December 30, 2024 Draft

Physical Principles in Biology

Biology 3550/3551

Spring 2025

Chapters 1–7

David P. Goldenberg
University of Utah
goldenberg@biology.utah.edu

Contents

Chapter 1

The Scale of Things: Units and Dimensions

Aside from establishing links among the various sciences, a goal of this class is to help strengthen some of the skills that are required in all of the sciences, especially quantitative skills. Working with dimensions and units is one of the most important of these skills.

Historically, one thing that has tended to distinguish biology from the physical sciences is the extent to which mathematics is used. This distinction is diminishing, but there is certainly a strong tradition in biology that is very descriptive. The greatest of all biologists was (arguably) Charles Darwin, who used little or no mathematics.

What is so good about using mathematics in science? Is there anything that Darwin should have used math for? Two major things that math brings to biology and other sciences are that:

- Math provides a way to formalize a description, or "model" a phenomenon.
- Mathematical models have the power of prediction, both to test the theory and make useful predictions. Predictions are at the heart of the current debate about climate change (at least at one level). How good are the predictions?

Interestingly, Darwin's successors, evolutionary biologists, are among current biologists who use math most extensively. For instance, determining the evolutionary relationships among different species using DNA sequence data is a major application.

Most, but not all, applications of math in science involve measurable quantities, such as length, area, volume, mass, time, or concentration. Thus, working with the units of these measurements is one of the most important basic math skills for scientists, and you have, no doubt, had experience with this in many of your classes. None the less, many students continue to find this kind of calculation challenging, and I want to spend some time on this subject before moving on in the class. Even if you are already comfortable with this kind of calculation, you may find that there are some interesting subtleties that you may not have thought about before.

1.1 Measurements as comparisons

Although most measurements are expressed in terms of units, such as meters, grams, liters, *etc.*, mathematics usually deals just in numbers. How do we bridge measurement and numbers? To start, it is useful to consider that nearly all measurements involve comparisons. For instance, we measure length by comparison with some sort of ruler, as illustrated in Fig. 1.1A. Similarly, we measure mass by comparing the gravitational force on an object (its "weight") with the gravitational force of a reference mass (Fig. 1.1B).

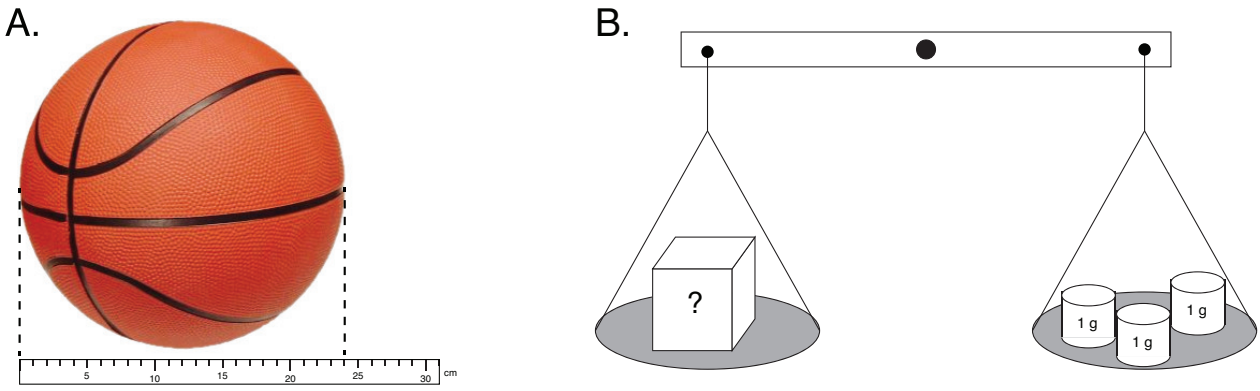


Figure 1.1 Measurement of the diameter of a basketball (A) and the mass of a cube of something (B). In each case the measurement is based on a comparison with some reference object.

The object that we use for comparison then defines the units for the measurement. What is the most natural unit for length? A human body or part of a human body! In the United States, we still use the foot as a unit of length, and the basic unit of length in the metric system, the meter, is on the order of the length of human body.

The somewhat arbitrary nature of unit definitions is illustrated by a famous prank played at the Massachusetts Institute of Technology (MIT) in 1958. A fraternity at MIT decided to use the body of a freshman pledge, Oliver Smoot, to measure the length of the Harvard Bridge, which connects Cambridge, near the site of MIT, to Boston on the other side of the Charles River (and isn't actually close to Harvard University). To do so, they laid Smoot down with his feet at one end of the bridge and his head pointed towards the other. They then made a mark indicating the position of the top of his head, moved his feet to this position and repeated the process until his head reached the other side of the bridge. From this process, the length of the bridge was determined to be 364.4 smoots, ± 1 ear[1] , as commemorated in a plaque shown in Fig. 1.2.

The punch line to this story is that Oliver Smoot went on to a distinguished career in the discipline of *metrology*, the science and technology of measurement. He served as both chairman of the American National Standards Institute (ANSI) and president of the International Organization for Standardization (ISO), the U.S. and international agencies that define the standards of measurement.

Although any unit for length, as an example, can be used as a reference for length at any scale, it is convenient to have units that are appropriate for different ranges. In addition, it is very helpful to have your own "internal rulers" to aid in thinking about the very different scales that we encounter in the sciences, especially when we can't directly experience them on the scales of our body. Fig. 1.3 shows a few examples of biological and manufactured objects on a wide range of length scales. In this figure, lengths are indicated both in meters (at the top) and in units derived from the meter.

[1]The names of units are never capitalized, even when they are derived from a person's name, such as the newton or tesla (a unit of magnetic field strength, not the car). On the other hand, the abbreviations of these names are capitalized, such as N, T or (I presume) S.

Figure 1.2
The plaque commemorating the measurement of the the length of the Harvard Bridge in smoot units.

https://en.wikipedia.org/wiki/Smoot

https://alum.mit.edu/news/AlumniNews/Archive/smoots_legacy

Some typical lengths that are relevant in biology are:

- 1 meter (m) ≈ length of an adult human arm
- 1 millimeter (mm) = 10^{-3} m ≈ length of some of the smallest multicellular animals, *e.g.,* the nematode *C. elegans*. Also about the diameter of a sharp pencil point.
- 1 micrometer (μm or just μ) = 10^{-6} m ≈ length of a bacterial cell.
- 1 nanometer (nm) = 10^{-9} m ≈ radius of a small protein molecule.
- 1 angstrom (Å) = 10^{-10} m = 0.1 nm ≈ length of a covalent chemical bond.

All but the last of the units of listed above use the meter and one of the standard prefixes defined in the metric system. (We will get more specific about what we usually mean by the term "metric system" shortly.) The same prefixes are used for nearly all metric units and, you should be or become fluent in using the ones listed in Table 1.1 on the following page.

Although it may seem an arcane subject, the definition of units and the establishment of standards is of immense practical importance for science, technology and commerce. In the United States, Article 1, Section 8 of the Constitution gives Congress the power (among other things) "To coin Money, regulate the Value thereof, and of foreign Coin, and fix the Standard of Weights and Measures." To fulfil this responsibility, in 1830 Congress established the Office of Standard Weights and Measures, as part of the Department of the Treasury. This office was replaced 1901 by the National Bureau of Standards (NBS). Over time, the NBS took on a broader range of activities, and in 1988 it was replaced by the National Institute of Standards and Technology (NIST) and is now part of the Department of Commerce. Other nations have comparable agencies, and they collaborate to establish international standards through the International Organization for Standardization (ISO)[2]

[2]In English speaking countries it is often believed that ISO stands for International Standards Organization and should be pronounced "eye-ess-oh". But, ISO is defined by the organization as an official abbreviation to be used in all languages and pronounced "iso", which is derived from the Greek word for equal, isos. (I usually forget and say "eye-ess-oh", but it's wrong!)
https://en.wikipedia.org/wiki/International_Organization_for_Standardization

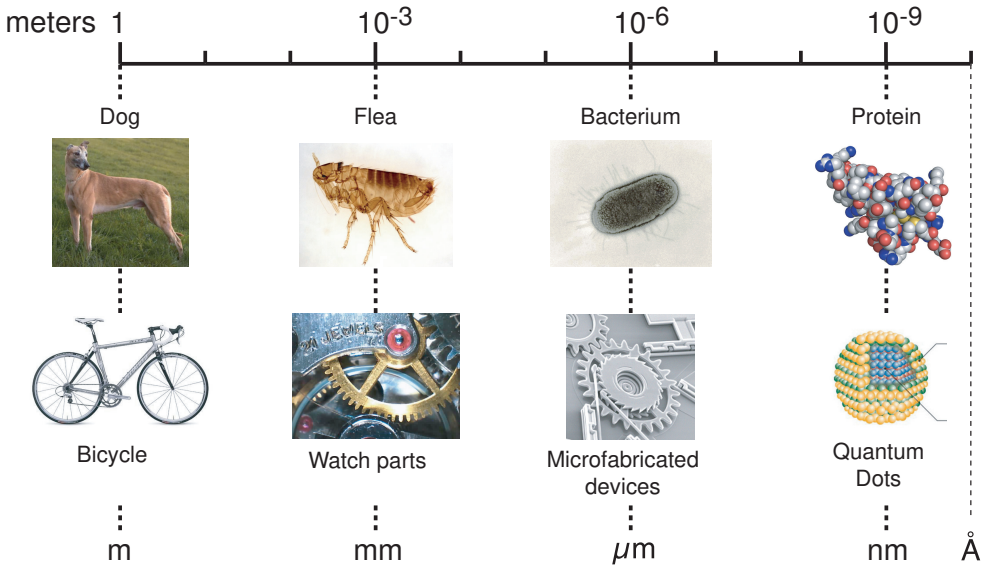


Figure 1.3 Some examples of biological and fabricated objects on a wide range os size scales.

Table 1.1 Standard prefixes for units in the SI metric system. See http://physics.nist.gov/cuu/Units/prefixes.html for prefixes covering the range from 10^{-24} to 10^{24}.

prefix	abbreviation	multiplier	examples
nano	n	10^{-9}	nm, ng
micro	μ	10^{-6}	μm, μg
milli	m	10^{-3}	mm, mg
centi	c	10^{-2}	cm, cg
deci	d	10^{-1}	dm, dg
kilo	k	10^{3}	km, kg
mega	M	10^{6}	Mm, Mg

1.2 Units versus dimensions and a brief history of the metric system

The types of quantities described in the previous sections imply a built in reference object for comparison, such as an object 1 m long. These quantities are called *units* and they are distinguished from another kind of quantity called *dimensions*. A dimension is a quantity like length, mass *etc.* that can be expressed in different, but interchangeable, units. We can compare 1 m and 1 smoot, but we can't compare 1 m and 1 g. (In principle we could have a unit of mass defined as the mass of Oliver Smoot, but this would be very confusing!) Quantities that can be compared directly, such as length, have the same dimension, even if they are given in different units, such as 1 km and 1 mile.

Modern systems of measurement recognize (a minimum of) 5 basic dimensions:

- Length, L
- Mass, M
- Time, T
- Temperature, Θ
- Electric charge, Q, or current, I

Note that the symbols for these quantities are written with italic (or Greek) characters, which follows the mathematical typesetting convention that most variables are represented in this way. The dimensions are usually used in more abstract expressions where specific values are not assigned.

Other dimensions can be derived from the five listed above. Some examples are listed below.

- Area: $A = L^2$
- Volume: $V = L^3$
- Velocity: distance per unit time: $v = L/T$
- Acceleration: change in velocity per unit time: $a = (L/T)/T = L/T^2$
- Force: defined by Newton's second law of motion, $f = m \cdot a$: $f = ML/T^2$

Notice that the dimension of volume is defined in terms of length, as L^3, but the liter (L) is a *unit* of volume. This is a particularly tricky case where it is important to make sure we are talking about a unit or a dimension. Also, the liter is somewhat of an oddity, because its abbreviation is an upper-case letter but is not derived from a person's name.

