interface surface, the electric field component of the light induces movement of the electrons on the surface of the gold. These behave as though they are quantized and are called plasmons. This electron movement neutralizes the reflection of the incident light and there is no reflected beam. At angles just a little more or less than the critical angle, there is appreciable reflection by the interface, but at the critical angle, there is none.

In this phenomenon the evanescent wave plays a role. The index of refraction in the evanescent wave zone strongly affects the critical angle. Consequently, very small changes in the index of refraction in the very thin evanescent wave zone significantly alter the angle at which the excitation of surface plasmons block reflection from the interface. Changes in the angle of the null of the reflection can then be sensitively measured using a split sensor detector centered at the critical angle, Figure 6-5.

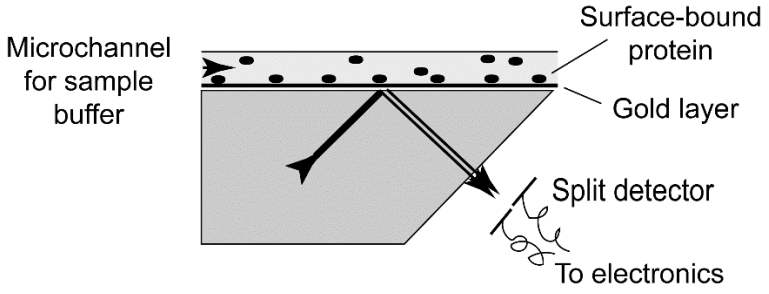



Figure 6-5. Schematic of an spr instrument showing the incident beam that is totally reflected except at the precise angle at which surface plasmon resonance occurs. On the

other side of the thin layer of gold, buffer containing sample is flowed across the glass with the thin layer of gold. The difference in the two outputs of a split sensor at the nominal position of the null in the internal reflection provides a signal proportional to the shift in the position of the null in the internal reflection. This signal is proportional to the amount of material that has been immobilized in the evanescent wave zone.

Finally, it is possible to complete the description of how spr is used to detect and quantitate binding between two proteins. One of the proteins is immobilized on the gold surface. The other is contained in a solution that is passed over the gold surface in a microchannel. As molecules of the second protein bind to those of the first protein, the index of refraction near the surface changes and is detected by the angle-sensitive detector.

**Equilibrium Constant Measurement with SPR**

Within suitable ranges of the rate constants, spr-based instruments can be used to measure equilibrium dissociation constants. For an interaction between molecules A and molecules B, let molecules B be attached to the gold surface or to an intermediary molecule attached to the gold surface in an spr instrument. At the start of the measurements, only buffer is flowed through the microchannel above the gold surface. Then buffer is shifted to one containing a known concentration of molecule A. As molecules of A bind to B and change the index of refraction close to the gold, the angle of null reflection from the surface changes in proportion to the amount of A bound. After further increase in the response ceases and equilibrium in the reaction of $A + B \rightleftharpoons AB$ has been reached, the flow is shifted back to buffer only. As $A$ now dissociates from $B$, the spr response falls back to the original buffer only level, Figure 6-6.

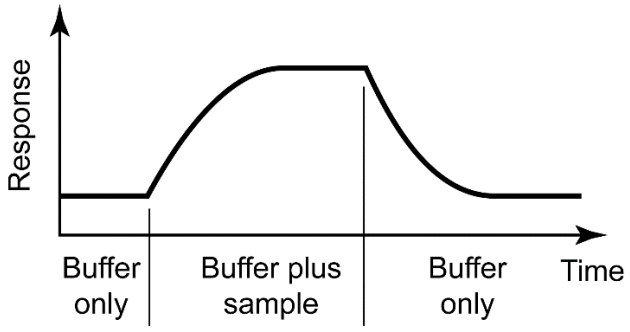



Figure 6-6. The association and dissociation phases of a binding experiment in an spr instrument.

As higher and higher concentrations of $A$ are used in the experiment, values are eventually reached where all $B$ becomes fully bound with $A$ and the spr response is saturated, Figure 6-7.

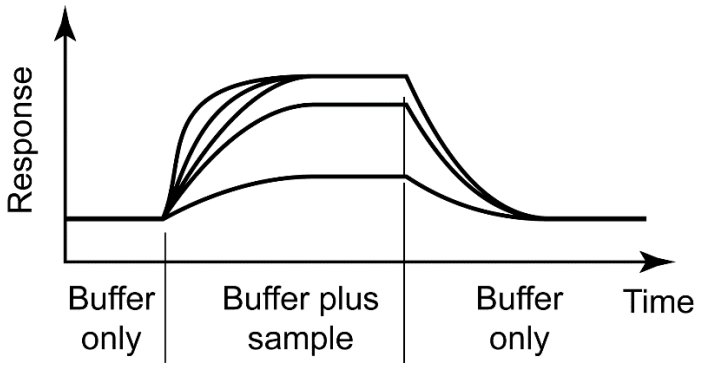



Figure 6-7 Association and dissociation at concentrations of A approaching and then exceeding saturation in binding to B

Next, concentrations of $A$ are found which lead to an equilibrium response that is half of the saturated response. By the simple Michaelis-Menten binding equation derived and discussed earlier and written appropriately for this situation,

$$\frac{Response(A)}{Saturated\ Response} = \frac{[A]}{K_d + [A]}$$

so that when $[A] = K_d$,

$$\frac{Response(A)}{Saturated\ Response} = \frac{K_d}{K_d + K_d}$$

$$= \frac{1}{2}$$

thus, providing the value of $K_d$.

**Measuring Association and Dissociation Rates with SPR**

In a binding reaction of $A$ binding to $B$, the net rate of change in the concentration of $AB$ is the rate of association of $A$ with $B$ minus rate of dissociation of already formed $AB$. These rates are governed by the association and dissociation rate constants, $k_1$ and $k_{-1}$.

$$A + B \rightleftharpoons AB$$

$$\frac{d[AB]}{dt} = k_1[A] \times [B] - k_{-1} \times [AB].$$

Earlier, this relationship was used in relating the forward and reverse rate constants to the equilibrium dissociation constant by waiting until equilibrium in such a reaction has been reached at which point $\frac{d[AB]}{dt} = 0$. In an spr measurement, $A$ is in the flow buffer, and if the experiment suitably designed, its concentration remains nearly constant. $B$ however, is immobilized on the sensor chip, and as $A$ binds to it, the concentration of unbound $B$ is

diminished, and equals total $B$ minus liganded, that is, $[B_0] - [AB]$. Substituting this in the net rate equation and rearranging in the standard form for solving a differential equation gives

$$\frac{d[AB]}{dt} + (k_1 \times [A] + k_{-1}) \times [AB] = k_1 \times [A] \times [B_0]$$

This differential equation can be solved by going to a web site for solving differential equations, or by a few manipulations to yield

$$[AB(t)] = [B_0] \times \frac{k_1 \times [A]}{k_1 \times [A] + k_{-1}} \times \left(1 - e^{-k_1 \times [A] t - k_{-1} t}\right)$$

This is not as opaque as it first looks. Its predictions for two extreme conditions can easily be checked. At zero time, when the buffer containing species A just begins to flow, the concentration of $AB$ should be zero, and this is consistent with the solution. Also, if the concentration of $A$ is very high, then for large $t$, the concentration of $AB$ should be $B_0$, as is also shown by the solution.

Often, what is desired from an spr experiment are the rate constants $k_1$ and $k_{-1}$. During the dissociation phase of a binding experiment, when $[A] = 0$, the kinetics are determined only by the dissociation constant $k_{-1}$. Its value can readily be extracted from the data. Then, using the numerical values of $k_{-1}$ and of $A$ and the data from the association portion of the curve, $k_1$ can be determined numerically, also using the equation derived above. Usually, spr instruments come with software that numerically and rather effortlessly finds values for $k_1$ and $k_{-1}$ that best fit the data. Once these values are known, their ratio provides the equilibrium association or dissociation constant.

**Electrophoretic Migration Retardation**

The electrophoretic migration retardation assay, colloquially known as the gel retardation assay, is widely used for the sensitive detection and quantitation of protein-DNA interactions. The principle of the assay is that a protein bound to a DNA fragment of length less than about 2,000 base pairs appreciably retards the migration through acrylamide or agarose gels of the protein-DNA complex compared to the migration of free DNA. This allows their separation and quantitation.

The DNA used in the assay can be sensitively detected following electrophoresis by staining the gel with a DNA-specific dye like ethidium bromide or by using DNA that has been labeled with a radioactive atom or with a fluorophore. Because of the separation achieved in gel electrophoresis and the sensitivity of DNA detection afforded by labeling the DNA, binding can be detected when as little as 0.01 of the DNA has been bound by the protein. This allows the detection and study of DNA binding when the sample containing the binding protein is far from pure. Sometimes, binding by a specific protein to a specific sequence in a DNA fragment can be studied using crude cell extracts.

A key feature of the electrophoretic migration retardation assay is the fact that the dissociation rate of a protein from DNA during the electrophoresis step can be adjusted to be much slower than its dissociation rate under normal physiological conditions. Therefore, proteins that, under physiological conditions, would completely dissociate from DNA during a 30–180-minute electrophoresis step can easily be studied.

Two factors make it possible to extend the lifetime of protein-DNA complexes. The more important results from the negatively charged phosphate groups along the DNA phosphodiester backbone. Their charges are neutralized by cations like $Na^{+,}$ $K^{+}$, and $Mg^{++}$ that, because of the relatively high charge density of the DNA, are localized near the phosphates. Binding of a protein to the DNA displaces these counterions from the DNA binding site. Therefore, the displacement or replacement of these ions should be considered in the binding equation between a protein, $P$, and DNA, $D$, in which effectively, $n$ counterions are displaced. Because the counterions are not perfectly localized, $n$ is about 0.6 of the number of phosphates that the protein comes close to upon binding. Considering just those counterions that are displaced upon protein binding, the binding equation and equilibrium dissociation constant are:

$$[P] + [D \cdot nNa^{+}] \rightleftharpoons [P \cdot D] + nNa^{+}$$

$$K_{real} = \frac{[P] \times [D \cdot nNa^{+}]}{[P \cdot D] \times [Na^{+}]^{n}}$$

Overall, it proves convenient to omit the explicit inclusion of the counterions of the DNA by first writing

$$K_{real} = \frac{[P] \times [D]}{[P \cdot D] \times [Na^{+}]^{n}}$$

and recognizing that the factors

$$\frac{[P] \times [D]}{[P \cdot D]}$$

look like an apparent dissociation constant, $K_d$, giving

$$K_{real} = \frac{K_d}{[Na^{+}]^{n}}$$

and upon replacing the constant $K_{real}$ with C to emphasize that it is a constant,

$$K_d = C \times [Na^{+}]^{n}.$$

Thus, the apparent dissociation constant between the protein and DNA is a function of the salt concentration in the buffer used in the experiment. The lower the salt concentration, the more tightly the protein binds. When just the apparent equilibrium constant is being used, it is essential that at the same time, the concentration of the counterion also be given. Normally this counterion is either sodium as has been used above, or potassium. A dimeric DNA binding protein may displace as many as ten counterions as it binds so that $n \approx 6$ and a two-fold change in the salt concentration in the binding buffer can change the apparent dissociation constant by more than a factor of fifty.

At last, we can see how DNA binding proteins can be forced to remain tightly bound to DNA during an electrophoretic separation of free and protein-bound DNA. Use very low salt concentrations in the separation gel and in the electrophoresis buffer. Typically, a small protein-DNA sample in a buffer with near physiological levels of salt is loaded onto a separation gel. Rapid diffusion of the ions reduces the salt concentration near the DNA in just a few seconds, and the protein is then effectively welded onto the DNA.

The second factor contributing to the stability of protein-DNA complexes during electrophoresis is confinement by the gel of a protein in the vicinity of its original binding site on the DNA. A molecule of the protein that has dissociated from the DNA is appreciably restricted in its diffusion away from its DNA binding site, and it is highly probable that it will rebind to its binding site rather than diffuse away altogether.

**Problems**

1. In a microcalorimeter experiment, what does the titration curve look like if instead of $[P_{total}] \ll K_d$, $[P_{total}] \gg K_d$?
2. What are the reactions giving rise to bubble formation at the electrodes of an electrophoresis apparatus? What experimental problem is generated by these electrode reactions?
3. Explain in words why it is possible in an spr experiment to obtain the association and dissociation rates without ever explicitly knowing the concentration of the immobilized species.
4. Fluorescence anisotropy is another valuable technique for measuring binding between biological molecules. Explain the basic principles by which this method works.

# 7. Structure and Function of DNA and RNA

**What DNA Must Do**

DNA is the cell's information library. Cells require that information be stored, replicated, and on demand, read out. Evolution has developed a linear coding scheme so that the information is stored in the linear sequence of four bases along the two sugar phosphate backbone strands of a DNA molecule. Figure 7-A shows the structures of the four bases found in DNA and the uracil base that substitutes for thymine in RNA. A base joined to the five-carbon sugar 2'-deoxy ribose or to ribose is called a nucleoside, and when the nucleoside is joined to one or more phosphate groups, the molecule is called a nucleotide.