For most of history, measurements have been made using a mish-mash of units chosen for different purposes, such as cubits, furlongs, feet, miles, *etc.*. This is still true to a degree, but far less so.

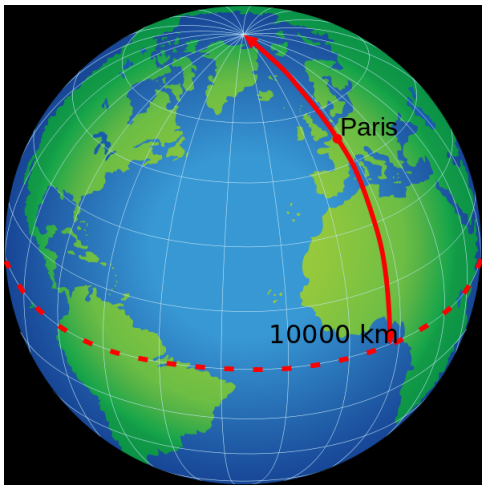


Figure 1.4
The original definition of the meter (1795) as as one ten-millionth (10^{-7}) of the distance from the Equator to the North Pole, along the meridian passing through Paris. Illustration from: https://en.wikipedia.org/wiki/Metric_system

1.2.1 Early metric systems

Although there were earlier precedents, the origins of our current metric system lie in the French Revolution of 1789–1799. Although you might not think that the details of measurement would be an important issue in a political revolution, one of the grievances that led to the revolution was inconsistency among tax collections in different parts of France and with other countries, in part because of the use of numerous different units for measurement of goods. The leaders of the French Revolution also placed a high value on rationality (as they saw it) and wanted a measurement system based on powers of 10. They even went so far as to introduce a decimal calendar and clock. Although the decimal time system didn't catch on, the decimal metric system definitely did, and the United States is now one of just a few countries that use derivatives of what are commonly called "English" units (which, themselves, are not entirely consistent among the countries that use them).

In the 1790s, the French defined two basic (as we would consider them now) units, the *métre* (meter in English) and the *gramme* (gram). Traditionally the definition of any unit has required some standard object that can be used for comparison, and the best standard objects are the most universal ones, so that they are accessible, in principle, to anyone. In this vein, the French chose the Earth itself as the reference object for length, and they defined the meter as one ten-millionth (10^{-7}) of the distance from the Equator to the North Pole, along the meridian passing through Paris, as illustrated in Fig. 1.4. The gram, in turn, was defined as the mass of water in a volume defined by the meter, specifically $1\,\mathrm{cm}^3 = 10^{-6}\,\mathrm{m}^3$.

Although these definitions are clear and based on well-defined physical objects, the definition of the meter, especially, is obviously problematic for practical measurements of, say, the height of a house. Not only is the specified distance along the surface of the Earth absurdly impractical for comparison to a house, the distance was not very well-determined at the time. To address the latter problem, a survey expedition was commissioned to measure a fraction of a meridian that *almost* passes through Paris (from Dunkirk, France to Barcelona, Spain). This turned out to be a rather heroic endeaver that lasted seven years, and the calculations ignored what are now recognized to be significant asymmetries in the shape of the Earth. None the less, the survey produced a defined distance. To create a practical reference for

the meter, a platinum bar, the *mëtre des Archives*, based on the meridian measurement was fabricated and placed for safekeeping in the French National Archives. This reference was then used to create secondary references, which were then used to make additional references, and so on. As it turned out, the platinum bar was about 0.02 m shorter than defined by the meridian, but it remained the standard reference for many decades, probably because everyone was so tired of the whole business by then!

At the same time, a reference object for the gram, a cylinder of platinum with mass equal to that of 1000 cm^3 of water, representing 1000 g=1 kg, was fabricated and placed in the French National Archives. This object was designated the *kilogramme des Archives*. The French also defined units for area and volume based on the meter.

In 1875, an international treaty, the *Convention du Mëtre* was signed by 17 nations and called for replacement of the *mëtre des Archive* and the *kilogramme des Archive* with new standard reference objects, the international prototype meter (IPM) and the international prototype kilogram (IPK). Like the earlier reference objects, these were made of a platinum alloy and were based on the French standards. Importantly, however, the meter and the kilogram were now defined directly in terms of the IPM and IPK, as opposed to the length of a meridian or the mass of a given volume of water (or other substance). This eliminated any question of how closely the reference objects matched the the official definitions of the units. On the other hand, this placed extraordinary importance on the objects themselves, and the *Convention du Mëtre* and the organizations it created, established detailed protocols for maintaining the standards and replicating them. Copies of the IPM and IPK were made and distributed to all of the treaty signatories, which then used these as references for their own countries. The *Convention du Mëtre* also established an international organization to continue work on refining measurement standards and to organize conferences for this purpose, the General Conference on Weights and Measures (Conférence générale des poids et mesures, CGPM). Since then, there have been 26 CGPM meetings.

The IPM remained the definition of the meter until 1960, when the meter was redefined in terms of the wavelength of light corresponding to a specific electronic transition in krypton 85 atoms. This represented the long-sought goal of a standard that was independent of a single object and was, in principle, accessible anywhere. However, the IPK continued to serve as the international standard for mass until 2018, when a major revision of the SI was adopted, as discussed further below.

During the century following the establishment of the French metric system (and related ones in other countries), some of the major advancements in the physical sciences were in the fields of thermodynamics, electricity and (closely related to electricity) magnetism. These fields required the the development of entirely new classes of dimensions and units. It took much of the nineteenth century to define the basic quantities for these disciplines, and even then there were two basic approaches for relating electricity and magnetism to force as defined by Newton's second law[3]. In 1873, a committee of the British Association for the Advancement of Science proposed a system of units that unified units for electricity and magnetism with previously defined units for length, mass, time and temperature. This system came to be known as the centimeter-gram-second (cgs) system, which was widely

[3]The two approaches differed by whether force was related to electricity in terms of the electrostatic interaction between two charges or the magnetic interaction between currents flowing through two wires.

adopted and was the official "metric system" for several decades. Even with the introduction of the cgs system, however, there were internal inconsistencies involving electrostatic and magnetic forces. As a consequence, there were actually two branches of the cgs system (each with further variants). This can still be a source of confusion, especially when reading older publications and trying to convert values to the modern conventions.

1.2.2 Establishment of the Modern Metric System, the Système International d'unitès (SI) and Further Revisions

The next major revision to the metric system was enacted in 1960, when the name Système International d'unitès (SI) was introduced at the 11^{th} CGPM.. Among other things, the SI finally settled on a single, consistent way of dealing with electricity and magnetism. The SI also replaced the centimeter and gram as the basic units of length and mass, respectively, with the meter and kilogram. As a consequence, the SI is sometimes referred to as the MKS (meter, kilogram, second) system, but this is really a more general designation that includes some predecessors to the SI.

Rapid developments in physics and electronics (especially the invention of the laser) led to another redefinition of the meter in 1983. The meter is now defined as the distance travelled by light in a vacuum during 1/299792458 of a second.

This redefinition of the meter was more profound than earlier changes in units, because it depended on establishing an essentially arbitrary definition of a fundamental constant of the universe, the speed of light, as opposed to a measurement of the constant using defined units. At the same time that the meter was redefined, the speed of light was declared to be exactly 299792458 m/s. The second had previously been defined, in 1967, as the time of 9,192,631,770 cycles of the radiation associated with a specific quantum transition[4] of cesium 133 (^{133}Cs) atoms (The frequency of this radiation is designated $\Delta\nu_{\text{Cs}}$). Like the speed of light, the second was defined by setting the value of a physical constant, in this case $\Delta\nu_{\text{Cs}}$. Although the numbers associated with these definitions may seem rather arbitrary and not very convenient, they were chosen to give the best possible match to the original standards.

The new definition of the meter represented a change from an *explicit-unit* to an *explicit-constant* basis for defining units. Though this indirect approach is admittedly rather awkward, it allows the continuous refinement of numerical values of units by making more precise physical measurements of the constants. For instance, if the speed of light were to be more precisely measured, the official value for this constant will remain exactly 299792458 m/s, but the standard for the meter or second would be adjusted to reflect the improved measurement. Furthermore, these measurements can, in principle, be made by anyone in any location. For instance, it would be possible to for an extraterrestrial civilization to implement our definition of a meter, provided only that they know our definitions and can measure the speed of light and the frequency of the ^{133}Cs transition used to define the second.

In contrast to the definition of the meter, the IPK proved to be remarkably difficult to replace as the standard for mass, which was finally accomplished in 2018. In 2007 the CGPM called for a complete shift in the definition units to an explicit-constant basis. The 2011 CGPM further decided to base the definition of the kilogram on a fixed value of the Planck

[4]The ground-sate hyperfine transition

constant, h, which defines the relationship between the energy of a quantum transition, E, and the frequency, ν, of the electromagnetic radiation absorbed or released during that transition

$$E = h\nu$$

As discussed below, the units of energy are expressed in SI units as $\mathrm{kg \cdot m/s^2}$, so that the units of the Planck constant are given by:

$$h = \frac{E}{\nu} = \frac{\mathrm{kg \cdot m/s^2}}{\mathrm{s^{-1}}} = \mathrm{kg \cdot m/s}$$

Once the Planck constant is given a fixed value (along with the defined values for the meter and second), the kilogram can be defined as:

$$\mathrm{kg} = \frac{h \cdot \mathrm{s}}{\mathrm{m}}$$

The 2011 CPMG defined the value of the Planck constant to be exactly $6.62606 \times 10^{-34}\,\mathrm{kg \cdot m^2 s^{-1}}$. However, the kilogram was not redefined at this time, because there was not yet a method deemed sufficiently precise for implementing the new definition, which requires being able to actually make measurements of mass that are directly related to the Planck constant. The first step in that process was to establish very precise measurements of the Planck constant in terms of the existing definition of the kilogram. The criteria established in 2011 specified that the Planck constant be determined by three independent experiments, using two independent methods, with uncertainties less than 50 parts per billion (5×10^{-8}). At that time, there were two methods available for measuring the Planck constant with such precision, both of which are somewhat round-about.

The first method is based on the value of Avogadro's number (also called the Avogadro constant, N_A), the number of atoms, molecules or ions in a mole of a substance. Avogadro's number is directly related to the Planck constant, through other physical constants that have been measured with extremely high precision (less than one part per billion). Thus, a measurement of Avogadro's number is equivalent to a measurement of the Planck constant. Prior to 2019, Avogadro's number was defined as the number of atoms in 12 g of pure carbon 12 (^{12}C). But, determining an actual value for N_A directly from that definition is highly problematic, because carbon readily undergoes a variety of chemical reactions and even pure carbon can exist in multiple forms. Instead silicon, which forms very stable and well defined crystals were used.

The method used to determine Avogadro's number is referred to as an x-ray crystal density (XRCD) measurement and used crystals of silicon highly enriched (99.995%) in ^{28}Si. Implementation of the XRCD method involved machining two spheres of crystalline silicone ^{28}Si (to provide two of the three independent measurements called for). The sphere is the three-dimensional shape that can be most precisely manufactured (using a lathe), and its size is defined by a single measurable length, the diameter. The spacing between atoms in the crystal can be determined with very high precision by x-ray crystallography, so that the number of atoms in the sphere can be determined with very low uncertainty. However, to correlate the number of atoms with the mass of the crystal, it was also necessary to determine

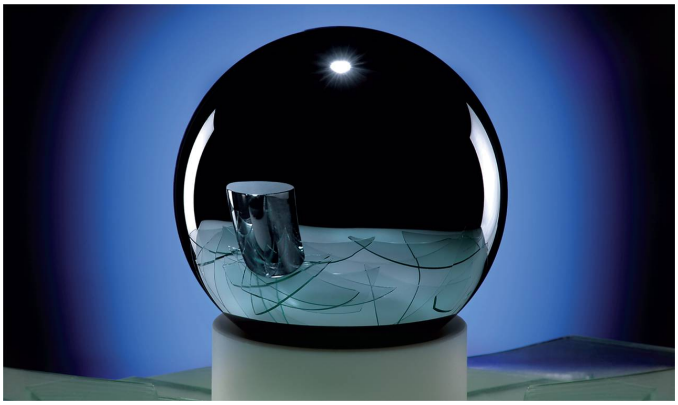


Figure 1.5
One of two silicon spheres, 99.995% enriched in ^{28}Si, used to establish Avogadro's number. A reflection of an old kilogram standard can be seen in the surface of the sphere. Illustration from: Cho, A. (2018). World poised to adopt new metric units. *Science*, 362, 625–626. http://doi.org/10.1126/science.362.6415.625 Photograph from the Physikalisch-Technische Bundesansttalt (PTB), Germany.

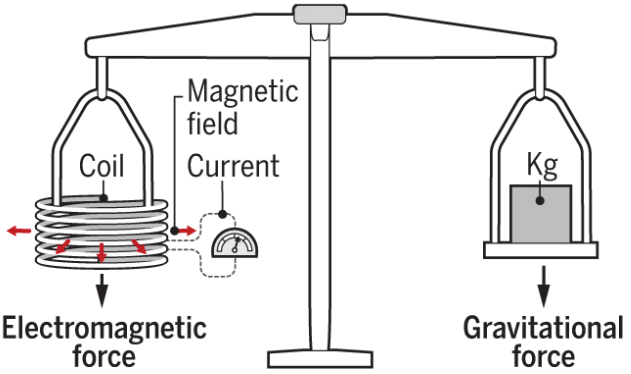


Figure 1.6
Schematic diagram of a Kibble (or moving-coil watt) balance. The gravitational force acting on the mass to be determined (on the left) is balanced by the magnetic force between a stationary magnetic field and the field generated by electric current passing through the coil attached to the right-hand side of the balance. The electric current required to balance the two forces is measured and used to calculate the mass of the object. Adapted from:
Cho, A. (2017). Plot to redefine the kilogram nears climax. *Science*, 356, 670–671. http://dx.doi.org/10.1126/science.356.6339.670

precisely the isotopic composition of the silicon, which was done by mass spectrometry. Over a period of several years, measurements of the dimensions and isotopic composition of these spheres were refined, until the reproducibility could be established to be 1 part in 50 million. Among other precautions, this required extensive polishing to remove virtually all traces of contamination from the surface. These spheres have been described as "the world's roundest objects", and a photograph of one of them is shown in Fig. 1.5. From the measurements of the silicon spheres, the value of Avogadro's number was defined, at the 2018 CGPM, to be exactly 6.02214076×10^{23}.

The second method for measuring the Planck constant uses a special electronic balance called a *Kibble balance*, illustrated in Fig. 1.6. In this device, the force of gravity acting on the object to be weighed is balanced by the magnetic force generated by a coil of wire in a magnetic field, and the mass is determined by the electric power, expressed in watts, required to balance the two forces. Although this basic idea for a balance is not new (and is commonly incorporated in electronic laboratory balances), an important refinement was introduced in 1975 by Bryan Kibble, who devised a method to internally calibrate the coil and magnetic field. Kibble died in 2016, just before his invention was expected to enable a new definition of the kilogram, and the device (previously described as a moving-coil watt balance) was named in his honor.

The power required to balance the mass represents the product of the electrical current, measured in amperes (A) passing through the coil and the electrical potential across the coil, measured in volts (V). The current and potential can be directly related to the Planck constant through measurements of two quantum-mechanical phenomena, the quantum Hall effect and the Josephson effect. Using a kilogram standard (as then defined), it was possible to precisely measure the Planck constant in a way that is independent of the XRCD analysis of the ^{28}Si spheres.

By 2018, the criteria established in 2011 for measuring the Planck constant, with an uncertainty of less than 50 parts per billion, had been met by multiple experiments using both XRCD method for determining Avogadro's number and the Kibble balance. This milestone, then enabled the 2018 CGPM to finally redefine the kilogram in terms of the Planck constant, the second (defined by $\Delta\nu_{\mathrm{Cs}}$) and the meter (defined by the second and the speed of light). At the same time, Avogadro's number was redefined to be exactly 6.02214076×10^{23}.

In principle, internally calibrated Kibble balances can now be used to independently measure masses in laboratories, or more practically, Kibble balances can be used to measure reference masses that can be used as secondary standards. Alternatively, spheres of ^{28}Si characterized by XRCD could be used a mass references. At present, however, both of these techniques are very challenging in practice, and only a few nations have the resources to implement them. The XRCD method is particularly difficult to implement, and for the near future the Kibble balance is expected to be used to calibrate secondary standards, including the IPK, which will be used to calibrate additional reference masses. As a consequence, the redefinition of the kilogram will have practical consequences at only the highest level of metrological standardization. It should also be noted that the Kibble balance is actually a bit less precise (about 20 μg/kg) than the best conventional balances (with precisions as low as (1–2 μg/kg), so that reference objects calibrated with the new method are expected to be slightly more variable than before. To place this in perspective, however, the uncertainties from the Kibble balance are on the order of one millionth of one percent.

1.2.3 The base dimensions of the SI and their current definitions

The SI defines seven "basic" dimensions and their standard units, as summarized in Table 1.2. Of the seven basic units, the first five in the table are independent of one another. However, the amount of a substance (mole) and luminous intensity (*candela*) can be defined in terms of the the other five basic dimensions and are not strictly necessary for a complete set of units. They are included in the SI largely as a matter of convenience and consistency with older definitions.

The *luminous intensity* and its SI unit, the candela, may be the dimension and unit that are least familiar. As suggested by its name, luminous intensity is a measure of the brightness of visible light sources and, specifically the intensity along a specific direction. The origin of the candela, and its name, is reflected by the fact that the luminous intensity of a conventional wax candle is approximately 1 candela. (The Latin for *candle* is *candela*.) Until 1948, various countries had different defined units of luminous intensity defined in terms of very specific light sources, including in some cases a wax candle of defined size and composition. The dimension of luminous intensity (and the several other quantities related

Table 1.2 The seven basic dimensions and units defined in the SI. For details, see http://physics.nist.gov/cuu/Units/units.html

Dimension	SI base Unit	Abbreviation
length	meter	m
mass	kilogram	kg
time	second	s
electric current	ampere	A
thermodynamics temperature	kelvin	K
amount of substance	mole	mol
luminous intensity	cendela	c

to light intensity) are also unusual because they were originally used to express a human response to a stimulus, rather than a specific amount of light, as measured for instance by the number of photons of a specified energy.

By 2019, all of the SI basic units were defined by the values of seven physical constants, which are now set to exact values, as listed in Table 1.3. There is not a simple one-to-one relationship between the basic SI units and the constants listed in Table 1.3, as several of the units are defined in a somewhat hierarchical way, as listed in Table 1.4.

From the seven basic units, there are an almost limitless number of derived units that can be used to specify any measured quantity that has so far been conceived. Some examples of derived SI units are listed in Table 1.5.

Although the the basic units in the SI now well established as the foundation for measurement throughout the world, the system continues to be updated to incorporate new technologies and new applications. Going forward, this will involve making decisions about when to update the numerical values used to define the basic units, while keeping the the seven physical constants fixed to their standardized values.

Table 1.3 The seven physical constants used to define the SI basic units, and the their exact, defined values. For details, see http://physics.nist.gov/cuu/Units/units.html

Constant	Symbol	Exact value
The ground-state hyperfine transition frequency of ^{133}Cs	$\Delta\nu_{\mathrm{Cs}}$	$9.192631770 \times 10^{9}\,\mathrm{Hz}$
The speed of light in vacuum	c	$2.99792458 \times 10^{8}\,\mathrm{m/s}$
The Planck constant	h	$6.62607015 \times 10^{-34}\,\mathrm{J \cdot s}$
The elementary charge	e	$1.602176634 \times 10^{-19}\,\mathrm{coulomb}$
The Boltzmann constant	k	$1.380649 \times 10^{-23}\,\mathrm{J/K}$
The Avogadro constant	N_{A}	$6.02214076 \times 10^{23}\,\mathrm{mol}^{-1}$
The luminous efficacy of monochromatic radiation of frequency $540 \times 10^{12}\,\mathrm{Hz}$	K_{cd}	$683\,\mathrm{lm/W}$

Table 1.4 Definitions of the basic SI units in terms of the seven physical constants listed in Table 1.3. For details, see http://physics.nist.gov/cuu/Units/units.html

Unit	Defined by:
second (s)	Setting $\Delta\nu_{\rm Cs}$ to be 9.192631770×10^{9} in units of $\rm s^{-1}$.
meter (m)	Setting the speed of light (c) to be 2.99792458×10^{8} in units of m/s, with the second defined in terms of $\Delta\nu_{\rm Cs}$.
kilogram (kg)	Setting the value of the Planck constant to be 6.62607015×10^{-34} in units of $\rm kg\cdot m/s$, with the meter and second defined in terms of $\Delta\nu_{\rm Cs}$ and c.
ampere (A)	The relationship $1\,\rm A = 1\,coulomb/s$, with the coulomb defined so that the elementary charge (e) has the exact value $1.602176634\times10^{-19}$ coulomb, and the second is defined in terms of $\Delta\nu_{\rm Cs}$.
kelvin (K)	Setting Boltzmann's constant (k) to be 1.380649×10^{-23} when expressed in units of $\rm kg\cdot m^2s^{-2}K^{-1}$, with the kilogram, meter and second defined in terms of $\Delta\nu_{\rm Cs}$, c and h.
mole (mol)	Making 1 mol equal to Avogadro's number of elementary entities.
candela (cd)	Setting the value of $K_{\rm cd}$ to be 683 when expressed in units of lm/W, which is equivalent to units of $\rm cd\cdot m^{-2}kg^{-1}s^2$.

Table 1.5 Examples of derived SI units.

Dimension	SI unit
Area	$\rm m^2$
Volume	$\rm m^3$
Acceleration	$\rm m/s^2$
Force	newton (N) $= \rm kg\cdot m/s^2$
Energy	joule (J) $= \rm N\cdot m = kg\cdot m^2/s^2$
Electric charge	coulomb (C) $= \rm A\cdot s$

1.2.4 Other Units

The non-metric units still used in the United States for consumer products and some other purposes are sometimes referred to informally as "English units", but this term, like "metric system", actually covers several historic and closely related systems. The units used in the United States are more properly designated the "United States customary units". The other major system of English units is the Imperial system, which was officially established in 1824, and still has limited use in the United Kingdom, Canada and a few other British Commonwealth nations. The United States customary and Imperial systems are largely similar, but with some distinctions.

In order to simplify conversions with the SI, while not deviating too noticeably from the traditional definition of length units, the U.S. Customary yard is defined directly in terms of the meter, as exactly 0.9144 m, which makes the inch exactly 0.0254 m = 25.4 mm. Similarly, the U.S. customary pound is defined as exactly 453.59237 g,[5] and the US gallon (for liquid measure) is defined as exactly 3.785411784 L. There are also more informally defined "English" units of fluid and dry volume, as typically found in food recipes. These are not so formally defined, and probably don't need to be.

1.3 Using units in calculations

For students and practitioners of science, the important point about units and dimensions is that their proper use is a critical skill! Although students in this course should have had lots of experience in this already, I often find that many are rather rusty.

The simplest problems involving dimensions often have the form, "How many feet are there in a kilometer?", and one can find tables that provide instructions such as, "To covert kilometers to miles, multiply by 0.621371." Instructions like this one are often called *conversion factors* and there are many published tables and websites with conversion factors for different units. Most of these are probably correct, though care is sometimes required, especially when the same word is used for different measurements; "ounces" is a particularly confusing one. One convient and quite comprehensive website for conversions is:

http://www.digitaldutch.com/unitconverter/

But, I can provide no guarantee for the reliability of the information on this site! One can also use Google and other web-search engines to quickly look up conversions, with queries such as:

Miles to feet

More authoritative references for conversion factors are provided at the end of this chapter.