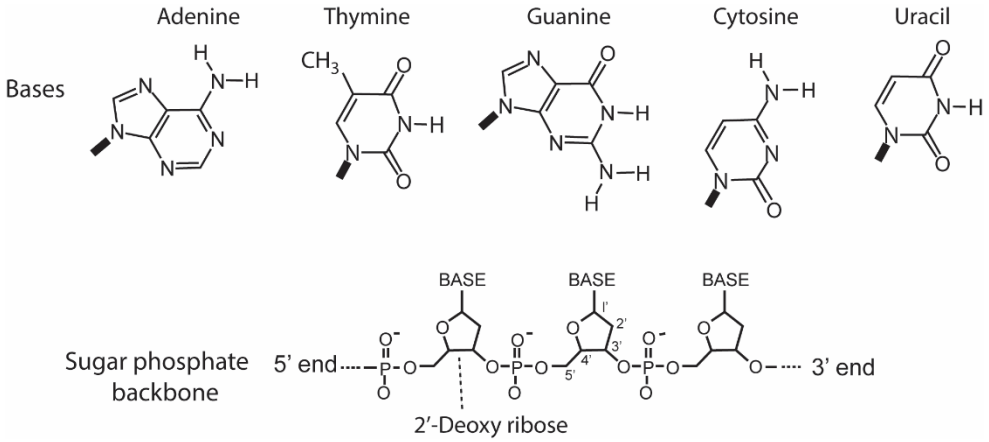



Figure 7-A The components of DNA. The bases adenine, thymine, guanine, and cytosine are found in DNA. In RNA, uracil replaces cytosine and the sugar is ribose rather than 2'-deoxy ribose.

A monomer of a polymeric strand consists of a 2'-deoxy ribose whose 3' and 5' carbon atoms are bonded to oxygen atoms of flanking phosphate groups, the 1' carbon is bonded to a nitrogen atom of a purine or pyrimidine base, and the 4' carbon bonds with the oxygen of the deoxyribose to form a five-membered furanose ring.

Because most of biology is three dimensional, processes separate from the direct handling of information are utilized in the conversion of information stored in one dimension to three dimensional structures with various biochemical activities. The major topic of the next chapter, protein folding, addresses this conversion for proteins, which fulfill most of the important activities required by a living cell.

To carry out its functions four different activities are associated with DNA. One, as already mentioned, is carrying information. Two, the information must be duplicated as cells grow and divide or as blocks of information are copied into RNA. Three, blocks of information needed by a cell must be readily identifiable by gene regulatory proteins, and four, the DNA must be managed or handled, packed, or unpacked, and recombined largely without regard to the information in the region of DNA being handled.

The structure of RNA is very similar to that of DNA. Its backbone contains ribose rather than the 2'-deoxy ribose of DNA and instead of the base thymine, it contains a similar base, uracil. RNA plays a structural and functional role in ribosomes, and a functional role in transfer RNA, tRNA. Cells also utilize the information storing capability of RNA. One of its important roles is as a short-lived carrier of information from DNA to ribosomes where its base sequence is translated into protein

**The Structure of DNA**

DNA is usually found in a two-stranded double helical structure where the atoms in each of the oppositely oriented strands are covalently connected. The two strands of DNA are held together by inter-strand hydrogen bonds between the complementary bases that form Watson-Crick base pairs, and Van der Waals interactions between successive base pairs, Figure 7-1.

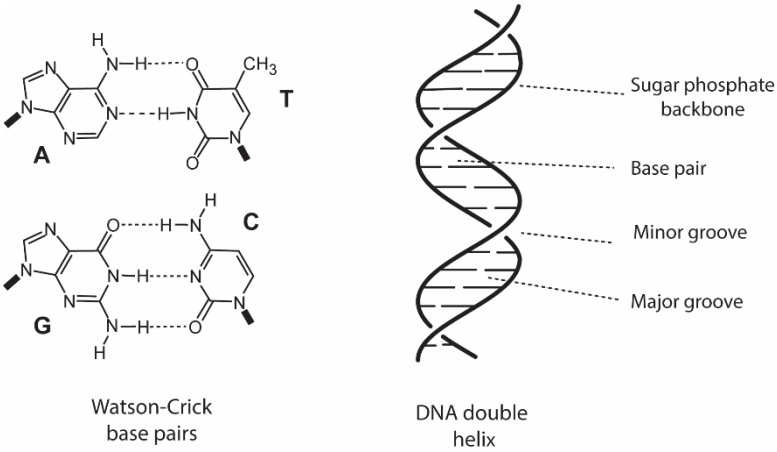



Figure 7-1. The structure of DNA showing the structure of the Watson-Crick base pairs, how the base pairs fit between the sugar phosphate backbones, and the minor and major grooves.

As can be seen in Figure 7-1, from the positions and orientations of the bonds between the bases and the sugar of the complementary A-T and G-C base pairs, both can fit nearly identically into the double helical framework of its two sugar phosphate backbones. Thus, mispaired bases that may result from replication errors can be detected by the resulting distortions in the phosphodiester backbone. Such errors can be corrected by DNA repair enzymes. The complementarity of the two strands of DNA duplicates the genetic information, which allows a damaged strand to be corrected by reference to its strand of complementary sequence.

While it is often useful to think of DNA as a rod or cylinder, a higher "resolution" image of DNA reveals grooves in the DNA cylinder that wrap around the cylinder like the threads of a screw. Figure 7-2 illustrates the origin of the grooves in DNA. It shows a cross section of a DNA helix

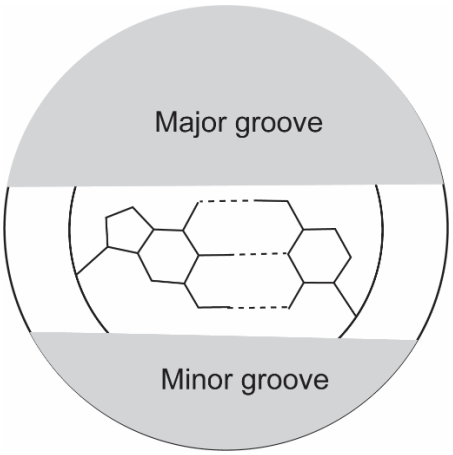



Figure7-2 Cross section of a DNA double helix in white also showing the overall cylindrical shape of the DNA.

including a base pair and the associated cross section of the sugar phosphate backbone of deoxyribose and phosphate. The grooves arise because a base pair and its backbone atoms are not circular discs. If the DNA were a smooth cylinder without grooves, the cross section would be a circular disc. In DNA, the sections missing from a disc form the grooves. With 10.5 base pairs per turn, the next base along the DNA is rotated about 34 degrees from the previous base. The missing sections similarly are rotated about 34 degrees. These indentations wrap around the cylinder, and constitute the grooves.

The duplication of DNA or the copying of DNA segments into RNA requires that the double-stranded DNA be locally disrupted so that replication or transcription can utilize Watson-Crick base pairing to determine the sequence of the DNA or RNA being copied. Melting and base pairing cannot be used however, for sequence identification by the proteins that must identify specific sequences in the DNA. Such sequence identification is required for the regulation of DNA replication and of transcription of specific genes. Proteins can recognize specific DNA sequences in double-stranded DNA by reading the patterns of hydrogen bond donors and acceptors in the major or minor grooves of DNA.

**Melting and Annealing DNA**

Raising the temperature of DNA in solution can lead to loss of DNA's double helical structure and separation of the two strands in a process called melting. The hydrogen bonds of base pairs and the stacking interactions between successive base pairs are broken by thermal energy at higher temperatures. Because the three hydrogen bonds of a G-C base pair are stronger than the two hydrogen bonds of an A-T base pair, the melting temperature of DNA, $T_m$, depends on the base composition of the DNA. The surrounding salt concentration also affects DNA's melting temperature because the higher the salt concentration, the more that the charges on the mutually repulsive negatively charged phosphate groups are shielded from one another. Higher salt concentrations result in greater stability of the double stranded structure and a higher melting temperature.

Amazingly, holding the temperature of a melted solution of DNA at a temperature a little lower than the temperature at which half of the helical structure melted, allows the separated strands to associate in register with their Watson-Crick partners and reanneal to reform the original double stranded DNA structure.

The ability of DNA or RNA strands of complementary sequence to anneal and reform the original double stranded structure is the basis of some of the more important techniques of molecular biology. These include the *in vitro* replication of DNA by using DNA oligonucleotides of the appropriate sequence to prime the replication from a specific point, the polymerase chain reaction to detect and/or amplify specific DNA sequences, and the detection of specific sequences by the formation of double stranded structures that are resistant to single-strand specific DNases or RNases. This was used in Southern transfers, a now classical technique for the detection and quantitation of specific RNA's. Regions of complementary sequence within a single RNA or DNA molecule can also anneal to each other to form double-stranded regions of intra-molecular structures, most notably, the cloverleaf structures of tRNA.

Since DNA is composed of four bases, on average, a sequence of $n$ bases occurs $N/4^n$ times in a random sequence of N bases. This means that a sequence of only about 16 bases will occur on average once in a random sequence the size of the human genome of $3.2 \times 10^9$ base pairs. The genome of any organism is far from random and it is quite possible that a few short sequences will have been selected against and be absent from a genome. Nonetheless, in general for sequences less than 30 bases, it is a good assumption that genomes are random. Therefore, as given by the Poisson distribution discussed earlier, if the average number of occurrences in a genome of a particular size is one, the probability of its not occurring at all is 0.368, of occurring once is 0.368, of occurring twice is 0.1839, and three times is 0.0613 etc.

The annealing of complementary sequences to each other can be highly specific if it is performed under conditions where the energetic cost of a single mispaired base prevents annealing. Often this is possible because the pairing of complementary bases and the formation of two or three hydrogen bonds contributes around 4 kcal/mol of binding energy. Thus, under the right conditions, DNA oligonucleotides of around 20 nucleotides can be expected to bind specifically to a single complementary sequence in the human genome despite the presence of a huge number of regions that would form complexes containing several mispaired bases.

**The Flexibility of DNA**

The path of a long homogeneous polymer will resemble a random walk. If one end of many such polymers is placed at the origin, the average positions of the other end will, also be at the origin. Instead of considering all such random walks beginning at the origin, consider just those that start off in the *x* direction. In this case, the average of the positions of the other end will be displaced some distance in the *x* direction. This displacement is defined as the persistence length of the polymer.

A single nucleotide of DNA contains five single bonds along the phosphodiester backbone about which rotation is possible. A single strand of DNA therefore, is highly flexible. As discussed in the next chapter, this contrasts with proteins, in which there are only two bonds per amino acid about which rotation is possible. Additionally, in proteins, significant steric hindrances limit the ranges of these two rotations. Flexibility is completely different in the case of double-stranded

DNA. The base stacking interactions present in double stranded DNA as well as the hydrogen bonding between bases generates a rigid and very stiff cross-braced structure that contains relatively small base pair to base pair structural differences. Whereas the persistence length of single-stranded DNA is about 15 bases, the persistence length of double stranded DNA is about 150 base pairs.

**Protein-DNA Interactions**

Multiple important biological processes involve proteins binding to DNA. Amongst these are processes in which proteins must recognize specific sequences in the genome. A large class of such proteins are transcription factors that are involved in regulating the expression of different genes. As mentioned above, in order that a prokaryotic gene regulatory protein bind specifically to a unique site on the chromosome that is involved in regulation of its target gene, the protein must be able to read DNA sequences of 12-20 base pairs. Because of the greater complexity of eukaryotic cells, their gene regulation generally involves multiple DNA binding proteins that also recognize and bind to DNA sequences of about this size. The next chapter discusses some of the variety of protein structures that are found for reading sequence by contacting DNA in the major groove.

In many of the cases for which data is available, the relevant biological activity of a DNA binding protein is governed by the extent of DNA binding as described by equilibrium dissociation constants. This means that wide fluctuations in the protein-DNA dissociation constant of a regulatory protein would be problematic. Therefore, cells will have evolved mechanisms for minimizing variation of a protein's DNA binding affinity. One of the most important potential causes of changes in DNA binding affinity is salt concentration. For this reason, cells maintain constant intracellular salt concentrations. In contrast, experimentalists studying protein-DNA interactions are not confined to using just one salt concentration in their buffers. Thus, one should be aware of and correct for the sometimes enormous salt concentration effects on the binding of proteins to DNA. The previous chapter discussed and analyzed salt concentration effects on DNA binding proteins.

Below, this important effect is further discussed.

When two atoms or molecules, $A$ and $B$ associate to form a complex, $A \cdot B$, the equilibrium dissociation constant is

$$K_d = \frac{[A] \times [B]}{[AB]}.$$

If another molecular species, $C$, is present that does not associate with $A$ or $B$, its concentration cancels out of the equation. Seemingly, sodium or potassium would be in this category. In fact, however, sodium or potassium ions participate halfway in protein-DNA binding events and they must be included in any complete description of the binding reaction.

The high concentration of negatively charged phosphate groups along both strands of double-stranded DNA sufficiently attracts positively charged ions like sodium or potassium so that, on average, some reside close to the negatively charged phosphate groups. As a result, they participate somewhat in protein binding and dissociation reactions because the binding of a protein to DNA can displace a number of these DNA-associated ions.

Neglecting for the moment the fact that a counterion neutralizing the negative charge of a phosphate group is not fully associated with the phosphate, the equilibrium dissociation constant for a protein, $P$ binding to the DNA, $D$, and displacing $n$ positively charged sodium ions from the DNA is

$$K_1 = \frac{[P]_1 \times [nNa^+ \cdot D^{n-}]}{[Na^+]_1^n \times [P \cdot D]_1}.$$

Typically, one measures and directly considers the concentrations of protein and DNA, but not the concentration of a counterion like sodium. Thus, an apparent dissociation constant is defined as,

$$K_1' = \frac{[P]_1 \times [nNa^+ \cdot D^{n-}]}{[P \cdot D]_1},$$

In terms of the real dissociation constant, $K_1$, the apparent dissociation constant at a salt concentration of $[Na^+]_1$ is

$$K_1 = \frac{K_1'}{[Na^+]_1^n}$$

Since the real equilibrium constants at two different salt concentrations $[Na_1]$ and $[Na_2]$ are the same,

$$\frac{K_1'}{[Na]_1^n} = \frac{K_2'}{[Na]_2^n}$$

or the apparent equilibrium constant at a salt concentration of $[Na]_2$ equals the apparent equilibrium constant at a salt concentration of $[Na]_1$ times $\left(\frac{[Na]_2}{[Na]_1}\right)^n$. Hence, when reporting apparent dissociation constants from DNA, the salt concentration at which the measurement was made must be provided.

Now we can return to the question of how much a sodium ion participates in the binding and dissociation reactions. Experimentally, the apparent number of sodium ions participating in a reaction is found to be about 0.6 times the number of phosphate groups closely approached by a binding protein.

The lower the salt concentration, the tighter a protein binds to DNA. This can be understood as follows. When a protein dissociates from DNA, the newly exposed negatively charged phosphate groups must be neutralized by positively charged ions in the buffer. The lower the concentration of these ions, the harder it is for the phosphates to be neutralized, and therefore, the tighter the protein binds.

**Supercoiling**

As seen above the molecular structures of the bases of DNA are critical in the recognition of DNA sequence by cellular machinery. In the case of single stranded DNA, as at a replication fork or in a transcription bubble, sequence recognition uses the atoms of the bases that are involved with Watson-Crick base pairing. Sequence recognition in unmelted double-stranded DNA uses the exposed edges of base pairs. These are accessible in the major and minor

groves of helical DNA. In addition to these highly important and structure-sensitive roles, the gross physical structure of DNA is also biologically important.

Three measures help describe the gross physical structure of double-stranded DNA. The role played by each of these measures is most easily understood by first considering two linked circles, where each circle can represent a strand of single-stranded DNA. If the circles remain covalently closed, the net number of times that one strand winds around the other strand is fixed and cannot be altered by distorting one or both circles Figure 7-3. This is called the

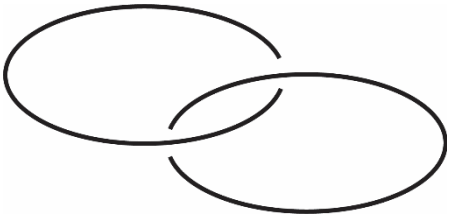


Figure 7-3. Two interlocked circles cannot be separated without opening one of the circles. The linking number of this structure is one.

linking number and is usually abbreviated as $Lk$. It is a topological invariant. What makes the linking number a valuable structural indicator is that under physiological conditions, double-stranded DNA has a strong tendency to engage in base pairing and form a helix where stretches of base paired DNA have a right-handed twist of about 10.5 base pairs per turn. This local structure of the DNA is called the twist, $Tw$. It is the number of times that one strand appears to twist about the other strand as you move along the axis of the helix. In a double-stranded DNA circle, while the linking number is fixed until a covalent bond along the backbone is broken, the number twists is not fixed and need not equal the linking number. Macro structures or global structure of the DNA can lead to changes in the total amount of twist. It is convenient therefore, to define a third variable, writhe, $Wr$ as the difference between linking number and twist, that is, $Lk = Tw + Wr$. While the linking number and twist may be easily determined by inspection for many DNA structures, methods for the direct evaluation of writhe are more complex and will not be described here.

**Examples of Supercoiling**

The interplay amongst the three variables $Lk, Tw,$ and $Wr$ is nicely illustrated with the following example. Suppose we have a 10,500 base pair linear DNA molecules. At the standard 10.5 base pairs per turn, this DNA contains 1,000 turns. If the hydrophobic planar molecule ethidium bromide is added to a solution of the DNA, it intercalates between bases because the molecules prefer the hydrophobic environment there to the more hydrophilic water. This intercalation somewhat unwinds the DNA, perhaps so that the total twist is 950 turns. In this state, let the DNA be circularized by the enzyme DNA ligase. At this point the DNA has $Lk = 950,\ Tw = 950,$ $Wr = 0$. If the ethidium bromide is now removed by providing a still more hydrophobic environment than the DNA, $Lk$ remains 950, but the twist of the DNA returns to its preferred value of 10.5 base pairs per turn, giving $Tw = 1{,}000$. In the process of accommodating these

two constraints, the DNA twists upon itself in what are called DNA supercoils, and $Wr = -50$. Figure 7B shows a supercoiled molecule.

Figure 7B. Double stranded DNA with a moderate number of supercoils.

Activities associated with DNA replication forks further illustrate supercoiling and how cells can add or remove supercoils. Consider a portion of a melted bubble of DNA containing a DNA replication fork moving to the right, Figure 7-4.

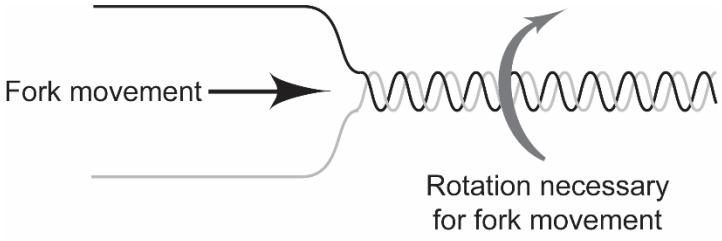



Figure 7-4. Fork movement to the right tries to rotate the DNA as shown. If rotation of this DNA is blocked or constrained, the DNA becomes overwound and positive supercoiling is generated. In living cells, the free rotation is blocked and therefore the positive supercoiling (overwinding) must be removed. requires that the double helix rotate as indicated. This rotation generates positive supercoiling that must be removed.

As the fork moves to the right, as viewed from the point where the strands separate, the DNA ahead of the fork must rotate counter clockwise. This increases the twist of the DNA ahead of the fork, but because of the strong tendency of the DNA to return to its relaxed twist of 10.5 base pairs per turn, twist decreases back towards 10.5 base pairs per turn. Since the linking number remains constant, writhe must therefore increase. That is, movement of a replication fork introduces positive supercoiling ahead of the fork. Similarly, behind the fork, negative supercoils are introduced as the DNA there is underwound. As such supercoiling would soon lead to humongous tangles and impede movement of the fork, cells use proteins called DNA topoisomerases to reduce both forms of supercoils. DNA topoisomerase I removes negative supercoils, and DNA topoisomerase II, which is also called DNA gyrase, removes positive supercoils. Note that this activity is the same as that required to making negatively supercoiled DNA more negatively supercoiled which can assist cellular processes that require melting of sections of DNA. DNA gyrase makes a double-strand break, passes a more distal portion of the

double-stranded DNA through the break, and rejoins the break. Such an operation reduces the linking number by two, and in the case of gyrase, the direction of the transfer across the break is to remove positive supercoils. DNA topoisomerase I increases linking number by one and does so by breaking one strand, passing the uncut strand through the cut, and rejoining the cut ends.

**Supercoiling and Melting**

Supercoiling density is used to facilitate comparison of supercoiling on different sizes of DNA circles. It is defined as

$$\frac{L_k - L_{k0}}{L_{k0}}$$

where $L_k$ is the linking of the DNA and $L_{k0}$ is the linking number the DNA would have if it were relaxed and possessed no supercoiling. Supercoiled DNA extracted from bacteria frequently is found to have a supercoiling density of about -0.06, that is, for DNA that, if relaxed, would possess a linking number of 100 has a linking number of 94. Thus, in the DNA extracted from the cells, $Lk = 94,\ Tw = 100,\ Wr = -6$. Rather than being willingly knotted up with the negative supercoils, the supercoiling exerts a restive force acting to reduce the net amount of supercoiling. That is, the "pressure" is to reduce net supercoiling from $-6$ to $-5$. This is equivalent to a pressure reducing the twist so that the existence of the negative supercoils pushes the DNA toward $Lk = 94,\ Tw = 99,\ Wr = -5$. That is, the negative supercoiling pushes the DNA to be less twisted, which has the effect of assisting melting. In other words, negative supercoiling helps melt the DNA. This property may explain why cells adjust the supercoiling density of DNA to be somewhat negative—it helps melt the DNA and facilitates the formation of transcription bubbles and replication forks.

**Supercoiling and Linear DNA**

Although some bacterial chromosomes are circular as are smaller bacterial genetic elements called episomes, some bacterial chromosomes and most eukaryotic chromosomes are linear. Nonetheless, the concepts of linking number, twist, and writhe are still helpful in understanding properties of the DNA.

Consider a point on a small circle of double stranded DNA, Figure 7-5. Linking number, twist,

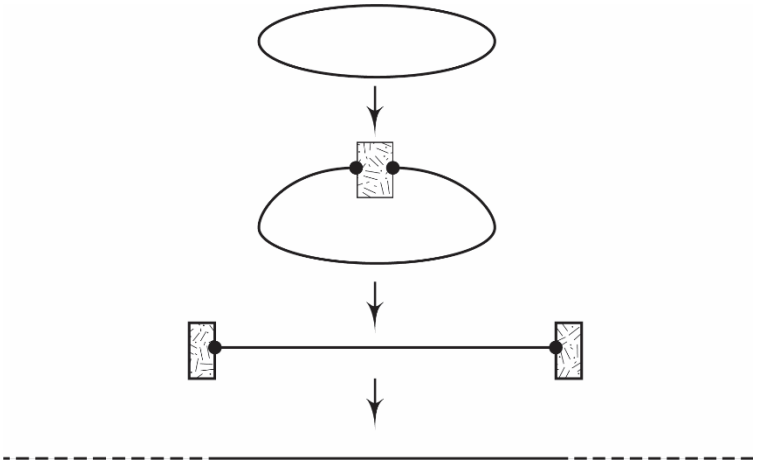


Figure 7-5. With respect to linking number, twist, and writhe, the behavior of DNA far from the ends of a linear DNA molecule is much the same as that in a circle.

and writhe can all be well defined for the circle. In this circle, the base pairs on either side of the point cannot significantly rotate or wrap with respect to one another. The same would be the case if the circle were opened at the point and each strand at each free end were attached to something large and immobile. This would remain true of the two ends and whatever they are attached to were moved apart. The DNA of the now linearized circle could still have a linking number, twist, and writhe defined for it. Finally, suppose that the DNA is vastly extended at each end. Then, any of the topological changes would not easily reach the ends and for considerations of supercoiling, the DNA is a circle and can display many of the properties that can be displayed in small circles.

**Energy Required to Bend DNA to Form a Nucleosome**

As mentioned earlier, packing the very long chromosomal DNA molecules into the nucleus, and at the same time leaving a good portion of the chromosomal genes available for transcription poses special problems. Several levels of organization can be recognized in the chromosomal DNA, but it is not yet known to what extent this organization pays a role in the regulation of gene activity. The first organizational level is wrapping stretches of several hundred base pairs of DNA around a protein core to form a nucleosome. The nucleosomes themselves then pack in several additional organizational themes.

Wrapping the DNA around the histone core of a nucleosome involves a balance of energies. The relatively stiff DNA molecule must be forced to bend into small circles. Counteracting the forces to keep the DNA straight rather than curved are electrostatic forces and hydrogen bond Interactions between the DNA and the histone core of a nucleosome. These hold DNA to the histone surface. The various energies involved are large and some cannot be well calculated. Therefore, the objective of this section will be only to crudely estimate the energy that is required to bend the DNA around a nucleosome core by making use of the properties of random walks that were derived earlier.

If we ignore natural bends and base composition effects, then at absolute zero, DNA molecules would be straight. Therefore, at physiological temperatures, assuming the DNA is not kinked

and contains no melted regions, the curvy nature of DNA comes from the accumulated temperature induced flexing at each base pair. In most physical systems, the force acting to restore the system to equilibrium is proportional to the displacement from equilibrium. In the case of bending DNA, we will assume this to be the case so that the force required to generate a bend will be proportional to the size of the bend.

It should be mentioned here that the base composition and sequence of DNA can and does introduce moderate bending. A naturally bent segment of DNA can significantly assist the formation of a nucleosome as less "external" bending would be required. Biology utilizes this property to position some nucleosomes to facilitate recognition of gene and control regions of DNA. As a result, the main conclusion of this section is that the energies involved in nucleosome formation are large, but must be decently balanced in order that nucleosomes can be added to and removed from DNA.

To begin the energy estimation, approximate a base pair as consisting of a rigid link and a joint about which bending is possible, but that the DNA resists bending with a force proportional to the amount of bend from the equilibrium position. As derived earlier, by the equipartition theorem, in this case, the average thermal bending energy in the DNA per base pair will then be $(KT/2)$ where $K$ is the Boltzmann constant and $T$ is the temperature. Since there are two directions the DNA can bend, this should be multiplied by two. In a mole of the DNA molecules, the average bending energy at one base would be $RT$.