The use of units in calculations has a fancy name, *Dimensional Analysis*. The basic idea of dimensional analysis is to treat units as part of the algebraic terms representing quantities.

[5] More specifically, this defines the pound avoirdupois (derived from a French phrase for "weights and measures") in the U.S. Customary system. There are actually two groups of units for mass in this system, the other being the Troy system, which is still sometimes used for precious metals, and even more are found in other English systems.

This idea can be developed quite formally, but for practical purposes this is not necessary. We can think of conversion factors as recipes, such as "To convert kg to g, multiply by 1,000.". But we can also write them as equations, such as

$$1\,\mathrm{kg} = 1000\,\mathrm{g}$$

We can re-write this equation as

$$\frac{1\,\mathrm{kg}}{1000\,\mathrm{g}} = 1$$

or:

$$\frac{1000\,\mathrm{g}}{1\,\mathrm{kg}} = 1$$

Remember that any number multiplied by 1 (or divided by 1) is itself. This is expressed more formally by the statement, "Multiplication by 1 is an identity operator." So, if we want to convert 37 kg to g, we can write this algebraically as:

$$37\,\cancel{\mathrm{kg}} \times \frac{1000\,\mathrm{g}}{1\,\cancel{\mathrm{kg}}} = 37000\,\mathrm{g}$$

This isn't very exciting, but the important point is that the units are treated just as any other algebraic term, like a variable x or a, would be. We know that we did things properly because the kg units cancel out properly to give us an answer in g. If we had divided instead of multiplied we would get:

$$37\,\mathrm{kg} \times \frac{1\,\mathrm{kg}}{1000\,\mathrm{g}} = 0.037\,\mathrm{kg}^2/\mathrm{g}$$

Mathematically, this is correct, but the answer we get doesn't make sense in terms of the units we are looking for.

As a potentially more interesting example, consider the scale and dimension of a bacterial cell. One of the major bacterial species in our gut is *Escherichia coli*, and the cells of this species can be approximately described as cylinders $2\,\mu$m long and $1\,\mu$m in diameter, as illustrated in Fig. 1.7.

Expressing the relationship in terms of dimensions (as opposed to specific units), the volume of a cylinder is calculated as:

$$V = L \times A$$

where L is the length of the cylinder and A is the area of the "caps" at each end. From the equation for area of a circle,

$$V = L\pi R^2$$

where R is the radius of the cylinder. For our bacterium, L=2 μm and, R=0.5 μm, so that we can replace the dimensions with values with specific units:

$$\begin{aligned} V &= 2\,\mu\mathrm{m} \cdot \pi \cdot (0.5\,\mu\mathrm{m})^2 \\ &= \pi \cdot 0.5\,\mu\mathrm{m}^3 \\ &\approx 1.57\,\mu\mathrm{m}^3 \end{aligned}$$

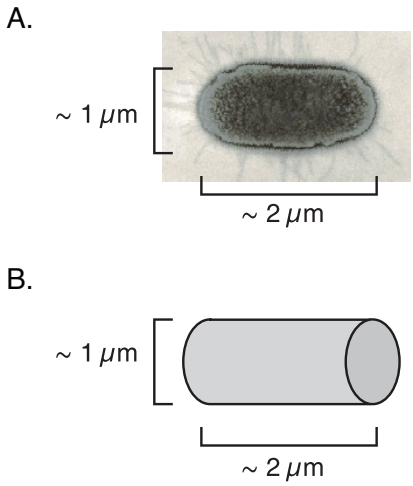


Figure 1.7
A. Light-microscope image of an E. coli bacterium, from http://eduspace.free.fr/ibbiology2007_14/02_cells/e_coli.html.
B. Approximation of the bacterial cell as a cylinder.

So the answer is $\approx 1.6\,\mu\text{m}^3$, but cubic micrometers are not units of volume that are very easy to relate to!

A more conventional unit of volume is a liter. So, how do we get from μm^3 to L? An easy to remember conversion factor for volume is based on the cubic centimeter, or "cc", which is equal to 1 mL. We can use this to derive a conversion factor from mL to m^3.

$$
\begin{aligned}
1\,\text{cm} &= 0.01\,\text{m} \\
(1\,\text{cm})^3 &= (0.01\,\text{m})^3 \\
1\,\text{cm}^3 &= 10^{-6}\,\text{m}^3 \\
1\,\text{mL} &= 10^{-6}\,\text{m}^3
\end{aligned}
$$

Notice that we start with a relationship between two units of linear distance and raise both sides of the equation to the third power to obtain a relationship between units of volume. Importantly, the entire expression on each side of the second line is raised to the third power, not just the units. This follows the standard rules of algebra, and raising only the units to the third power will lead to an incorrect result. We can then manipulate this result further to obtain the relationship between 1 L and m^3.

$$
\begin{aligned}
1\,\text{mL} \times 1000 &= 10^{-6}\,\text{m}^3 \times 1000 \\
1000\,\text{mL} &= 10^{-3}\,\text{m}^3 \\
1000\,\cancel{\text{mL}} \times \frac{1\,\text{L}}{1000\,\cancel{\text{mL}}} &= 10^{-3}\,\text{m}^3 \\
1\,\text{L} &= 10^{-3}\,\text{m}^3
\end{aligned}
$$

This conversion factor is one that is worth committing to memory, as we will frequently want to relate volumes to lengths on a range of scales.

For the case of the bacterial cell, we can use this factor to express the volume in liters:

$$1.6\,\mu\text{m}^3 \times \left(\frac{10^{-6}\,\text{m}}{1\,\mu\text{m}}\right)^3 = 1.6 \times 10^{-18}\,\text{m}^3$$

$$1.6 \times 10^{-18}\,\text{m}^3 \times \frac{1\,\text{L}}{10^{-3}\,\text{m}^3} = 1.6 \times 10^{-15}\,\text{L}$$

A typical laboratory culture of *E. coli* contains about 10^9 (1 billion) bacteria per mL. The total volume of these bacteria is:

$$10^9\,\text{bacteria} \times 1.6 \times 10^{-15}\,\text{L/bacterium} = 1.6 \times 10^{-6}\,\text{L}$$
$$1.6 \times 10^{-6}\,\text{L} \times 10^3\,\text{mL/L} = 1.6 \times 10^{-3}\,\text{mL}$$

So, about 0.2% of the culture volume is occupied by bacteria.

1.4 Units of Concentration

The concept of concentration is central to chemistry and is critical to much of biology and physics. The particular issues that arise in dealing with units of concentration are thus deserving of some review here. There are a variety of different ways that concentrations can be expressed, but we will focus on the ones that are most common and appropriate for the topics we will cover in this course.

1.4.1 Different ways of expressing concentration

For most purposes, the most convenient units of concentration are those that express the amount of solute present in a given total volume of solution. This amount of solute might be expressed as a mass or as the number of moles, to give units, for instance, of g/L or mol/L (M). In practical terms, these definitions mean that we would make, for instance, a 10 g/L solution by measuring 10 g of the solute and dissolve it in somewhat less than 1 L of solvent and then add more solvent to make up a final volume of 1 L. At first glance, this seems like a rather awkward definition and procedure: It would be easier to make a solution by dissolving 10 g of solute in 1 L of solvent. However, for calculations, it is much easier to work with concentrations defined in terms of the total volume, rather than the amount of solvent used. This is because it is straight forward to calculate the amount of solute in a given volume of solution, or conversely the volume that would contain a given amount of solute. For instance, if we have 50 mL of a 50 g/L solution of glucose in water, we can readily calculate the number of grams in the solution as follows:

$$50\,\cancel{\text{mL}} \times \frac{1\,\text{L}}{1000\,\cancel{\text{mL}}} = 0.05\,\text{L}$$

$$50\,\text{g}/\cancel{\text{L}} \times 0.05\,\cancel{\text{L}} = 2.5\,\text{g}$$

Consider, on the other hand calculating the number of grams in a solution made up by dissolving 50 g of glucose in 1 L of water. We would expect that the total volume of this solution

would be greater than 1 L, but knowing how much more requires additional information, and this is not a simple calculation! When compounds, even as liquids, are mixed together in a solution their volumes do not necessarily add together in a simple way. The balance of interactions among the different kinds of molecules can bring some pairs closer together and lead to repulsions between others. As a consequence, the final volume can be either smaller or greater than the sum of two volumes of different liquids. When a solid is dissolved in a liquid, the problem is even more complicated. Furthermore, water, the solvent most relevant to biology, is a particularly complicated liquid, a point that we will return to later in the course. So, if we were to make up a solution by mixing a given amount of solute with a given amount of solvent, we would probably have to measure the final volume in order to relate volume to the amount of solvent. (For a few, particularly well-characterized solute-solvent pairs, very precise measurements have been made that can be used to predict volumes and densities of solutions.)

There are some special applications for which it is advantageous to use units of concentrations based on the amounts of solute and solvent, rather than total volume. In particular, molal units are used often in chemical thermodynamics. A 1 molal solution is prepared by dissolving 1 mole of solute in 1 kg of solvent. The advantage of a concentration defined this way is that it does not change with temperature or pressure; the masses of solute and solvent remain the same. In contrast, a change in temperature or pressure can change the total volume of a solution and, therefore the molar concentration. The behaviors of the two solutions aren't really different, it's just that the definition of a molar concentration depends on volume, and the molal concentration doesn't. As in many things, it is a matter of what is most important to keep track of for a particular purpose.

Now, it should also be noted that the practical difference between solutions defined by the amount of solvent versus a given total volume depends greatly on just how concentrated the solutions are. If, for instance, the solute makes up less than 1% of the total volume of the solution, just mixing the solute with, say, 1 L of solvent may not introduce a significant error for many purposes. In biochemistry labs, solutions are often made up to quite small volumes (usually because the reagents are very expensive), and this kind of error is frequently considered acceptable. On the other hand, the solute may make up as much as half of the total volume of some solutions, and the difference between, say, a 5 M solution and a 5 molal solution is very significant, indeed.

In some situations, solution concentrations are expressed as percentages, and here there is also an important distinction. Percentage solutions can be specified as either mass per volume or volume per volume. A percent mass/volume,%(m/v), concentration, sometimes identified as %(w/v), is defined as the number of grams of solute dissolved in enough solute to make 100 mL total volume. As with molar concentrations, it is easy to calculate the amount of solute in a given volume of solution for a %(m/v) solution. Percent volume/volume, %(v/v), concentrations are usually used for solutions made by mixing volumes and are defined by the number of mL of one liquid mixed with a second liquid to give a final volume of 100 mL. Depending on the densities of the liquids, there may be a significant difference between the percent concentrations expressed as m/v and v/v for the same solution.

1.4.2 Units of atomic and molecular mass

A variety of terms are used to describe the masses of atoms, ions and molecules, and there is a bit of confusion and inconsistency in their definitions and use. At first glance, it might seem that atomic and molecular masses should be expressed in the usual SI unit of mass, the kilogram. But, the mass in kilograms of a single atom or molecule is a very small and awkward number for most uses. So, instead we have a special unit, which goes by different names in different contexts. Both of the following refer to the same unit:

- Unified atomic mass unit, abbreviated as u or amu
- dalton, abbreviated as Da

Niether of these equivalent units is defined in the SI, but are defined by the International Union of Pure and Applied Chemistry (IUPAC). The term amu is widely used in the field of mass spectrometry (where molecular masses are sometimes measured with precision of a fraction of an amu), whereas the dalton is more widely used in the molecular biosciences.

These two equivalent units are defined as the mass of an atom, ion or molecule divided by the mass of an atom of carbon 12 (^{12}C) divided by 12. By this definition, then, an atom of ^{12}C has a mass of exactly 12 Da. 1 amu, or 1 da is approximately equal to 1.66054×10^{-27} kg.