A result from polymer physics is that the expectation value of the cosine of the angle $\theta$ between a vector that is tangent to the polymer at position 0 (zero) and a tangent vector at a distance $L$ away from position 0, along the contour of the chain falls exponentially with distance as

$$< cos\theta >= e^{-L/P}$$

where $P$ is the persistence length. Double stranded DNA in physiological buffers has a persistence length of about 150 base pairs. Therefore, at a distance along the DNA of 150 base pairs, the average total bend from its initial direction will, on average, be about $68°$ since

$$\cos(68°) = e^{-1}$$

Therefore, in a two-dimensional random walk (the two orthogonal angles of a bend at each base pair) of 150 base pairs, the average amount of bending will be 68°.

Recall that the distance from the achieved of a random walk is the step size times the square root of the number of steps taken. Therefore,

$$68° = Average\ bend\ per\ base\ pair° \times \sqrt{150}$$

so that the average bend per base pair under conditions where the persistence length is 150 base pairs, is about 5.5°. Therefore, the work required to bend one base pair of DNA by 5.5° is about $KT$, or for a mole of base pairs, $RT$.

Since 146 bp of DNA wrap 1.65 times in forming a nucleosome, if there is equal bending at each base pair, the bend per base pair is 1.65 x 360°/146 = 4.1°. In systems where the restoring force is proportional to the displacement, as we are assuming here for bending, the work required to displace the system from equilibrium is proportional to the square of the

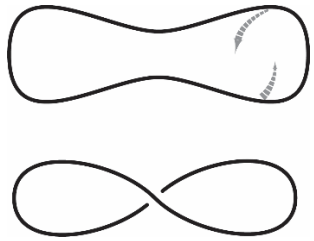
displacement. Therefore, to bend 146 base pairs of DNA, each base pair by 4°, the work required is

$$146 \times \left({}^{4}/_{5.5}\right)^{2} KT = 78KT$$

In this estimation we have approximated a base pair of DNA by a single rigid link and a single bendable joint whose restoring force when bent is proportional to the amount of the bend. It we approximate a base pair as possessing two rigid links and two joints with equal flexibility, the calculated energy should be very close to the single link single joint value. Repeating the above calculations with a persistence length of 300 links and joints shows this to indeed be the case.

**Problems**

1. Show that the pattern of hydrogen bond donors and acceptors in the major groove of DNA permits unambiguous reading of the four possible base pairs, but not so in the minor groove.
2. Why is RNA more suited than single-stranded DNA of the same sequence for the formation of tertiary structures?
3. Why when annealing longer strands of DNA is it sensible to slowly decrease the temperature rather than jump from a melting temperature to an annealing temperature?
4. Show that cutting both strands of DNA, passing a more distal section of the DNA through the cut, and then rejoining the cut ends changes the supercoiling of the DNA by two.
5. The single line in Fig. 7-C represents two strands of DNA. In the top figure, twist is zero so that $L_k = T_w = W_r = 0$. If the right half of the molecule is twisted one half of a full rotation to produce the bottom figure, what now are the values of $L_k$, $T_w$, and $W_r$?



# 8. Protein Structure and Function

## Proteins, the Universal Tools

Most of the thousands of different chemical reactions that take place in cells are each catalyzed by a different protein. Proteins also play multiple structural roles in cells. As proteins are linear chains consisting of only 20 different types of amino acids and occasionally a few specially modified amino acids, what accounts for the diversity of their activities? Many sequences of amino acids, as specified in DNA by the genetic code, can fold into complex three-dimensional structures that are each capable of binding to specific molecules and often capable of participating in the chemistry of reactions catalyzed by the protein. This wide variety of activities necessitates that the collection of 20 amino acids include a variety of sizes, shapes, physical properties, and chemical properties.

## Structure Determines Function, or Does It?

While it is nice to be able to catalog the genes, proteins, RNAs, and reactions that are assisted by proteins and RNA's, still more useful would be the ability to design proteins and RNAs that can carry out any reasonable chemical or physical activity. Beyond this almost utopian molecular understanding, would be the further goal of understanding the consequences to cells, organisms, and populations of changes in proteins and RNAs.

Our experiences with the physical world--screw drivers, wrenches, pens, paper and so on--have taught us that structure determines function. Therefore, it is not surprising that ever since we have known that at least some proteins possess structure, we have sought to determine structures as the first step in determining the mechanism of action of a protein, and perhaps from there to develop the tools and understanding necessary to being able to design proteins with new and desirable characteristics.

One hint that a protein's static structure may not be the complete story has been the finding that some proteins, and some parts of other proteins appear not to possess regular structures, but instead are disordered. Whether these disordered regions briefly acquire structures adequate to perform some activity or whether disordered regions code for a collective physical property like inducing a phase transition is currently being actively investigated.

## How Can Structure Be Determined?

Much effort has been devoted to developing physical methods for determination of the structure of a protein, and despite the recent development of a handful of powerful tools for structure determination like high intensity X-ray beams, high resolution NMR machines, and cryo-electron microscopes, experimental determination of a protein's structure is still an expensive and lengthy process. Biochemists have also long sought to determine three-dimensional protein structures merely from the one-dimensional sequence of their amino acid constituents. To this end, a surprisingly large fraction of the biochemical and biophysical research over the past fifty years has been explicitly and implicitly directed towards the prediction of structure from sequence. As explained below, highly significant progress has recently been made in predicting structure from sequence, and therefore much research effort will shift to the determination of function from structure.

Crucial to the quest of predicting protein structure is the requirement that the sequence itself plus the laws of physics and chemistry be sufficient to determine a protein's three-dimensional structure. Thus, the demonstration by Christian Anfinsen that many proteins can indeed spontaneously refold to their native structure from a denatured state without the presence of other cellular components was highly important. Since Anfinsen's crucial experiments, it has also been possible chemically to link together amino acids in the same sequence as is found in active enzymes and for the synthetic protein to fold on its own into an enzymatically active structure.

Not all naturally occurring proteins can refold on their own. Some proteins do not refold upon removal of denaturing conditions, and *in vivo* folding of many of these requires auxiliary cellular proteins. Even in these pathological cases however, it is thought that all the information necessary for a string of amino acids to be folded into a specific three dimensional structure is contained in its amino acid sequence.

### Plan of the Chapter

Because the study of protein structure and function has been extensively studied for many years, much is known on the topic. Biochemistry and Molecular Biology texts contain much of what has been learned, and they should be consulted for detailed information. The material in this chapter summarizes some of what is known and is organized to encourage the viewpoint that real understanding allows one to deduce a sizeable amount of additional information.

Proteins fold and function in response to forces acting on the protein's individual atoms. These forces are electrostatic, hydrogen bond, Van der Waals, and hydrophobic. Their properties and origins are well described and explained in standard biochemistry texts and also will not be repeated here.

### Amino Acids, the Basic Building Block of Protein

The vast majority of naturally found proteins each contain all of the following 20 alpha-L-amino acids, Figure 8-1. These consist of an amino group ($-NH_2$), a central carbon atom designated as the $C_\alpha$ carbon, with a side

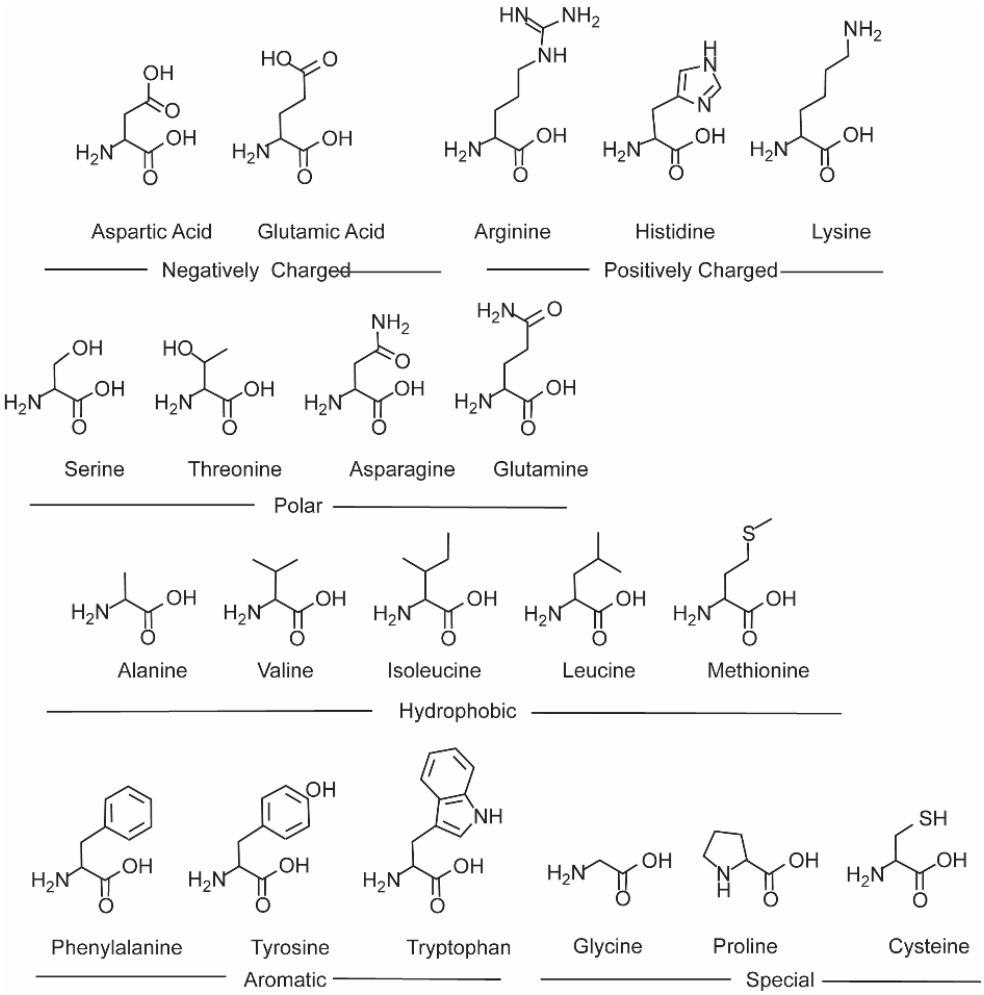



Figure 8-1 Structures and classification of the twenty amino acids. The structures as drawn do not exist in nature, as at physiological pH's, the carboxyl groups are deprotonated and have a negative charge, and the amino groups are protonated and have a positive charge.

group (R) that is unique to each amino acid, and a carboxyl group (-COOH). The fact that these twenty amino acids have been retained through eons of evolution means that each amino acid possesses unique and important properties. In thinking about protein structure and function, it is convenient however, to consider subsets of the 20 organized by their structures and prominent chemical properties.

Amino acids can be acidic, basic, or neutral depending on the side group R. Acidic amino acids have an additional carboxyl group in the side group (-COOH), making them negatively charged at physiological pH, while basic amino acids have an additional amino group (-NH2), which makes them positively charged at physiological pH.

Neutral amino acids can also be hydrophobic or hydrophilic. Hydrophilic amino acids have polar or charged side groups that can interact with water whereas hydrophobic amino acids have nonpolar side groups and tend not to interact with water.

The aromatic amino acids possess aromatic side groups, and a few amino acids, glycine, proline, and cysteine, are best considered individually.

**The Peptide Bond and Polypeptide Backbone**

The consecutive amino acids in a protein chain are joined by peptide bonds, Figure 8-2. As a result of the partial double bond character of the bond due to resonance of the C=O double bond and the C-N single bond with C-O and N-H, the peptide bond is stable and planar, or

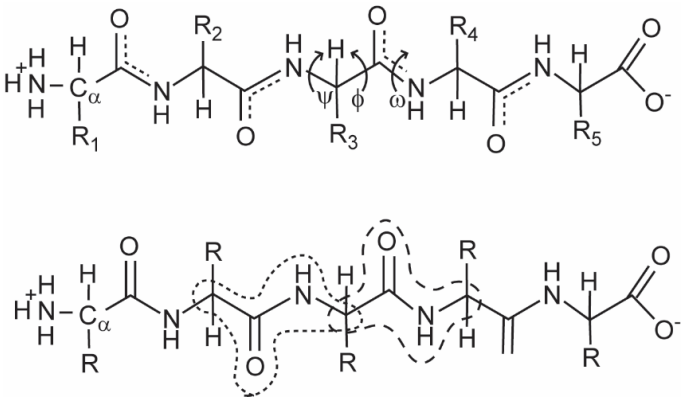


Figure 8-2. Two views of a pentapeptide. The top shows the resonance character of the peptide bond between the C and N atoms and the bonds flanking the alpha carbon about which the $\phi$, $\psi$, and $\omega$ angles are defined. In the bottom view the two-enclosed areas include the atoms, all of which lie in the same plane. This is a result of the planar structure of the peptide bond and the lack of free rotation about the peptide bond. All the peptide bonds in these diagrams and in most proteins are *trans,* which means that the $C_\alpha$ carbon atoms and the R-groups connected to them lie on opposite sides of the intervening peptide bond.

almost so, although structural constraints in folded proteins can sometimes distort the peptide bond to be a few degrees from planar. The two remaining bonds along the peptide backbone allow free rotation until such rotation brings other atoms in the protein into contact. Magnitudes of these two rotations, denoted by $\phi$ and $\psi$, for each of the amino acids of a protein largely define the three-dimensional structure of the protein.