Although we tend to think of the atomic mass of an atom as being an integer, equal to the total number of protons and neutrons in the nucleus, the masses of the protons and neutrons do not add exactly, because of the presence of other subatomic particles. So, the masses of atoms other than ^{12}C generally differ slightly from an integer. In addition, when a sample of an atomic or molecular species is considered, there are usually more than a single isotope present, with different masses, so that the average atomic or molecular mass of the species usually deviates significantly from an integer. The molecular masses that are usually cited in articles, books and on the labels on bottles of chemicals are based on the average ratios of the various isotopes found in nature (on our planet). But, these ratios can differ slightly for natural reasons and can be altered greatly by artificial enrichment.

Because the amu and dalton are defined as the ratio of two masses, they are really units without dimensions, and is common and legitimate to represent atomic and molecular masses as pure numbers without units. To emphasize the relative nature of atomic and molecular masses, IUPAC defines the terms *relative atomic mass* and *relative molecular mass*, with the symbols A_r and M_r, respectively. The terms *atomic weight* and *molecular weight* are also deeply ingrained in many of us, but aren't strictly correct, since weight is a measure of (gravitational) force, rather than mass.

Two other commonly used terms are *molar mass* and *relative molar mass*, both of which refer to the mass of one mole of a substance[6]. Logically, molar mass would simply have the units of g or kg, but it is usually expressed as g/mol, which is a bit redundant. The

[6]The definition of the molar mass has been muddied somewhat by the redefinition of the mole in 2019, but not in a way that is likely any to have any practical differences. Prior to 2019, the mole was defined as the number of ^{12}C atoms in exactly 12 g. This meant that, by definition, the molar mass of ^{12}C was exactly 12 g/mol, and the molar masses of other entities were similarly linked exactly to their atomic or molecular masses. With the mole now redefined as exactly $6.02214076 \times 10^{23}$ particles, the direct connection to the atomic mass and molar mass of ^{12}C is broken. But, any discrepancy is on the order of parts per billion!

relative molar mass is defined by IUPAC as the molar mass divided by 1 g/mol, making it dimensionless and equal to the relative molecular mass, M_r.

So, we have the following terms (for molecules) which all have the same numerical values and convey the same information:

- Molecular mass (or weight), with units of daltons (da) or unified atomic mass unit (u or amu).
- Relative molecular mass (or weight), without units and represented as M_r.
- Molar mass, with units g/mol.
- Relative molar mass, without units and equivalent to relative molecular mass.

These distinctions are all rather picky, but for clarity one should be careful to use the appropriate units when using one of these terms. The important point, for practical purposes is this: Given the molecular mass in any of these forms, one has in hand the conversion factor for converting between mass in grams and the number of moles. If we have a molecular species with a relative molecular mass, M_r, we can write:

$$1\,\text{mole} = M_r\text{g}$$
$$M_r\text{g/mol} = 1$$

So, for instance, if we have 30.0 of glucose with a molecular mass of 180.16 Da, we can calculate the number of moles as:

$$30.0\,\text{g} \div 180.16\,\text{g/mol} = 0.166\,\text{mol}$$

1.4.3 Special units of concentration for hydrogen and hydroxide ions

There are two ionic species that receive special attention whenever water is involved, and these are the hydrogen (H^+) and hydroxide (OH^-) ions. These two are always present, though usually at quite low concentrations, in aqueous solutions because water itself has a tendency to dissociate to produce both:

$$H_2O \rightleftharpoons H^+ + OH^-$$

The forward dissociation reaction rate is actually quite slow, so that the average time for a given water molecule to dissociate is about 11 h. However, even a small volume of water contains a large number of water molecules, and re-association of H^+ and OH^- is very fast, occurring essentially instantaneously once the the two ions colide in solution. As a consequence the forward and reverse reactions reach a balance in a fraction of a second. We will discuss equilibrium constants in more depth later in the course, when we study thermodynamics, but for now it is sufficient to say that the reaction quickly reaches an equilibrium state such that the concentrations of H^+ and OH^- are related to one another according to:

$$[H^+]_{eq}[OH^-]_{eq} = 10^{-14}\,\text{M}^2$$

where $[H^+]_{eq}$ and $[OH^-]_{eq}$ are the equilibrium concentrations of the two ions. Because the dissociation reaction reaches equilibrium very rapidly, we generally assume that the concentrations of H^+ and OH^- in a solution satisfy the equilibrium condition. An important consequence of this relationship is that if the concentration of either H^+ or OH^- is known, the concentration of the other is also determined. In an absolutely pure sample of water, the dissociation reaction should be the only source of H^+ and OH^-, and their concentrations should be equal. This defines what we describe as a neutral solution, and the equilibrium expression is satisfied under these conditions when the concentrations of H^+ and OH^- are 10^{-7} M.

Because the concentration of H^+ and OH^- can vary over a wide range, it is convenient to use special representation for their concentration, pH and pOH, respectively. The pH and pOH of a solution are defined by"

$$\mathrm{pH} = -\log[H^+]$$
$$\mathrm{pOH} = -\log[OH^-]$$

From this definition and the discussion above, you should be able to readily demonstrate that the pH and pOH of a neutral aqueous solution are both equal to 7 and that the sum of pH and pOH is equal to 14 for any solution. Although either pH or pOH can be used to describe the equilibrium concentrations of H^+ and OH^- in a solution, pH is, by far, the more commonly used parameter.

Earlier, we estimated the volume of an *E. coli* bacterium to be about 10^{-15} L. An interesting implication of this very small volume is that the number of some molecules and ions in a single cell are surprisingly small. For instance, we can ask: How many hydrogen ions are in a bacterium? If the pH in the cell is 7, then the concentration is

$$[H^+] = 10^{-\mathrm{pH}}\mathrm{M} = 10^{-7}\,\mathrm{M} = 10^{-7}\,\mathrm{moles/L}$$

The number of moles of H^+ ions is then calculated as

$$10^{-7}\,\mathrm{moles/L} \times 1.6 \times 10^{-15}\,\mathrm{L} = 1.6 \times 10^{-22}\,\mathrm{moles}$$

To calculate the number of ions, the number of moles is multiplied by Avogadro's number, which can be thought of as having the units of particles/mole

$$1.6 \times 10^{-22}\,\mathrm{moles} \times 6.02 \times 10^{23}\,\mathrm{ions/mole} \approx 100\,\mathrm{ions}$$

That's not very many!

There are also bacteria that grow at pH 9. How many hydrogen ions are present in one of these bacteria?

1.5 Further reading

For an authoritative reference on the SI units and conversion factors, see:

- Thompson, A. & Taylor, B. N. (2008). Use of the international system of units (SI). NIST Special Publication 811, National Institute of Standards and Technology, Gaithersburg, MD.
  http://physics.nist.gov/cuu/Units/bibliography.html

A convenient online unit conversion tool:

- http://www.digitaldutch.com/unitconverter/
  Keep this disclaimer in mind: "While we try really hard to make all calculations accurate, we do not guarantee that the results you get are correct. If you do find any bugs I highly appreciate it if you email us at info@digitaldutch.com"

Wikipedia contains a number of good articles on the metric system, including its history and the current SI. As a publicly edited secondary source, some caution is always advised when using Wikipedia (or any source, really), but these articles are well referenced, so that the primary sources can be checked.

- https://en.wikipedia.org/wiki/Metric_system
- https://en.wikipedia.org/wiki/Metre
- https://en.wikipedia.org/wiki/Kilogram
- https://en.wikipedia.org/wiki/International_System_of_Units
- https://en.wikipedia.org/wiki/2019_redefinition_of_the_SI_base_units

Wikipedia articles on English measurement systems:

- https://en.wikipedia.org/wiki/English_units
- https://en.wikipedia.org/wiki/Imperial_units
- https://en.wikipedia.org/wiki/United_States_customary_units

Articles on the redefinition of the kilogram and mole with new technologies:

- Cho, A. (2018). World poised to adopt new metric units. *Science*, 362, 625–626.
  http://doi.org/10.1126/science.362.6415.625
- Cho, A. (2017). Plot to redefine the kilogram nears climax. *Science*, 356, 670–671.
  http://dx.doi.org/10.1126/science.356.6339.670
- Robinson, I. A. & Schlamminger, S. (2016). The watt or Kibble balance: a technique for implementing the new SI definition of the unit of mass. *Metrologia*, 53, A46.
  http://dx.doi.org/10.1088/0026-1394/53/5/A46
- https://www.nist.gov/physical-measurement-laboratory/silicon-spheres-and-international-avogadro-project

Chapter 2

Probability

2.1 An example of a random process: Brownian Motion

I. Some history

Although we tend to think of biological movement, especially that of animals, as being rather directed, at microscopic scales motion in both living and non-living systems can be quite random. The discovery of this kind of motion is credited to a Scottish botanist, Robert Brown (1783–1858). Brown served on a expedition to Australia in 1801–1805 and spent many years afterwards characterizing the plants that he collected on this expedition. Brown was a particularly skilled microscopist, and in 1826 described the motion of tiny particles within pollen grains. He was not the first to observe this kind of motion, but others who had seen it assumed that it reflected some kind of living process. By carefully describing the motions and showing that they could be seen in materials that were clearly not living (like particles of coal dust suspended in water), Brown showed that the motions represented a physical process, rather than a biological one.

A movie of small particles, ≈100 nm diameter, undergoing Brownian motion is available online:
https://www.youtube.com/watch?v=cDcprgWiQEY

A theoretical model explaining Brownian motion was presented by Albert Einstein in 1905. This was just one of four major papers that Einstein published in his *annus mirabilis* (miracle year). The others concerned special relativity and the photoelectric effect (for which he won the Nobel Prize in 1921). The paper on Brownian motion would have probably made the reputation of just about any other scientist, but for Einstein it seems almost a footnote.

Einstein's explanation for Brownian motion was that particles in a liquid are constantly being bumped into by molecules, and each collision causes a small motion of the particle. Every once in a while, an imbalance in the number of molecules colliding from one side or the other causes a larger movement in a random direction. This is illustrated in the drawing below:

This drawing comes from an online simulation of Brownian motion.
http://galileoandeinstein.physics.virginia.edu/more_stuff/Applets/Brownian/brownian.html

On the left, a large particle moves randomly because of collisions of more rapidly moving small molecules. On average, the forces on the particle average to zero, but at any instant, there may be an imbalance of collisions from different directions, leading to a small motion. The right-hand panel shows the trajectory of the particle over time, as viewed at lower magnification.

Importantly, Einstein did not just describe this model qualitatively (which others before him had done), but developed a mathematical treatment that made quantitative predictions that could be tested by experiments. Experimental confirmation of this theory provided critical support for the existence of atoms and molecules, an idea that was still contentious at the beginning of the 20th century.

II. A mathematical description - random walks

1. A detailed, exact mathematical description of this process, with explicit descriptions of the behavior each molecule, would be almost impossible.
2. An important aspect of science is deriving an abstract description of a process that captures important elements but is as simple as possible. To quote Einstein (roughly): Theories should be a simple as possible, without being so simple that they fail to account for important observations.
3. The key element of Brownian motion is that in a given time interval, there is an equal probability that a particle will move one way or the opposite way.
4. We can describe the overall behavior of a particle undergoing Brownian motion as a random walk: A process made up of multiple steps separated by random changes in direction.
5. A one-dimensional random walk:
   - Flip a coin.
   - If the coin lands heads-up, take a step to the right. If the coin lands tails-up, take a step to the left.

- Repeat.

6. The Galton probability machine or "Galton board" - aka Plinko: A mechanized demonstration of a one-dimensional random walk, or any process made up of a sequence of binary random events.