Another crucial property of the peptide bond is that for each amino acid residue of a protein the peptide bond leaves the oxygen atom of the carbonyl as a good hydrogen bond acceptor and the nitrogen as a good hydrogen bond donor. Therefore, hydrogen bonds can form between these backbone atoms of the protein and with other donors and acceptors within the protein or with water. The energetics do not strongly favor one class or the other. Why then, is it found that in folded, structured proteins, that virtually all the hydrogen bonds involving the backbone amide nitrogen and carboxyl oxygen atoms are to other donors and acceptors within the protein rather than to water? One important factor in favoring the formation of bonds within the protein

rather than to water is the chelate effect that was discussed earlier. When a protein is partially folded, some of the carbonyl and amide groups are well positioned to form a hydrogen bond, and because formation of the bond under these conditions does not suffer an entropy penalty, the strength of such a bond is particularly strong.

### Rotamers of Side Chains

Examination of the structures of many proteins has revealed that the side chain atoms strongly tend to be confined to favorable conformational isomers, which are called rotamers. The numbers of rotamers of the different amino acids varies from three for cysteine, serine, and valine to 27 and 34 for lysine and arginine. Therefore, only a relatively small number of rotamers need to be considered when predicting the structures of the side chains of amino acids in proteins.

### Secondary Structures

The repeating pattern of hydrogen bond donors and acceptors along the backbone of a polypeptide chain allows several basic regular, meaning each residue of the peptide has the same values of the phi and psi angles, structures of polypeptides that maximize their hydrogen bonding. Figure 8-3 shows the hydrogen bonding patterns between two parallel polypeptide segments and two antiparallel segments. Additional segments can be aligned alongside to form extended sheets. Such parallel and antiparallel segments are called beta-sheets, beta because they were the second regular structure predicted by Pauling and Corey after the alpha helix. In the sheet structures the R groups extend above and below the sheets.

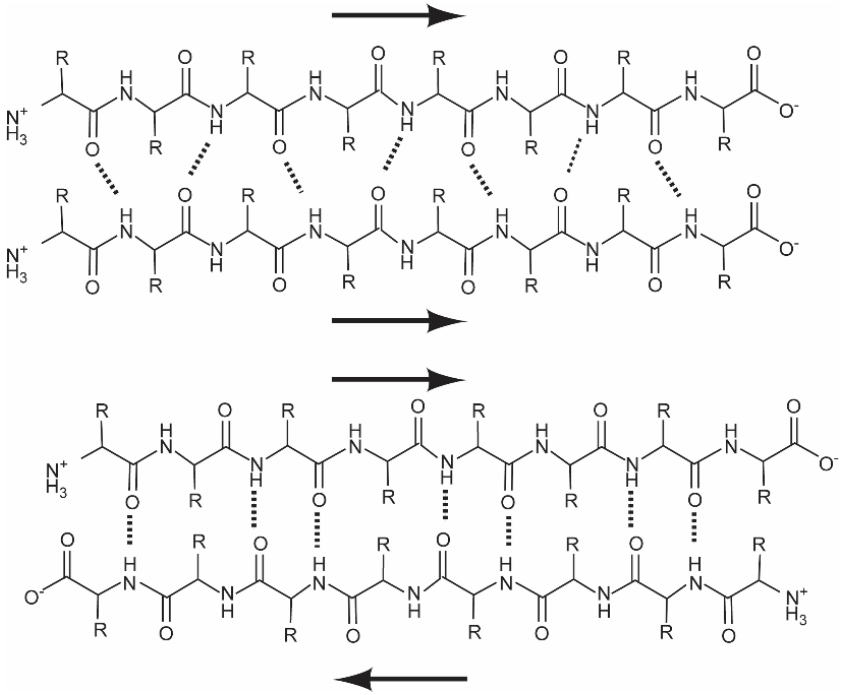


Figure 8-3. Hydrogen bonding patterns of parallel and antiparallel β-sheets.

Another regular secondary structure that maximizes hydrogen-bond formation is the alpha helix, Figure 8-4. In it, the polypeptide backbone winds in a right-handed helix such that hydrogen

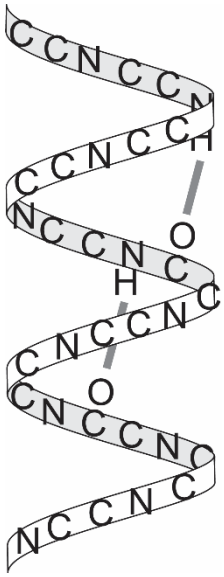


Figure 8-4. The backbone atoms of an alpha helix with the carboxyl and amide hydrogen atoms shown for two backbone atoms and the hydrogen bonds between them. The geometric complexity of the helix plus that of the tetrahedrally bonded alpha carbon of each amino acid means that the more accurate the drawing, the less intelligible it becomes. Better understanding of the structure can be obtained by working with a molecular display computer program like PyMol.

bond donors and acceptors are aligned nearly one turn apart. Because an alpha helix contains 3.6 (nearly 3 2/3) residues per turn, but the hydrogen bond donors and acceptors of any hydrogen bond are only 3 1/3 residues apart, the hydrogen bonds are somewhat tipped relative to the axis of the alpha helix. The side chains in an alpha helix are directed outward, and atoms of the protein backbone largely fill the core of the helix. Another distinguishing feature of the alpha helix is that all the hydrogen bonds are directed towards the same end of the helix. This plus the unpaired hydrogen bond donors and acceptors at each end of the helix give the alpha helix a polarity where its N-terminus possesses a partial positive charge and the C-terminus a partial negative charge.

**Higher Order Structure**

Primary structure is the sequence of amino acids in a protein, and secondary structures are the regular structural units discussed in the previous section. Tertiary structure is the full three dimensional structure of a polypeptide. Because proteins consist of closely packed atoms, all of which have about the same diameter, when viewed as space filling models with the sizes of atoms to scale, a protein appears to be a hopeless mess. But because proteins do many important things and their structures seem to be so important, scientists have devised ways of representing the atoms and amino acids, and secondary structures of proteins in ways that look beautiful and provide useful pictures of the organization of a protein's structure. Computer graphics also greatly aid in understanding protein structure because the ability to rotate computer-drawn models facilitates understanding the three-dimensional structure as does the color coding of representations of atoms and of a protein's secondary structures.

Quaternary structure of a multisubunit protein is the arrangement of the subunits. For example, a homotetrameric protein can either be tetrahedral or square. The arrangement of capsid proteins in viruses often is helical or based on the icosahedron. In either case, the relation of each subunit to its immediate neighbors is the same or very nearly the same throughout the capsid.

**Considerations for Predicting Protein Structure**

How could one predict the folded structure of a polypeptide, that is, predict the three dimensional structure of a protein from its primary sequence of amino acids? One way would simply be to mimic the way nature does it, that is, by calculating all the forces on each atom of the polypeptide chain and of a surrounding shell of water molecules and use Newton's equations of motion for each atom and follow the protein's structure with time until it is folded. This is not feasible however, because the forces are not known with sufficient precision, and even more importantly, because we lack sufficient computational power.

Another approach in predicting protein structure is to expect that the folded structure will be that with the lowest energy. While this may be true for most proteins, it cannot be rigorously demonstrated to be true for all proteins. For example, consider a very short polypeptide chain whose N- and C-terminal amino acid side chains can bond to circularize the peptide. Suppose that the work required to bend the peptide into a circle in order that the bond can form is greater than the strength of the bond itself. In this case, the energy of the folded, circularized, state is higher than that of the unfolded linear state.

For the majority of proteins however, a reason for thinking that the folded state is the lowest energy state is that adding energy to a protein by heating it causes it to unfold. Therefore, such a protein must be folded before heating. Taking this to the extreme, presumably, the protein would also be structured in the limit as the absolute temperature approaches zero.

An entirely different way of looking at the question helps explain the conundrum of why a protein might spontaneously reach a conformation of lowest free energy. Consider a protein molecule that has folded to its lowest energy state. What is the probability at room temperature at any instant that a particular one of its intramolecular hydrogen bonds is momentarily broken and neither partner of the former bond has formed new bonds to water. This can be calculated from the Boltzmann distribution where the probability of such a case is $P(\Delta E) = e^{-\Delta E/kT}$. Take the energy of the intramolecular hydrogen bond to be around 4 kcal/mole or about $0.66 \times 10^{-20}$ calories per molecule. The value of $kT$ is 9.83 x $10^{-22}$ cal. Therefore $\Delta E/kT \cong 6.7$ and the probability that a bond is broken is about 0.001. What this means is that even at room temperature the vibrations of thermal energy have a low probability of breaking a hydrogen bond in the protein. As the temperature increases, the Boltzmann equation predicts that the probability of the bond being broken increases. Amusingly, the Boltzmann equation also allows a protein molecule to be a thermometer. The effective temperature of a folded protein could, in theory, be determined from the fraction of protein that is found in a higher energy state, that where a hydrogen bond is broken.

Predicting the structure of a protein by finding the structure with the lowest energy sidesteps the problem of knowing at what temperature a simulated folding should be done as this is equivalent to determining the structure at absolute zero. Since proteins unfold at elevated temperatures, if a simulation were done at too high a temperature, there would be no hope that

a correct prediction would yield a folded structure. Another virtue of this approach is that as the temperature approaches absolute zero, the lowest energy state or states if there are more than one, is fully occupied and alternative conformations need not be considered. Finally, at absolute zero, entropic contributions to free energy are zero, and it suffices to consider only the enthalpic contributions to a protein's free energy. That is, knowing the strengths of only the atom-atom interactions is required and the counting of equivalent microstates is avoided.

**Side Issues Related to Predicting Protein Structure**

It is reasonable to ask why the equilibrium structures of many molecules is their lowest energy state. After all, if a polypeptide is simulated as immersed in a buffer at a temperature above absolute zero, the polypeptide can lose energy to the buffer or the buffer can lose energy to the protein. Why is the energy flow predominantly in the direction of flow to the buffer? The answer is that in the protein while there are potentially only a few thousand interatomic bonds, vibrational states, or strained states that can take up energy from the buffer, in the buffer there are on the order of Avogadro's number of states that can take up energy from the polypeptide. Therefore, it is far more likely that the system of protein plus buffer will be found in a state where it is the buffer that has taken up some energy from the protein rather than the reverse. It should be noted however that at sufficiently high temperatures, the direction of the energy flow will become two way and the polypeptide can be found in higher energy states, in most of which the polypeptide is unfolded.

Another side issue is that a few proteins are known that unfold in the cold. Furthermore, a careful accounting of the thermodynamics of protein-water interactions shows that many proteins are somewhat destabilized with lowering temperature. Experimentally these are not known to cold denature only because water freezes first. The cold sensitivity issue however, can be avoided in computations by not including the water-protein interactions that show a strong temperature effect due to structuring and destructuring within the water itself.

A third side issue concerns the relationship between a physical description of a system at equilibrium using the Boltzmann distribution energies only and a thermodynamic description that uses energies and entropy. In a Boltzmann description, the probability that a system containing n possible states occupies the mth state is

$$P_n(m) = \frac{e^{-\frac{E_m}{kT}}}{\sum_{i=1}^{i=n} e^{-\frac{E_i}{kT}}}.$$

In the thermodynamic description, the equilibrium constant for being in the mth state vs. the nth state is

$$K = e^{-\frac{(\Delta H - T\Delta S)}{RT}}$$

where $\Delta H$ is the enthalpy difference between states $m$ and $n$, and $\Delta S$ is the difference in entropies between these two states. The trivial difference between the two formulations is that the physical description uses energies per molecule and the thermodynamic description uses quantities per mole. The more relevant difference is that the physical description seems to omit consideration of entropy. This is not the case however as entropy is $kln(\Omega)$ where omega is the number of equivalent microstates of the system. In the physical description there are then

omega identical contributions to the sum over states term or in the numerator term in the probability equation. Note that this is the same as the heuristic introduction of the concept of entropy that was developed in Chapter 3.

**The Problem with Energy Minimization for Structure Prediction**

When the temperature is not too high, most proteins are capable of folding on their own to their final, active, and stable conformation. As discussed above, the "final" conformations possess lower free energy than other partially folded or unfolded conformations. However, in the folding process, as thermal motion jostles a protein, and its atoms move about, occasionally the protein may reach a conformation of lower free energy that is not a global energy minimum, but only a local energy minimum. The protein or part of the protein will briefly remain stuck in this conformation. Eventually random thermal motion will kick it out of this potential local energy minimum and the protein can proceed on towards a global energy minimum. This process is under way everywhere in the protein and gradually the protein gives up most of its free energy as it settles into accessible low energy conformations and reaches thermal equilibrium with its surroundings. Higher energy conformations of the protein will be occupied by the protein in accordance with the Boltzmann distribution as discussed above. In light of these properties, the prediction of the three-dimensional structure of a protein from its primary sequence can be thought of as an energy minimization problem.

Finding the lowest energy conformation of a protein should be very simple, merely calculate the energy of each conformation and take the lowest one. Levinthal, however, pointed out the futility of this approach. Assume a protein has 100 amino acids. This gives 100 phi and 100 psi angles to consider. If we are being really coarse, consider that each can assume one of ten possible values, that is, 0, 36, 72, 109,… degrees. The number of conformations available to the protein then is $10^{200}$. As the number of atoms in the entire universe is estimated to be *only* around $10^{80}$, dealing with a list of each of the individual conformations is not possible.

Instead of listing each conformation and calculating its energy, why not let the system run down its energy surface to the energy minimum in a process called gradient minimization? To do this, merely minutely change the phi and psi angles of each amino acid in directions that reduces the energy of the protein. From time to time in the process, the different rotameric states of the amino acids can be tested and chosen also to minimize energy. This process could be repeated many times until the protein arrives at its energy minimum, a point from which any change in any phi or psi angle and any change in the rotameric state of any amino acid always increases free energy. While this energy minimization process works for simple molecules, it miserably fails for more complicated molecules. The reason for the failure is that the energy surface, which is the free energy as a function of all the relevant variables, namely the phi and psi angles of all the amino acids and of all the rotameric states, is not a simple declining slope with the energy minimum at the bottom. It contains a huge number of mountains and ridges and deep valleys and basins. Once the structure falls into a local energy minimum, it has no way to escape.

Nature solves the problem generated by local energy minima by using random thermal motion. A protein in the process of folding will fall into local energy minima, but eventually, by chance, the relevant part of the protein will temporarily have sufficient energy to escape from the local minimum and allow the folding process to proceed. The probability that the part of the protein that is stuck in the energy well acquires the necessary energy, $\Delta E$, to escape is given by the

Boltzmann energy distribution given earlier, $P(E) \propto e^{-\Delta E/kT}$. Thus, the higher the energy barrier to surmount, the less likely or more protein escape.

### A Computational Method to Escape Local Energy Minima

It is not hard to devise a computational scheme that mimics Nature's thermal escape method. The Monte Carlo method with the Metropolis criterion does this. In this, from one conformation, a random and small change in the configuration of the protein is generated and the new free energy of the system is calculated. If the new energy is lower than the previous energy, the new coordinates are accepted. If the new energy is higher than the previous energy, the previous coordinates are retained. However, and this is the key point, with a probability proportional to the Boltzmann value, $e^{-\Delta E/kT}$, where $\Delta E$ is the energy difference between the old and the possible new state, the new state with higher energy is chosen. Hence, the system can jump out of shallow energy wells with a few tries, and can jump out of deep energy wells if given enough tries.

If Nature uses essentially a Metropolis Monte Carlo method to fold proteins, why can't we? The main problem is numbers. The number of local minima in the energy surface and the height of the ridges and mountains would require more computations than can reasonably be performed. Nature, however, "can do the calculations" because it is simultaneously running a huge parallel processor in which each atom of a system consisting, at the minimum, of the protein plus a surrounding shell of water perhaps 20 angstroms thick, is seeking an energy minimum, but its energy is dependent upon the coordinates of all the other atoms in the system. Thus, it can perform enough computations to fold a protein.

Molecular dynamics is similar to Metropolis Monte Carlo in that individual atoms are considered and potential energies are calculated from the positions of atoms. Molecular dynamics more closely mimics the way nature folds proteins than does the Monte Carlo method. In it, initially, each atom is given a random velocity, and after that the motion of each atom is governed by Newton's equations of motion where the forces on each atom are given the electrostatic forces, Van der Waals forces, hydrophobic forces, and the interatomic forces generated by stretching, bending, and twisting the covalent bonds in the protein. Since Newton's equations cannot be solved analytically for even a three-body problem, they certainly cannot be solved for a system containing hundreds of thousands of atoms. Instead, the equations must be "solved" numerically in which time is incremented in very small steps and new positions, velocities and interactions are calculated at each step. The time steps must be small relative to the vibrational periods of the atoms, and typically they are on the order of $10^{-12}$ seconds. If a protein in Nature were to fold in one second, and the protein plus water shell contains 100,000 atoms, and only relatively local atom-atom interactions are calculated, even the fastest supercomputers are still not up to the task of folding a protein.

### Folding Methods that Do Work

Two rather different methods have been found to be successful in predicting the structure of proteins. One, named Rosetta explicitly utilizes information in the RCSB protein structure database and explicitly calculates trial conformations and calculates energies. It can be summarized as being based on physical interactions. As the predicted structure appears to close in on a final structure, it refines its computations, for example by shifting from gross

representations of amino acids to full atom representations, and instead of relying totally on Metropolis Monte Carlo methods, it uses energy gradient methods to reach a local minimum, which, if the preceding steps have been successful is sometimes also the global energy minimum.

Predicting protein structures with Rosetta has been greatly aided, particularly in the case of larger proteins, with the use of compensating mutations. An inactivating mutation in a protein can sometimes be compensated for by a change in an amino acid that lies near or touches the first mutation. Families of related proteins can be identified by sequence similarities. It is assumed, and has proven largely to be true, that members of such families are evolutionarily related. Therefore, if two amino acids are near one another or touching in the three-dimensional structure of one of the related proteins, the corresponding amino acid residues in other members of the family also likely lie near one another. As a result, a random mutation that changed one of the residues and inactivated the protein may then have been compensated for by a mutation in the nearby partner residue. Hence, in a collection of sequences of related proteins, pairs of amino acids that over evolutionary time have changed together are likely to lie close to one another in the tertiary structure. Several structural constraints like this provide great assistance to protein structure predictions and enable the formation of good structure predictions even for very large proteins.

A second approach to predicting protein structure from sequence is to make use of the more than 100,000 known structures that are contained in the protein data bank, PDB, to train a large neural network. The program alphafold has used some physical data and neural network methods to predict structures. Its protein structure predictions are considerably better than those of Rosetta. Alphafold was then succeeded by alphafold 2, which does not explicitly use physical data on proteins and amino acid, but uses the pdb and sophisticated artificial intelligence methods with a large neural network to make its predictions. In some cases, its predictions match or exceed the resolution of structures determined by X-ray diffraction methods. As the amount of calculation required for alphafold2 to make a prediction is not excessive, the program has been used to predict the structures of most of the proteins whose amino acid sequences are contained in sequence data banks. The program RoseTTA also uses large neural networks and artificial intelligence to predict protein structures. Its particular strength is its ability to predict structures of protein complexes whereas alphafold2 can predict structures for only single polypeptides.

**Neural Networks**

An artificial neural network is a simplified computer simulation of how portions of the brain are thought to act. Such a simulation consists of multiple layers of interconnected nodes. Input to the network is achieved by assigning values assigned to nodes in the first layer, called the input layer. Each of these nodes is connected to each of the nodes in the second layer, which is called a hidden layer. Similarly, each of the nodes in the first hidden layer is connected to each of the nodes of the next layer, and so on so that a “computation” from an input feeds forward layer by layer until the final layer, the output layer, is reached. Figure 8-5 diagrams a simple neural network containing only one hidden layer.

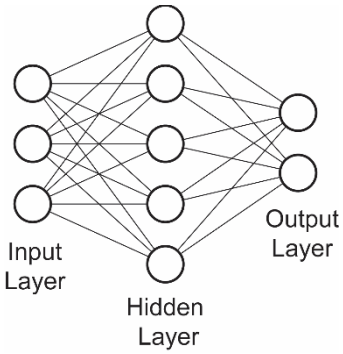



Figure 8-5. A three-layer neural network showing the nodes of the input, hidden, and output layers.

In a neural network, the strength of the connection between a node in one layer and a node in the next layer is determined by a weighting parameter specific to that pair of connected nodes. The input to the connected node in the next layer equals the output value of the first node multiplied by the weighting factor. Then, the values of all the inputs to a node are summed in the process of determining that node's output value. This value is then multiplied by the weighting factors of the connections from this node to the nodes in the next layer.

A node's output value depends on the sum of the inputs to that node and a transfer function. Multiple different transfer functions have been tried in neural networks. Generally, they are chosen to mimic biological neurons that "fire" when they receive sufficient excitatory inputs. That is, when the sum of the inputs to a node is low, a node's output is close to zero, and when the sum is high, the output is close to one.

As initially programmed, a neural network program is untrained and incapable of producing useful output for an input. Because of the enormous size and complexity required for neural networks to perform useful "calculations", deterministic assignment of weight values is impossible. Therefore, initially, the weights in a network are randomly assigned. The network is trained bit by bit by adjusting the weights of the internode connections so that the outputs in response to inputs in a training set more and more closely match what they should be. In training, an input, or a set of inputs are applied, and size of the error or errors between the network's output and what would be the correct output is obtained. After this, the values of the weights between the last hidden layer and the output layer are adjusted to reduce the size of the error. Then, the weights between the next to last hidden layer and the last hidden layer are similarly adjusted. The process is repeated, layer by layer until adjustments have been made to the entire set of weights. In training, the process of weight adjustment is repeated, over and over, and for many, many different inputs.

If all is well, the values of the different weights from successive training sessions converge to reasonable values, but if things are not well, each weight can bounce around and never settle down to a specific value or can explode to unrealistic values. For purposes of adjusting the weights, multiple ways of using the errors between the predictions and the desired output have been tried, and multiple different transfer functions have been used. Often, a network is not entirely feed forward. One way to give a network a memory of a previous output is to use the network's output as part of the input for the next operation of the network.

Suppose we wish to design a neural network for predicting whether the middle amino acid in a sequence of nine residues will possess phi and psi values consistent with an alpha helix. Such a network therefore could have nine groups of 20 input nodes. Each of the 20 nodes in the nth group would correspond to a particular amino acid in the nth position of the nine residues in the sequence. Setting the value of a node in the nth group of 20 to one indicates that its corresponding amino acid is present at position n in the input sequence of nine. It might be reasonable for the network to contain three hidden layers, each of around 30 nodes, and a final output layer consisting of two nodes A and B. The network's prediction would be considered helical of A>B and nonhelical if A<B.

Currently the neural networks that are being used with language input and output may consist of 20,000 input and output nodes, and 50 hidden layers, each consisting of $10^5$ nodes. Each of the different input nodes may correspond to a syllable or word. Training such a network can use enormous amounts of training data, take months, and cost tens of millions of dollars.

**Static Structure is not the Final Answer**

As noted earlier, knowing the structure of a protein likely is a necessary intermediate step in determining the mechanism of action of a protein. Only rarely does the structure of a protein clearly reveal the mechanism by which that protein achieves its function. One beautiful example of an obvious mechanism is the sliding clamp associated with DNA polymerases. These ring-shaped proteins encircle the DNA and prevent the polymerase from dissociating inappropriately. For most proteins, however, for which their structure is known, what they bind to and what they do with what they bind to is hard to deduce from their structure alone. Often, complex and difficult biochemical investigations are required to determine the mechanism of action of even the more simple proteins.

One difficulty in mechanism determination from structure is the likelihood that the static structures determined for many proteins are slightly different from the structure most relevant to the proteins' actions. That is, the activity of proteins depends upon some flexibility and flexing. Evidence for this comes from the fact that while one defective gear or even a single misshapen gear tooth will completely stop a mechanical watch, most amino acid substitutions in proteins either have little effect or reduce activity by a factor of 5 to 20, and do not totally abolish activity. This means that in the mutants, the protein's internal motions still bring the active site of the protein to its functional shape, but just for a smaller fraction of the time.

**Problems**

1. Why are there no linear homotetrameric oligomers?

2. Regular icosahedra possess 12 five-fold axes of symmetry where planar equilateral triangles meet at a pyramidal point. If six planar equilateral triangles meet at a point, the structure is planar. Thus, we could say that in a regular icosahedron, the angular deficit at each of the 12 vertices is 60 degrees and the total deficit is 720 degrees. In a cube, the angular deficit at each of its eight vertices is 90 degrees and the total deficit is 720 degrees. Comment and generalize.

3. Estimate the density of a protein relative to water by calculating the total volume occupied by each atom type, using the mass and radius of each atom type, and assuming that

the atoms in a protein and in water pack as closely as they do in random close packing of uniform spheres.

4. It has been proposed that the precise frequencies of atom vibrations and structural elements in proteins play a key role in the chemical reactions catalyzed by many enzymes. Argue that this idea is incorrect by starting from the fact that many organisms can survive nearly complete substitution of hydrogen by deuterium.

5. Suppose you have a set of, say, twenty amino acid sequences that are aligned by sequence. Devise an algorithm for the identification of coevolving pairs of the amino acids.

6. In the Metropolis Monte Carlo method for seeking a global minimum, when new values of the variables are chosen for a new step, why does the method fail if the changes in the variables are too large or too small?

7. What is a Ramachandran plot and what is it good for?

# 9. Hemoglobin

## Why Think About Hemoglobin?

How can energy be provided to the various tissues in the body? One way would be to synthesize high energy ATP from ingested food and then distribute it throughout the body to tissues that need it. Evolution, however, seems to have provided us with a different style of energy system. Most likely this is a result of the fact that multicellular organisms evolved from unicellular organisms that had already developed metabolic pathways for the generation of high energy intermediates like ATP. Perhaps it was easier to evolve pathways for the transport and utilization of sources of the metabolic energy than to distribute ATP itself.

After the great oxidation event on earth that occurred about two billion years ago of oxygenating the atmosphere, the oxidative phosphorylation system became possible. This energy conversion pathway generates ATP by the flow of electrons from glucose or fatty acids to oxygen with the ultimate formation of $CO_2$ and $H_2O$. The production of high energy bonds as the electrons move down the electron transport chain is analogous to a stream flowing down a hillside and driving multiple waterwheels. Using glucose or fatty acids as a cellular source of chemical energy is highly efficient. For example, one molecule of glucose and six molecules of $O_2$ can power the conversion of up to 36 molecules of ADP to ATP.