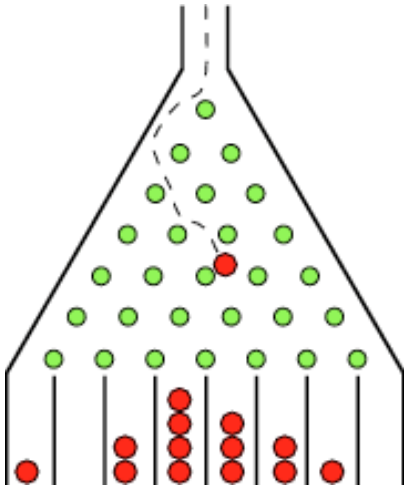


Illustration from http://mathworld.wolfram.com/GaltonBoard.html

- A triangular array of pegs placed on a vertical board. A ball is dropped on to the top peg and bounces to the left or right with equal probability. At each row, the ball hits a peg and moves to the left or right. The balls are collected in bins below the bottom row.
- Devised by Sir Francis Galton, 1822-1911: A cousin of Charles Darwin. Galton played an important role in developing the mathematical description of genetic variation and evolution, but was also a major advocate of the idea that society could be improved by the selective breeding of humans and gave this idea the name eugenics.
- Also known as a probability machine or "Plinko".
- Computer simulation: https://phet.colorado.edu/en/simulation/plinko-probability

2.2 Introduction to probability theory

I. Some introductory comments.

In terms of its relevance to science and every day life, probability is arguably one of the most important branches of mathematics. But, it is also has a bit of an odd position within mathematics, and it is, I think, severely under represented in our undergraduate curriculum. It is also one of the most challenging subjects to learn and teach.

1. Why is probability a misfit in the world of mathematics?

If you think about the traditional branches of mathematics, they are generally concerned with the properties of certain kinds of abstract objects:

- Geometry: lines, circles, polygons, planes, spheres, polyhedra, *etc.*
- Number theory: Integers
- Algebra: Polynomials
- Calculus (or analysis for the purists): Functions that change smoothly (usually).

Although all of these have applications in the real world, these branches of mathematics can be discussed completely in the abstract, and that is the way most mathematicians like it!

Probability, on the other hand, deals specifically with the description of real events of a certain type (or models of those events). In particular, probability deals with events about which we are, to some degree, ignorant. We use probability to describe things that have uncertain outcomes. If you think about it, there are lots of things like that!

2. Why is probability so difficult?

   A. One problem is that we constantly use the language of probability in our everyday lives, without necessarily paying attention to exactly what we mean. Some common expressions of a probabilistic nature:

      - It is likely that . . .
      - Chances are . . .
      - I'll bet that . . .

      We are also accustomed to hearing numbers associated with such statements, such as "There will be an 80% chance of rain tomorrow." What do statements like this mean, and where do they come from?

   B. Another problem is that discussions of unpredictable events often have large emotional component. For instance:

      - What are the chances that I will win the lottery?
      - What is the probability that I will get cancer?

      The probabilities of these events might (or might not) be similar to the probability of rain tomorrow, but our emotional responses to them are likely very different.

   C. The calculation of probabilities often involves some rather tricky counting, and the results often contradict our intuition.

   D. The answer to a probability question can depend on exactly how the question is framed. Make sure that you are answering the right question!

II. A coin toss

A typical probabilistic statement: If I toss a coin, the chances it will land heads-up are the same as the chances it will land heads-down.

What is implied by this statement?

- Ignorance: I don't actually know which way the coin will land.
- Knowledge: If I toss the coin a large number of times, the number of times it lands heads-up will be approximately the same as the number of times it lands tails-up.

These answers raise some more questions:

- Why don't I know which way the coin will land? Isn't this just Newtonian mechanics?
- How many times do I have to toss the coin before the number of heads will equal the number of tails? Will they ever be exactly equal?
- Can I say anything more specific about the expected pattern of heads and tail?

For now, we will try to address just one of these questions: Why can't I predict the outcome? The answer is that the final outcome (heads or tails) is extremely sensitive to a large number of small factors that we usually don't have control over. These factors include the exact force applied to the coin, the angle at which the force is applied, any air currents that affect the coin, exactly how the coin hits the surface when it lands. In principle if all of these factors could be controlled and measured, it should be possible to predict the outcome of the toss.

To some degree, the uncertainty of a coin toss is tied to the structure of the coin: The thin edge makes it almost certain to fall one way or the other, and the (near) symmetry makes it equally likely to fall either way. Of course, the coin may be bent or otherwise altered so that the probabilities of heads and tails are not equal.

III. A bit of mathematical formalism.

In order to develop a mathematical theory of probability, we have to make some careful definitions of quantities that we can manipulate. This will seem a bit much for a simple coin toss, but the definitions are important for keeping us straight as we move on to more complicated cases.

1. **Outcomes** For a given experiment, we define a set of distinct *outcomes*. For the coin toss, we define two outcomes, heads (H) and tails (T). Now, we could also consider other outcomes, like dropping the coin, but that makes things more complicated. So, what we usually do is simplify the situation by excluding things like dropping the coin or that it might land on its edge.
2. **Probabilities** For each of the possible outcomes, we define a *probability*, a number (p) constrained such that:
   - p for any given outcome must lie between 0 and 1, inclusive.
   - The sum of the probabilities for all of the possible outcomes is 1.

   For the coin toss, our experience and intuition says that:

   $$p(H) = 1/2$$
   $$p(T) = 1/2$$

What, exactly do we mean by this? This is not quite as obvious as it sounds, and there are actually two major ways of interpreting probabilities, reflecting some rather deep philosophical differences among probabilists. We will use the more intuitive and traditional view, called a "frequentist" interpretation.

A. The frequency interpretation of the statement, $p(H) = 1/2$, is simply that if a "fair" coin is tossed a large number of times, the fraction of times it lands heads-up will be approximately 1/2, and the fraction will, over time, get closer to 1/2 as the number of tosses is increased.
This general trend is called the "law of large numbers".

B. The alternative interpretation of probabilities is called "Bayesian", referring to Thomas Bayes, an 18^{th} century cleric and mathematician, who devised a very important equation concerning the probabilities of related events. Frequentists do not dispute Bayes' equation, but the Bayesians interpret and apply it more broadly. In brief, the Bayesian approach is used for situations in which we are asking questions for which there are not enough data to make a frequency estimate, such as, "What is the probability that it will rain tomorrow?" Since there has never been even one day exactly like tomorrow, there is no way to know the frequency of rain on such days. But, if we have an initial estimate of the probability, called a *prior* probability, additional information can be used with Bayes' equation to refine the initial estimate to create a *posterior* probability. The Bayesian approach is somewhat controversial, but it has become a very important tool in areas in which exact probabilities are not known, but there is a large amount of data with which to refine the estimates. One common example is filtering e-mail messages for spam.

   For our purposes, the frequency interpretation is most useful. It has a relatively intuitive basis, and most of the problems we will be considering, such Brownian motion and diffusion, involve large numbers of random events.

3. **Sample spaces** We call the set of all possible, distinct outcomes for a given experiment a *sample space*, S. For a single coin toss:

   $$S = \{H, T\}$$

   We will use curly braces, as above, to enclose the elements of the set. This gets a little more complicated when we consider more complicated experiments, such as multiple coin tosses. For two coin tosses, there are four possible outcomes, and we will define the sample set as:

   $$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

   where the outcomes are defined as ordered pairs, in parentheses, representing the results of the two independent coin tosses.
   This is a little bit arbitrary. We could define three outcomes defined in terms of the total number of heads or tails, irrespective of the order:

- Two heads: $2H$
- Two tails: $2T$
- One heads, one tails: $1H1T$

But, the case of $1H1T$ is actually the combination of two outcomes, (H, T) and (T, H), as initially defined.

The major difference between these two ways of defining the outcomes is that the probabilities of the individual outcomes are all equal for the first definition, but not for the second.

In general, we try to define the outcomes and the sample set to make assigning probabilities as simple as possible. This doesn't necessarily mean that the probabilities are all equal, though.

For the plinko, we would define the sample set as the set of all distinct paths through the pegs, not the set of all possible final bins.

Although there might be different ways of defining a sample space for a particular experiment, it must satisfy two requirements:

- The set must be complete, *i.e.,* it must include every possible way that things can end up.
- The items in the set must not overlap.

A consequence of these two requirements is that the sum of the probabilities of the outcomes must be exactly 1.

4. **Events** Formally, an *event* is defined as a subset of the sample set, *i.e.,* a set of zero or more of the possible outcomes.

   For example, with two coin tosses, we could define the events that we considered above:

   - Two heads: $2H = \{(H, H)\}$
   - Two tails: $2T = \{(T, T)\}$
   - One heads, one tails: $1H1T = \{(H, T), (T, H)\}$

   For the plinko, we could define an event as the ball falling into a given bin.

   As noted above, we generally try to define outcomes so that the probabilities can be easily calculated, and then use those probabilities to calculate the probabilities of events, or groups of outcomes.

   The choice of words, "outcomes" and "events", is pretty arbitrary, but the distinction between the two kinds of groupings is important. The outcomes of an experiment are events, but there are usually other events that can be defined as groups of outcomes. The outcomes defined in the sample space must satisfy the requirements specified earlier: They must include all possible outcomes of the experiment, and the sum of their probabilities are one, while there are no general requirements for events.

   We can often define a variety of different events, some of which may overlap. For instance we could define an event such that there is at least one heads.

   $$1^+H = \{(H, T), (H, H), (T, H)\}$$

This event overlaps the events $2H$ and $1H$.

There is no requirement that a set of events be complete or non-overlapping.

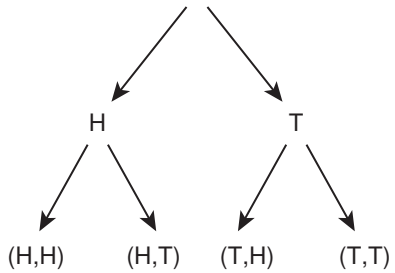
Often, it is the probabilities of events, as defined here, that is most important. For instance, we care about which bin the plinko ball falls in, but not necessarily the specific path it takes there. Thus, we often want to be able to calculate the probabilities of events from the probabilities of outcomes.

IV. Multiplying and adding probabilities.

Provided that we are careful in defining the sample space, the rules for calculating the probabilities of other events are relatively simple. For this discussion, it is useful to introduce another term, *trial*, to indicate a single probabilistic process or experiment. A trial that can have only two outcomes is referred to as a *binary trial* or *Bernoulli trial*, for the Swiss mathematician Jacob Bernoulli (1665–1705). It is also natural to refer to individual trials as events, but this leads to confusion with the definition of events as subsets of the sample set.

1. Sequential independent trials - the product rule.

   We can think of the experiment composed of two coin tosses as two sequential experiments, or trials, each with the sample space, $S = \{H, T\}$. In fact, just about any complicated process can be broken down in this fashion. It is often useful to draw a tree representation of the sequential trials, like the one below:



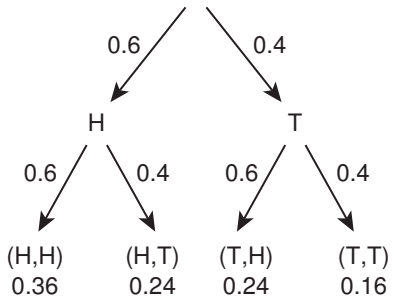
   It's not a coincidence that this looks like the plinko, but there is an important difference: All of the different outcomes are kept separate.

   For the first coin toss, $p(H) = p(T) = 1/2$. Therefore, if we do this experiment many times, we expect heads for the first toss half of the time. Consider just this half, for a moment. For the second toss, we also expect heads half of the time. So, out of all of the two-toss experiments, we expect the outcome (H, H) $1/2 \times 1/2$ of the time. Therefore, $p(H, H) = 1/4$. The same argument can be made for all of the outcomes of this experiment.

   The general statement of this result is that if we have two sequential and independent trials, then we can calculate the probabilities of the final outcomes of the compound experiment as the products of the individual probabilities. We call this the *product rule*. We can extend it to compound experiments of any length.

   Application of the product rule is often associated with the word "and." For instance, the outcome (H, H) can be described as "heads for the first toss *and* heads for the second toss."

In this particular case, the product rule leads to the conclusion that all of the outcomes in the sample space have equal probabilities, 1/4. But this is not always the case. Suppose that we are playing with a coin that has somehow been messed with so that the probability of landing heads-up is 0.6 and the probability of landing tails-up is 0.4. We can still use the same arguments and the product rule:



Notice that the sum of the probabilities is still 1.