Small wonder that cells evolved the highly efficient $O_2$ carrier, hemoglobin. In large multicellular organisms, glucose or fatty acids and oxygen must be delivered to cells to satisfy their energy needs. While quantities of glucose or fatty acids adequate to supply the energy needs of cells can be dissolved in blood plasma, that is, blood without red blood cells or other cell types, the requisite quantities of oxygen cannot be directly dissolved in water or buffers. Instead, the oxygen necessary for oxidative phosphorylation is carried in red blood cells on the oxygen binding protein, hemoglobin. This chapter concerns the special properties of hemoglobin that make it particularly suitable for this oxygen carrying function.

## Water Can't Carry Enough Oxygen

Oxygen delivery to tissues in the body is of great importance and must be highly optimized. Just what are these optimizations? Before considering the special and very interesting optimizations, we must consider the overall system. Oxygen is carried on the hemoglobin. The hemoglobin, and not much other protein, is contained in red blood cells that are the major cellular component of blood. The blood is oxygenated as it passes through the lungs, and then blood circulation to the body's various tissues delivers oxygen where it is needed by deoxygenation of the hemoglobin. Before turning to hemoglobin's special oxygen carrying features, which is the major point of this chapter, let us see why hemoglobin is needed.

The oxygen carrying capacity of water is low and only a small fraction of the oxygen in the blood is dissolved directly into the blood plasma itself. Most of the oxygen is bound to hemoglobin for transport to the tissues. Shortly we will return to this topic.

## Human's Adaptations for Oxygen Use and Delivery in Adverse Situations

As species evolve to thrive in their environments, selective pressures will tend to change any limiting traits or characteristics to increase the species' survivability. This will act to maximize

the efficiency of each trait or cellular component. That is, each component or capability will tend to be present in sufficient, but not excessive quantities. Not only is it of value to possess adequate capabilities, but it is also of value not to waste resources by eliminating excess capabilities and capacities that do not contribute to survivability of the species. This tuning of capabilities seems to have occurred in the case of oxygen distribution through the body and in the properties of hemoglobin.

Most peoples live at altitudes relatively close to sea level. When a person who has been living at sea level goes to an altitude of, say 10,000 feet (3000 m), where the air pressure is 70% of that at sea level, they experience significant discomfort and impairment of extreme physical exertion. This means that the oxygen delivery system was well adapted for life at sea level, but not so well adapted for life at 10,000 feet. Not surprisingly then, hemoglobin's affinity for oxygen is eventually adjusted for life at higher elevations by changing the levels of the small molecule 2,3-diphosphoglycerate that binds to hemoglobin and decreases its affinity for oxygen so that it can more easily be released in tissues requiring oxygen. Over a period of a few weeks at an elevated altitude, red blood cell levels increase as well, and a person becomes partially acclimated. Increasing the red blood cell count is not an optimal solution however, as with increased concentrations of red blood cells, the blood becomes more viscous and strains the heart.

What about populations that have spent hundreds of generations at high altitudes in parts of India or the Andes? Interestingly, two different types of solutions have evolved in these two locales. The Sherpas in India have evolved to modify the flows in the body's metabolic pathways to reduce oxygen consumption for a given amount of energy production. The Peruvian Indians however have evolved more efficient blood circulation and oxygen delivery methods. Some of these changes are larger lungs, heart, and blood vessels.

**Why Describe Gas Concentrations in Partial Pressures**

Our discussions below of oxygen binding to hemoglobin will concern concentrations and dissociation constants. A convenient measure of oxygen concentration is partial pressure, which is useful in situations involving mixtures of gasses, air for example, which is primarily a mixture of nitrogen and oxygen.

Why is it that partial pressure is a reasonable proxy for molar concentration even in mixtures of gasses? The reason is that with respect to pressure, and not at extremes of high pressure or very low temperatures, all gasses are identical. This property is implicit in the gas law that figures so prominently in high school chemistry classes, $PV = nRT$, that says that pressure times volume equals the number of moles of the gas involved times the universal gas constant and times the absolute temperature. This law is true for all normal gasses and for mixtures of gasses.

Why is the gas law true? At first it might seem that more massive the gas molecule, the greater would be its contribution to pressure. The reason that mass does not enter the relationship $PV = nRT$ is that the force with which a gas pushes on its container walls is proportional to the average change in momentum of a molecule colliding with the container wall, $\overline{mv}$, and the average number of molecules that collide per unit time, which is also proportional to the molecules' velocity $\bar{v}$, giving the force per unit area on the wall as proportional to $\overline{mv^2}$. By the equipartition theorem which was discussed and derived earlier, the average kinetic energy of a

gas molecule, $\overline{\frac{1}{2}mv^2}$ is $\frac{3}{2}KT$ where $K$ is the Boltzmann constant and $T$ is absolute temperature. Remarkably, as can be seen, this average energy is independent of the mass of the gas molecule, and hence, the universal gas law relates pressure, volume, temperature, and number of gas molecules, but is independent of the mass of a molecule of the gas. Therefore, the contribution to total gas pressure of a mixture of gasses by any component of the gasses is proportional to that component's molar fraction. This contribution to the total pressure is called a partial pressure.

In experiments with oxygen binding to hemoglobin, it is simpler to deal with partial pressures than molar concentrations. As atmospheric pressure is 14.7 lbs./square inch, 760 mmHg, 100,000 pascals, the 20% oxygen composition of air gives the partial pressure of oxygen at sea level as 2.9 lbs./square inch, 150 mmHg, 20,000 pascals, or 0.2 atm.

**Estimating the Oxygen Carrying Capacity of Blood Plasma Compared to Whole Blood**

To gain a better idea of the biological value of hemoglobin, let us compare the amount of oxygen that can be dissolved in a volume of water with the amount of oxygen that can be carried by the hemoglobin in an equal volume of blood. The solubility of oxygen at 20,000 pascals partial pressure (air at atmospheric pressure) in water is about 8 mg/liter. The tetrameric hemoglobin molecule with a molecular weight of 64,500 daltons is present in blood at a concentration of 160,000 mg/liter. Thus, the amount of the 32 dalton $O_2$ that can bind to a liter of blood is (4 x 32/64500) x 160000 mg/liter, or about 320 mg/liter, giving the oxygen carrying capacity of blood as about 40 times that of water. Note that these numbers compare total carrying capacity only. While the oxygen carrying capacity is important, even more important is the ability to deliver the oxygen which involves binding oxygen in the lungs and discharging oxygen in other tissue. It is of no value if the hemoglobin does not release the oxygen where it is needed. This issue will be discussed in the sections below.

The graphs and tables of gas solubility in water show that the solubility of most gasses in water is proportional to the pressure of the gas. Instead, why shouldn't the solubility of oxygen in water obey the Michaelis-Menten binding that we have considered before? It does, it is just that the solubility of oxygen is very low relative to the total number of binding sites present. That is, the concentration of $O_2$is very small compared to the $K_d$, for oxygen "binding" to the water, and consequently, the binding equations

$$H_2O + O_2 \leftrightarrow H_2O \cdot O_2$$

$$\frac{[H_2O \cdot O_2]}{[H_2O]} = \frac{[O_2]}{K_d + [O_2]}$$

when

$$O_2 \ll K_d$$

become

$$\frac{[H_2O \cdot O_2]}{[H_2O]} \approx \frac{[O_2]}{K_d}.$$

This is a linear relationship between oxygen concentration and oxygen “binding”. Formally, each of the quantities in these equations must be of the same dimension, which in this case is molar. Since oxygen partial pressure is proportional to molarity, dimensionally, the equations are fine, even if here we are thinking of the oxygen in terms of pressure.

### Oxygenation in the Lungs and Deoxygenation in Muscle

Oxygen in the alveoli in the lungs rather freely diffuses into the lung capillaries and into the red blood cells where its binding to hemoglobin is near equilibrium. That is, the amount of binding can reasonably be described with an equilibrium dissociation constant. The oxygen partial pressure in the alveoli is about 100 mmHg, and in exercising muscle, about 40 mmHg. Additionally, the binding of oxygen to hemoglobin is pH dependent, and in the lungs, the pH of blood is about 7.6 whereas in exercising muscle it is about 7.3. Figure 10-1 shows the oxygen binding of hemoglobin as a function of $O_2$ partial pressure at three values of pH. It shows that in the lungs, the hemoglobin is nearly fully saturated with oxygen, and in exercising muscle where the oxygen partial pressure is lower, the hemoglobin can discharge about half of its bound oxygen.

Two features of hemoglobin’s oxygen binding curves are of importance, the sigmoidal, or S-shape of the curves, and the strong pH dependence. Both properties increase efficiency of oxygen delivery above what could be achieved if hemoglobin bound and discharged oxygen with Michaelis-Menten kinetics.

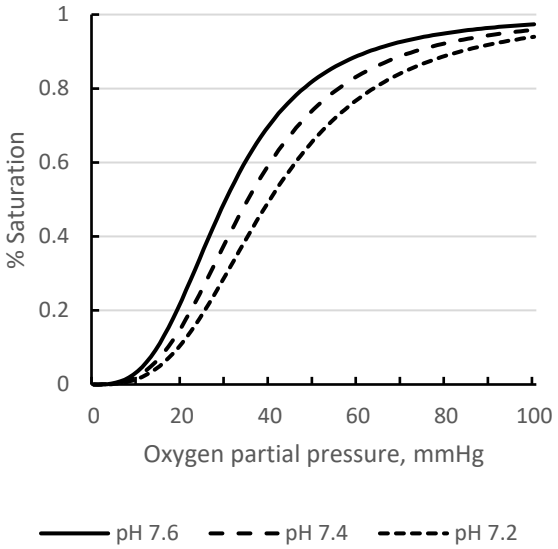



Figure 10-1. Saturation of hemoglobin as a function of oxygen partial pressure.

### Sigmoidal Binding Curves and Positive Cooperativity

Deviations from the hyperbolic shape of a Michaelis-Menten binding curve can only occur when a molecule of the binding protein contains several binding sites in communication with each other. This allows the binding of one ligand molecule to change the binding properties of that protein molecule for the subsequent binding of ligand molecules. Such an effect is called allostery and proteins that display such an effect are described as allosteric. If there were no

communication, each binding site would be independent, and the binding would obey simple Michaelis-Menten kinetics.

In the previous section we saw that a curve of fractional saturation of hemoglobin by oxygen has a sigmoidal or S-shape. This shape means that after the first molecule of oxygen binds, the binding of subsequent molecules is tighter than of the first. The reasoning is as follows: at a particular concentration of oxygen, an increment of oxygen concentration generates an increment in the amount of bound oxygen. That is, the slope of the curve indicates the effective affinity at that oxygen concentration. If at a higher concentration of oxygen, the slope is greater, the effective affinity at the higher oxygen concentration is greater than before. A slope that increases with increasing oxygen concentration is concave upwards, and overall, is sigmoidal.

**Biochemical Description of the Binding to Subunits of a Tetrameric Protein**

As implied above, a sigmoidal binding or saturation curve shows the presence of communication amongst the subunits of an oligomeric protein. Since we will be considering and comparing the binding affinities of individual subunits, it is easier to use microscopic rather than macroscopic dissociation constants. (See the more extensive discussion on this topic in an earlier chapter.) For example, in terms of a microscopic dissociation constant, $K_1$ for the binding of the first molecule of a ligand to a tetrameric protein, the equilibrium equation is written as

$$\frac{4[P] \times [L]}{[P \cdot L]} = K_1$$

where $[P]$ is the concentration of unliganded protein, $[P \cdot L]$ is the concentration of singly ligated tetramer, $[L]$ is the concentration of ligand, and $K_1$ is the dissociation constant for the binding of the first molecule of ligand. The factor of four results from the fact that there are four monomers for each tetramer of the protein. The equation for the binding of the second ligand molecule accounts for the fact that one tetramer of $P \cdot L$ contains three empty subunits to which ligand can bind and that a tetramer of $P \cdot L_2$ contains two monomers with bound ligand from which ligand can dissociate to yield $P \cdot L$. Thus, the equation for binding of the second ligand molecule is

$$\frac{3[P \cdot L] \times [L]}{2[P \cdot L_2]} = K_2.$$

Writing the remaining two equilibrium equations for the binding of the third and fourth molecules of ligand, and using the fact the total amount of $P$, that is $P_t$, equals the sum of free $P$ and the four liganded species of $P$ and then solving for $P$ yields for the fractional saturation of the protein,

$$\text{Fraction of saturation} = \frac{1}{4} \times \frac{4K_2K_3K_4[L] + 12K_3K_4[L]^2 + 12K_4[L]^3 + 4[L]^4}{K_1K_2K_3K_4 + 4K_2K_3K_4[L] + 6K_3K_4[L]^2 + 4K_4[L]^3 + [L]^4}.$$

As it should be, the equation is dimensionally correct, having no dimension, being $(Molar)^o$ on the left, and each term of the numerator and denominator on the right is $(Molar)^4$, making the right also $(Molar)^0$. In the extremes where we know what the result should be, the equation gives the expected results. As $[L]$ approaches zero, saturation approaches zero, and as $[L]$ becomes large, the fraction of saturation approaches one. Finally, if there is no communication

amongst the subunits and each subunit binds independently of the others, the four dissociation constants are all equal. The equation then simplifies to the simple Michaelis-Menten equation:

$$\text{Fraction of saturation} = \frac{[L]}{K + [L]}$$

The equation derived above with its four variables, $K_i$, is complicated and because of this, is not particularly useful. The relationships among the dissociation constants that are necessary to generate a sigmoidal binding saturation curve like that of hemoglobin are not at all apparent. Furthermore, the determination of the four dissociation constants from experimental binding curves would be subject to large errors due to unavoidable but small measurement inaccuracies.

Even worse is the fact that complicated as it is, the equation ignores the fact that a tetramer of hemoglobin consists of two different polypeptide chains, and other tetramers that are not tetrahedral could possess dissociation constants that depend both on binding order and on the positions of bound ligand molecules in the tetramer, *i.e*., ligands bound at opposite corners of a square or bound adjacent to one another.

### Thermodynamic Basis of Positive or Negative Cooperativity

Earlier the relationship was derived between the free energy of binding of a ligand and the tightness of binding as reflected in its equilibrium dissociation constant.

$$K = \mathrm{e}^{-\Delta G/RT}$$

This relationship makes it easier to understand how the tightness of binding of a ligand may be altered by steps in the binding process that either cost or release free energy. Suppose that in the process of ligand binding, the binding site must be opened by breaking some bonds that hold it closed or in an unfavorable shape. The opening will thus require the consumption of some of the free energy of binding. That is, the free energy available for binding will be the free energy of binding minus the free energy cost of opening the binding site. As a result, the ligand will bind less tightly than if its binding did not require breaking some bonds.

In allosteric proteins, the binding of one ligand molecule alters the tightness of binding of a later binding molecule. Suppose that in an homodimeric protein, both binding sites for the ligand require opening, but that when one site has been “opened” by the binding of a ligand, the conformation change induced by this first opening also opens the other binding site. As a result, all the binding energy of the second ligand molecule can go into holding the ligand in the binding site. That is, it will bind more tightly than the first ligand molecule. This is positive cooperativity. Similarly, if the binding of the first ligand molecule induces a conformation change that forces the second ligand molecule to use more of its interaction free energy than the first used, the effect will be negative cooperativity and the binding of the second molecule will be weaker than that of the first.

### How Oxygen Binds to Hemoglobin and What it Does

The human hemoglobin tetramer consists of two alpha chains and two beta chains, each containing a heme group with a coordinated $Fe^{2+}$. A molecule of oxygen can bind to, or more precisely, form a coordination bond with each of the $Fe^{2+}$ ions. Heme itself is a porphyrin ring

containing the $Fe^{2+}$ coordinated with the four nitrogen atoms of the porphyrin. In each subunit of hemoglobin, $Fe^{2+}$ makes a fifth coordination bond to a nitrogen atom of a histidine residue that is referred to as the proximal histidine, Figure 10-2. Because the radius of the five-coordinated iron is somewhat larger than the space available in the center of the porphyrin ring, the center of the iron atom lies somewhat below the plane of the four nitrogen atoms and towards the proximal histidine.

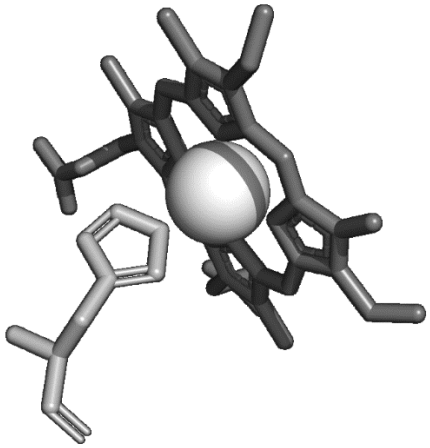


Figure 10-2. The heme group of deoxyhemoglobin. Heme is dark gray, the proximal histidine is light gray, and the iron is shown as a sphere.

When oxygen binds to the heme, it binds to the side opposite from the proximal histidine thereby increasing the coordination number of the iron atom to six. The chemistry of all that happens and why it happens is none too clear, but the changing coordination number changes the iron atom from a high spin state to a low spin state. Such a low spin state gives the iron atom a slightly smaller radius, shifting it from 0.92 Å to 0.77 Å. With its smaller radius, the iron now comfortably fits in the center of the porphyrin ring. Additionally, the porphyrin ring shifts from being slightly domed to being more planar. Altogether, the movement of the iron atom and the changed shape of the ring move the proximal histidine and the F helix of which it is a part, about 0.8 Angstrom. This movement is the beginning of the allosteric conformational change of the hemoglobin tetramer that leads to cooperative binding of the subsequently binding oxygen molecules.

**The Bohr Effect**

Hemoglobin possesses an additional property that further increases the efficiency of oxygen transfer from the lungs to oxygen depleted tissues. This is the Bohr effect which was reported in 1909 by Christian Bohr. (This Bohr, by the way, should be considered as a grandfather of quantum mechanics as it was his son Niels Bohr who first calculated the energy levels of the hydrogen atom and who continued to play an important role in the development of quantum mechanics.) Bohr, the elder, found that lower pH values decrease hemoglobin's affinity for oxygen in a range suitable for increasing the efficiency of oxygen transport to active tissues.

Recall that the endpoint of oxidative phosphorylation is $CO_2$. The $CO_2$ readily diffuses to red blood cells that in addition to hemoglobin, contain the enzyme carbonic anhydrase that converts the $CO_2$ to bicarbonate and an $H^+$ ion. Hence, the pH in active tissues is lower than in the lungs. Curves of oxygen binding as a function of pH were shown in Figure 10-1. In tissues with lower pH, the higher concentration of $H^+$ forces some $H^+$ to bind to the hemoglobin. This tends to shift the conformation of the hemoglobin towards its deoxy conformation where the oxygen is not bound as tightly as it is in the oxy conformation. Thus, the acidity of active muscle stimulates the release of oxygen in that muscle. The converse occurs in the lungs, where hemoglobin gives up the $H^+$ ions that it picked up in muscle. The combination of the sigmoidal shape of the oxygen binding curve and the shift in the curve caused by the pH difference allows about half of the bound oxygen to be released, much more than if the binding curve were not sigmoidal and if the Bohr effect were not present.

**What We Don't Know**

Hemoglobin displays a high degree of cooperativity in its binding of oxygen. Recall the earlier section on allosteric proteins where the three approximation descriptions of binding of a ligand to an oligomeric protein were considered. The simplest of these yields the Hill equation. In this approximation, n molecules of ligand are modeled as simultaneously binding to an oligomer with n subunits, yielding the binding equation

$$\frac{[P \cdot L_n]}{[P_{total}]} = \frac{[L]^n}{K^n + [L]^n}.$$

Experimentally it is often found that the Hill equation provides a rather good approximation of binding. Because no more than n molecules can bind to an oligomer with n subunits, the value of n in the equation must always be less than or equal to the number of subunits of the protein. The Hill coefficient for hemoglobin is found to be between 2.7 and 3.0, which is an amazingly high value that indicates a substantial degree of cooperation amongst the subunits. It is both surprising and disappointing that there is no understanding or at least no universal understanding of exactly what happens in a hemoglobin molecule upon the binding of successive oxygen molecules.

Quite often the approximation models of Monod, Wyman, Changeux, or of Koshland, Némethy, and Filmer have been discussed in considerations of oxygen binding to hemoglobin. As these models contain more adjustable parameters than the Hill model, they can produce curves that more closely fit experimental data. The differences between the qualities of the fits of the curves derived from the models and experimental data is generally too small however to demonstrate that one or the other of the models is a better approximation.

What about just looking at the structure of hemoglobin and reasoning out what happens in hemoglobin at the atomic level? As mentioned earlier, evolution's numerous tweaks have made this approach most difficult. The cause of the obscurity is that with time, evolution has applied nearly every possible little tweak and refinement that improves the system. What is present now is a very finely tuned molecular machine in which many little effects all contribute to the overall behavior. For this reason, it is almost pointless to expect hemoglobin to behave purely in accord with any of the approximation models for oligomers. A modern analog of many small refinements obscuring underlying principles is the contemporary internal combustion engine. A

yrinlook under the hood of a car with a low-polluting combustion engine reveals a complicated maze of tubes, pipes, and wires, the function of very few of which can be deduced by inspection. Fifty years ago, however, before pollution-induced evolution of the auto engine took place, the opposite was the case.

Molecular dynamics which was discussed earlier might seem to be a useful approach to understanding the internal mechanics of cooperativity in hemoglobin. After all, it should simulate a hemoglobin molecule's motions, contortions, and movements in response to the binding of oxygen. One difficulty is that the time scales for allosteric conformation changes in many proteins is thought to be on the millisecond scale, which is much too long to be simulated with current computational power. A second difficulty is that the atom-atom interaction potentials that must be used in molecular dynamics are simply not good enough. For example, the actual behavior of water molecules is exceedingly complicated and cannot be fully described with a small number of parameters. Since a hemoglobin molecule would need to be immersed in tens of thousands of water molecules in a simulation, small errors in water molecule parameters could lead to highly inaccurate simulations.

One question that molecular dynamics can answer however, is how oxygen reaches the iron atom in the heme. There is no channel from hemoglobin's surface to the iron. In molecular dynamics simulations of hemoglobin, oxygen can be seen diffusing through the protein using gaps and temporary holes generated by the vigorous thermal motion of the protein.

**Why Iron and Porphyrin**

If one were designing molecules of life today, we would be reluctant at best to make $Fe^{2+}$ the basis of oxygen transport to tissues. Fe and $Fe^{+2}$ are readily and almost irreversibly oxidized by atmospheric oxygen to $Fe^{+3}$, which is highly insoluble and also does not bind oxygen. Organisms like us must go to extraordinary lengths to take up and use iron ions. We would also shy away from requiring our oxygen-carrying protein to need the services of another complicated molecule like a porphyrin ring. Very likely the reason for our complex system for the uptake and distribution of oxygen results from the fact that iron atoms were much easier to utilize in biochemistry before oxygenation of the atmosphere. Once much of an electron transport system had evolved that made use of iron's versatility, which resulted from its ability to participate in multiple different coordination structures, it was too difficult to evolve other systems not using iron.

**Hemoglobin Gone Wrong, Sickle Cell Anemia**

Hundreds of variant hemoglobins have been identified, but the most important of these is the sickle cell variant. In it, in the beta chain, a glutamic acid residue on the protein's surface has been replaced by a valine. This change from a hydrophilic residue to a hydrophobic residue creates a hydrophobic patch that facilitates aggregation of hemoglobin molecules. When all the hemoglobin in the red blood cell is of the sickle cell variant type, hemoglobin in the deoxy state, but not the oxy state, polymerizes. The polymerization generates two linear chains that form a double helix, and seven of these double helices wrap together forming a cable in which almost all the mutant valines participate in hydrophobic bonding. These cables then form super-cables with the result being long, tough, inflexible fibers that extend across the red blood cells,

distorting many of them into sickle shapes. Sickled cells can clog the capillaries leading to severe pain and widespread tissue damage.

The sickle cell mutation dramatically shortens the life of red blood cells. This gives rise to sickle cell anemia but may also be the reason that the sickle cell trait confers some immunity to the malaria parasite. Possibly the shortened lifetime of the red blood cells interferes with the parasite's overall rate of reproduction.

Linus Pauling demonstrated that the sickling trait resulted from a change in an amino acid of the hemoglobin. This was the first demonstration of the molecular basis for a human disease.

### Additional Activities of Hemoglobin

In addition to carrying oxygen to cells, the tweaking by evolution of the body's oxygen delivery system has given hemoglobin additional capabilities beyond carrying oxygen and $H^+$. Hemoglobin helps carry some of the $CO_2$ that is generated in oxidative phosphorylation back to the lungs. Another molecule that interacts with hemoglobin is 2,3-bisphosphoglycerate. It can bind in the central cavity of a hemoglobin tetramer where it stabilizes the deoxy form. The body adjusts the levels of 2,3-bisphosphoglycerate to fine tune oxygen delivery to such conditions as acclimation to altitude. Finally, unfortunately, carbon monoxide can bind to the iron atoms in place of oxygen. Since it binds far more tightly than oxygen, the oxygen carrying capacity of the hemoglobin is poisoned, and the resulting carbon monoxide poisoning often is lethal.

### Problems

1. Why might the pH of soda pop change after opening its can?
2. Where do the $H^+$ ions go when hemoglobin takes them up in muscle tissue?
3. In addition to the proximal histidine in hemoglobin, another histidine residue near the heme group plays an important role. Where is it, and what does it do?
4. Suppose one had a purely Michaelis-Menten oxygen carrier operating to carry oxygen from the lungs to muscle. In partial pressure, what $K_d$ would give the protein maximum efficiency in terms of moles of oxygen delivered per mole of monomer? What efficiency is it?
5. What is an oximeter and what is its principle of operation?
6. Why is helium sometimes present in the breathing gas mixtures of deep-sea divers?
7. The structures of human oxyhemoglobin and deoxyhemoglobin suggest that functionally, the tetramer functions as a dimer of $\alpha\beta$ dimers. How do we know for certain that some allostery information is passed between the pairs of these dimers?

# Index