2. Groups of non-overlapping events - the addition rule.

   A. We have already used this rule implicitly.
   Consider the event we defined earlier, $1H1T$, *i.e.,* one heads and one tails, irrespective of order. This event is a composite of two outcomes:

   $$1H1T = \{(H,T),(T,H)\}$$

   The probability of $1H1T$ is calculated as the sum of the outcomes:

   $$p(1H1T) = p((H,T)) + p((T,H))$$

   Just as we said that the product rule is associated with the word "and", we can say that the addition rule is associated with "or". The event $1H1T$ can be described as being the result when (H,T) *or* (T,H) is the outcome.
   If $p(H) = p(T) = 1/4$, then $p(1H1T) = 1/2$.

   B. What is $p(1H1T)$ if $p(H) = 0.6$? What can we say in general about $p(1H1T)$ if $p(H)$ is not equal to $p(T)$?
   Consider two extreme cases:

   - $p(H) = 0$ and $p(T) = 1$. Then:

   $$\begin{aligned} p(1H1T) &= p((H,T)) + p((T,H)) \\ &= p(H)P(T) + p(T)P(H) \\ &= 0 \end{aligned}$$

   - $p(H) = 1$ and $p(T) = 0$. It should be apparent that $p(1H1T) = 0$, again.

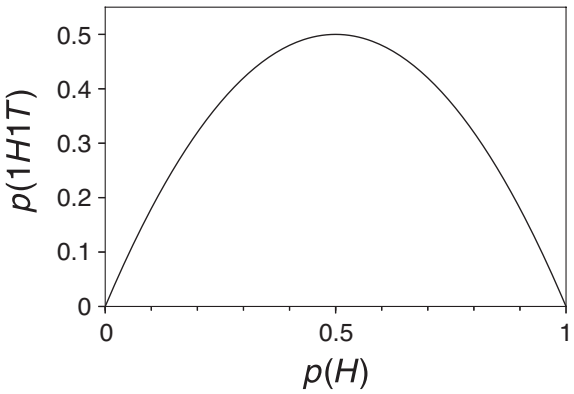
   We can write a general expression for $p(1H1T)$ as a function of $p(H)$, assuming that the two coin tosses are equivalent:

   $$\begin{aligned} P(1H1T) &= p((H,T)) + p((T,H)) \\ &= 2p((H,T)) \\ &= 2p(H)p(T) \end{aligned}$$

We also know that $p(T) = 1 - p(H)$, so:

$$\begin{aligned} p(1H1T) &= 2p(H)(1 - p(H)) \\ &= 2p(H) - 2p(H)^2 \end{aligned}$$

A graph of this function:



With a little bit of calculus, you should be able to confirm that 1/2 is the maximum probability of one heads and one tails. If the coin is biased either way, the probability is less.

Another example: Consider the event we defined earlier, one or more heads.

$$1^+H = \{(H,T),(H,H),(T,H)\}$$

We calculate the probability of this event as the sum of the probabilities of the three outcomes it represents:

$$p(1^+H) = p((H,T)) + p((H,H)) + p((T,H))$$

If the coin is fair, each of the outcomes has equal probability, and $p(1^+H) = 3/4$.

But, there is an even easier way to get this result. The only outcome that is not included in 1^+H is (T,T). Since the sum of the probabilities of all outcomes must be 1:

$$\begin{aligned} p(1^+H) &= 1 - p((T,T)) \\ &= 1 - p(T)p(T) \end{aligned}$$

If the coin is fair, then $p(T) = 1/2$ and $p(1^+H) = 3/4$. Sometimes it is important to consider which probabilities will be the easiest to calculate.

V. A final comment about independent events and the law of large numbers.

Consider the case of a long string of coin tosses. Suppose that 10 straight tosses turn up heads. Someone offers you a bet: If the next toss turns up tails, she will pay you \$1, if it turns up heads, you pay her \$1. Is this a "better-than-even" bet?

The law of large numbers says that eventually the numbers of heads and tails will be close to equal. So, is it time for the coin to show up tails?

No! The coin doesn't know or care about the law of large numbers! Each toss is independent, so the probability of tails for the 11^{th} toss is the same as for the first toss.

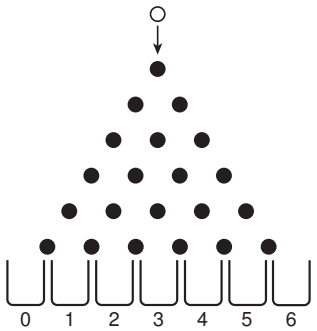
Thinking that "it's time for a tails" is known as the "gambler's fallacy", and has cost many people lots of money over the ages!

But, is there another way of thinking about this situation? What have we assumed about the coin (or its tosser)? If that assumption is called into doubt, how does that change our assessment?

2.3 Plinko probabilities: 6 rows

I. Formulation of the problem

A 6-row plinko:



The white circle represents the ball, and the black circles represent the pegs in the path of the ball.

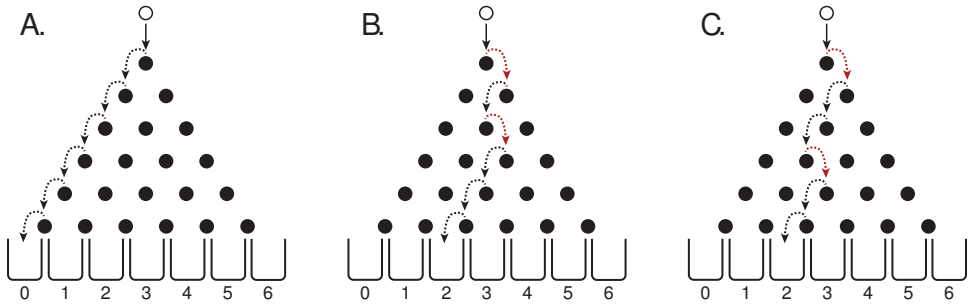
For a general n-row plinko, the bottom row of pegs will contain n pegs. Since a ball can fall to the right or left of each peg, there are $n + 1$ final positions, or buckets, for the balls to fall into. For convenience in what comes later, we label the buckets from 0 to n, or 0 to 6 for the 6-row plinko.

II. Outcomes We have some discretion in defining the outcomes and sample set, so long as we follow the basic rules:

- The outcomes in the sample set must include all possible outcomes.
- None of the outcomes in the sample set can overlap any other outcome.
- The sum of the probabilities of all of the outcomes in the sample set must equal 1.

At first glance, it might make sense to define seven outcomes, corresponding to a ball falling in bucket 0, 1, 2, 3, 4, 5 or 6. We know already, however, that the probabilities of these seven outcomes are not equal, and we will find that calculating them is rather involved. So, instead, we will start by defining the outcomes as all of the possible paths of a ball through the plinko, which all have equal probabilities. Then, we will use the elements in the sample set to calculate probabilities for the events corresponding to a ball landing in each of the buckets.

A few of the outcomes, individual paths, are shown below:



Notice that the paths labeled B and C both lead to bucket 2, but we are treating these as separate outcomes. Notice, also, that both of these paths include two turns to the right, where as the path to bucket 0 includes 0 turns to the right. More generally, any path leading to bucket k must include exactly k turns to the right.

First, we calculate the number of outcomes in the sample set and their probabilities. When the ball hits the single peg in the top row, there are two possible turns, left or right. Similarly, when the ball hits one of the pegs in the second row, there are two possible turns. Each of these turns are independent, just like a series of coin flips. Therefore, the total number of paths is equal to 2^n, where n is the number of rows. For the six-row plinko the number of outcomes is $2^6 = 64$. If the probability of a right or left turn at each peg is equal (0.5), then the probabilities of all of the outcomes are equal, and are equal to one divided by the number of possible outcomes. Thus the probability of each outcome for an $n-$row plinko is 2^{-n}. For the six-row plinko,the probability is 1/64.

III. Events

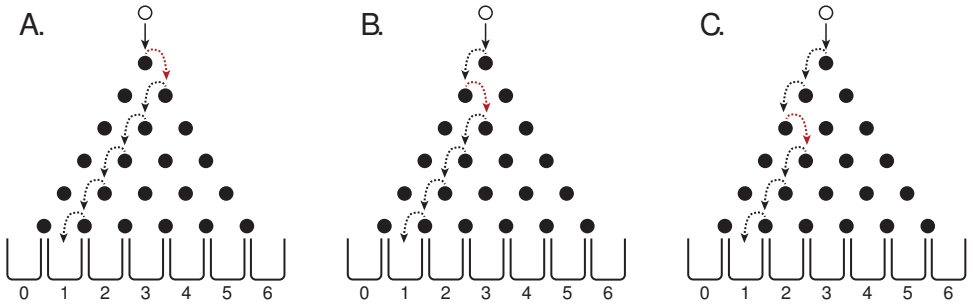
Next, we consider the events corresponding to the ball falling in one of the seven buckets, which we will call $E0$, $E1$, $E2$, $E3$, $E4$, $E5$ and $E6$. One way that we could do this is to write out all of the outcomes (paths) and sort these into those for which the ball lands in bucket 1, 2 and so forth. This would be quite tedious, however, and we would like to be able to do this for much larger numbers of steps. Therefore, we want a more general and efficient way to solve this sort of problem.

1. Paths to buckets 0 and 6

   If we consider first the possible paths to bucket 0, we quickly realize that the ball will reach this bucket only if all of the turns are to the left, as shown in panel A in the figure above. So, the probability of landing in bucket 0, $E0$, is 1/64. Similar reasoning can be applied to conclude that there is only one path to bucket 6, also with a probability of 1/64.

2. Paths to buckets 1 and 5

   In order for the ball to land in bucket 1, the ball must make 1 turn to the right and 5 to the left. Three such paths are shown below:

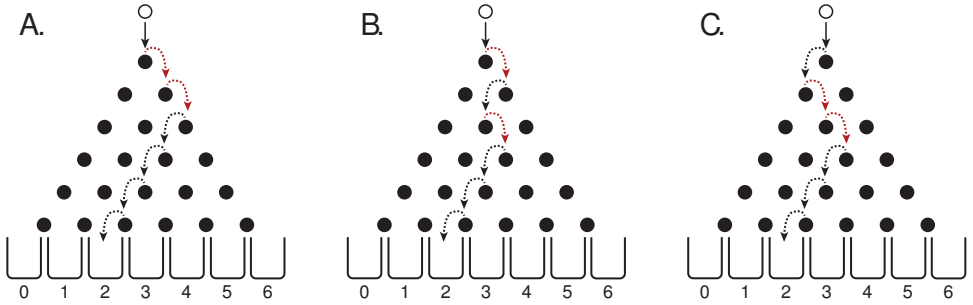


Since there are six rows, at each of which the single turn to the right can occur, there must be six different paths to bucket 1, making up the event $E1$. So, the probability of $E1$ is $6/64 = 3/32$. The same reasoning can be applied to the paths to bucket 5, and the probability of $E5$ is 3/32.

This result can be generalized to say that for an n-row plinko, there are n paths to bucket 1 and to bucket $n - 1$.

3. Paths to buckets 2 and 4

   Things get more complicated when we consider paths to bucket 2, where we must enumerate the possible paths that include exactly two turns to the right. The two turns can occur at any of the six rows, as shown in a few examples:



   In panels A and B, the first turn is to the right, and the second turn to the right is at row 2 (A) or row 3 (B). In panel C, the first turn is to the left, and the two turns to the right occur in rows 2 and 3, followed by 3 turns to the left.

   To determine the number of paths to bucket 2, without drawing them all out, we can calculate the number of paths as follows:

   - Consider the number of positions for the first turn to the right. This can happen at rows 1 through 5. (If the first turn to the right occurs at row 6, there is no chance for a second turn to the right.) If the first turn to the right is at row 1, then the second can occur at rows 2 through 6, corresponding to 5 paths. This is analogous to a 5 row plinko and the number of paths to bucket 1.
   - If the first turn to the right is at row 2, there are only 4 rows left at which the second right turn can occur.
   - Generalizing, the further down the ball moves before the first right turn, the fewer rows there are where the second right turn can occur. Specifically, if the

first turn to the right occurs at row i, then there are $i - 1$ possible locations for the second turn.

- For a 6-row plinko the total number of paths to bucket 2 is calculated as:

  $$5 + 4 + 3 + 2 + 1 = 15$$

  By considering the number of ways of placing two *left* turns, we can conclude that there are also 15 paths to bucket 4.

4. Paths to bucket 3

   We can now almost fill a table showing the number of baths to each of the buckets

Bucket	Paths
0	1
1	6
2	15
3	
4	15
5	6
6	1

   Since we have already concluded that the total number of paths to all of the buckets is 64, and 44 paths are accounted for so far, there must be 20 paths to bucket 3.

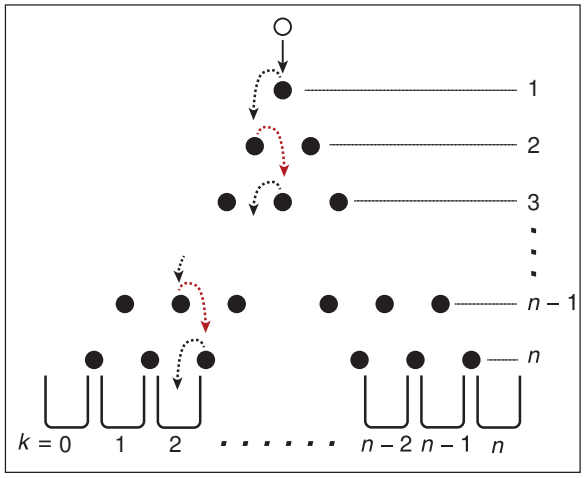
   The number of paths and probabilities for all of the buckets can now be listed:

Bucket	Paths	Probability
0	1	1/64
1	6	3/32
2	15	15/64
3	20	5/16
4	15	15/64
5	6	3/32
6	1	1/64

   Though we have been able to solve the problem for the 6-row plinko without too much trouble, you will likely guess, correctly, that enumerating all of the paths gets more and more complicated as the number of rows increases. To generalize the solutions to problems of this type of problem, we need to take a different approach.

2.4 Plinko probabilities: The general case for n rows

To keep track of the rows and buckets in the general case of an n-row plinko, we will label them as shown below:



Before trying to solve the general form of this problem, it is useful to step back and look at things a bit differently, and also consider some related probability problems.

I. Another way to count the paths to bucket 2 in a 6-row plinko.

Recall that we concluded that any path to bucket 2 must include 2 turns to the right and 4 turns to the left. A seemingly sensible (but flawed) way of looking at this would be to say that the first turn to the right can occur at any of the 6 rows, and the second turn to the right can occur at any of the 5 rows that are remaining. So, using the product rule, we would calculate the number of paths to bucket 2 as:

$$6 \times 5 = 30$$

Notice that this is twice the number that we calculated earlier! The reason for this is that this calculation has ignored the fact that one of the turns to the right has to come before the other. For instance, for the path that includes right turns at rows 2 and 5, the turn at row 2 has to come first. But, in our second calculation we included both this path and one in which the turn at row 5 comes before the one at turn 2, which is physically impossible! More generally, by simply taking the product of 6 and 5, we have counted twice each of the 15 paths that we counted earlier. But, if we take this into account, and divide 30 by 2, we get the right answer.

So a general strategy might be to calculate the number of all possible placements of the right turns, without worrying at first about the order of the turns, and then correct for the requirement that order does matter.

For the case of bucket 3, we can start by considering (ignoring order) that there are 6 rows where the first turn to right can occur, 5 where the second can occur and 4 where the third can occur. So the total number of paths (with over-counting) is:

$$6 \times 5 \times 4 = 120$$

But, how do we determine how many paths have been over-counted?

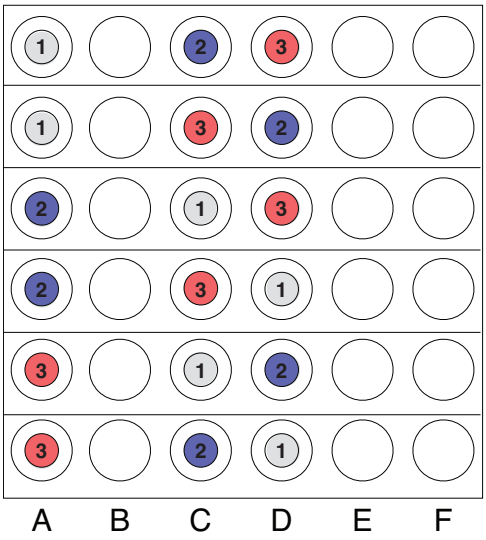
II. Labeled beans in cups

Though the connection may not be apparent just yet, it is useful to consider another type of problem that is popular among probabilists.

Suppose that we have 3 beans, each labeled with a number; 1, 2 or 3, and six cups. How many distinguishable ways are there to place one bean in one of the six cups? This is basically the same as the previous problem: There are 6 possible cups for the first bean, 5 for the second and 4 for the third. So the number of distinguishable different arrangements is:

$$6 \times 5 \times 4 = 120$$

The important point here is that these have *not* been over-counted, because the three beans are distinguishable. For instance, the following 6 arrangements are distinct:



For the general case of k labeled beans in n cups (assuming that $k \leq n$), the number of distinguishable arrangements is:

$$n(n-1)(n-2)\cdots(n-k+1)$$

You should be able to see where the first part of this product comes from, but it may not be so obvious that $(n-k+1)$ is the correct place to end the multiplication. So, you should try out a few examples to convince your self. For instance if $n = 10$ and $k = 6$, $(n-k+1) = (10-6+1) = 5$, and the number of distinct arrangements is:

$$\begin{aligned} n(n-1)(n-2)\cdots(n-k+1) &= 10 \times 9 \times 8 \times 7 \times 6 \times 5 \\ &= 151,200 \end{aligned}$$

There are six terms in the product, corresponding to the 6 labeled beans, and the final term, 5, represents the five empty cups available for the last bean. Notice, also, how quickly the number of possible arrangements has increased with a few more beans and cups!

III. The factorial function, permutations and combinations

Products like the ones used above arise frequently in probability and other areas of mathematics, and there is a function that is particularly useful for working with them. The factorial function is defined only for the integers greater than or equal to 0 and the factorial function of integer k is written as $k!$. The function is defined as:

$$k! = \begin{cases} 1, & \text{if } k = 0 \\ n(n-1)(n-2)\cdots 2 \cdot 1, & \text{if } k > 0; \end{cases}$$

Defining 0! as 1 may seem arbitrary (Why isn't it 0?), but this is important in order for the function to behave well when 0! appears.

An immediate application of the factorial function is that $n!$ is the number of ways arranging n labeled beans in n cups. This represents the special case, with $k = n$, of arranging k labeled beans in n cups. From the previous page, the number of distinct arrangements is:

$$\begin{aligned} n(n-1)(n-2)\cdots(n-k+1) &= n(n-1)(n-2)\cdots(n-n+1) \\ &= n(n-1)(n-2)\cdots 2 \cdot 1 \\ &= n! \end{aligned}$$

A distinct way of ordering all of the elements in a set is called a *permutation*. The items might be beans with distinct numbers, marbles with different colors or molecules with distinguishable covalent structures or conformations. So, we can say, "There are $k!$ permutations of k labeled beans." This is written as:

$$P(k) = k!$$

Note that we are using the upper-case P here to distinguish permutations from probabilities, written with the lower-case p. Another (mathematically equivalent) example of a set of permutations begins with k labeled marbles in a bag, and we draw all k of them from the bag. There are $k!$ different orders in which the marbles can be drawn.

An extension to this idea is to consider the number of ways of drawing k marbles from a bag starting with $n \geq k$ marbles. Strictly speaking, these are not permutations if $n > k$, because not all n elements are used, but they are often referred to as "k-permutations" of n, written as $P(k, n)$. The number of sequences is calculated as:

$$P(k, n) = n(n-1)(n-2)\cdots(n-k+1)$$

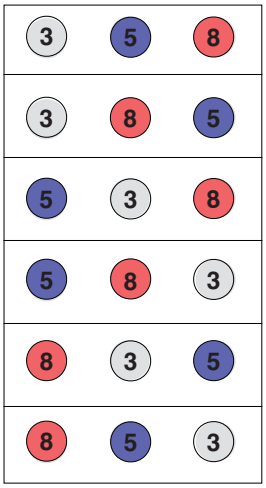
Another way of writing this is:

$$P(k, n) = \frac{n(n-1)\cdots(n-k+1)(n-k)(n-k-1)\cdots 2 \cdot 1}{(n-k)(n-k-1)\cdots 2 \cdot 1} = \frac{n!}{(n-k)!}$$

This is equivalent to the problem we considered in the previous subsection, the number of distinct ways of distributing k labeled beans into n cups, with only one bean per cup.

Now, we have a nice compact way of writing the result. And, if we have a calculator or computer programmed to calculate the factorial function, it is quite easy to do the calculation.

The term permutation is sometimes confused with *combination*. A combination is a distinct way of selecting a subset of a collection without regard to order. For instance, we might have a bag of 10 marbles, labeled 1 through 10, and without looking, choose 3 of them. From above, we know that there are $P(3, 10) = 720$ distinct ways of choosing the three marbles, *if* we treat the different orders of choosing the marbles as distinct from one another. For instance, there are 6 ways of choosing the marbles labeled 3, 5 and 8:



These represent the 6 permutations ($P(3) = 3!$) of the chosen marbles. For any other set of three marbles, there are also 6 permutations. Suppose that, after the 3 marbles have been drawn, the labels were to disappear. The six permutations of each group of 3 marbles would be indistinguishable, and the order in which they were drawn would no longer be discernable. So, to calculate the number of combinations in which 3 marbles can be drawn from a bag of 10, we can do the following:

- Calculate the number of ways in which 3 marbles can be drawn, distinguishing among the different possible orders. This is calculated as:

$$P(3, 10) = \frac{10!}{3!} = 720$$

- Calculate the number of ways in which 3 labeled objects can be ordered:

$$P(3) = 3!$$

- Divide the number of ways 3 objects can be drawn from 10 (distinguishing different orders) by the number of ways 3 objects can be ordered:

$$\frac{P(3, 10)}{P(3)} = \frac{10!}{(10-3)!} \div 3! = \frac{10!}{8!3!} = \frac{3,628,800}{5040 \cdot 6} = 120$$

To generalize, the number of ways of choosing k objects from a set of n, without distinguishing the order, is calculated as:

$$\frac{n!}{k!(n-k)!}$$

We will come back to say a little more about this function, and its more general applications on page 46.

In the meantime . . .

IV. Back to the plinko

Back on page 39, we suggested that a general way of calculating the number of paths to bucket k in an n-row plinko would be:

- Calculate the number of possible placements of the k turns to the right, without regard to order, realizing that this will lead to over-counting the real number of paths.
  This is is analogous to placing k labeled beans in n cups, with no more than one bean per cup. We have shown above that this is calculated as:
  $$P(k,n) = \frac{n!}{(n-k)!}$$
- Correct the number calculated above by recognizing that, for k turns placed at k specific rows, only one order is physically possible.
  For this part of the calculation, first consider the number of ways of placing k right turns in k specific rows. This is the number of permutations of k objects, $P(k) = k!$. But, we know that only one of these represents a physically possible pathway through the plinko. So, to get the total number of paths with k right turns, we divide the number of possible placements of the k turns to the right, without regard to order, $P(k,n)$, by $P(k)$:
  $$\frac{P(k,n)}{P(k)} = \frac{n!}{k!(n-k)!}$$

Thus, we have the desired result, the number of paths to bucket k in an n-row plinko is calculated as:

$$\frac{n!}{k!(n-k)!}$$

To test this result, you should apply it to the case of the 6-row plinko, for which we previously calculated the number of paths to each bucket (page 38).

The expression that we have derived for the number of paths to bucket k arises in a variety of situations and is commonly written as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

and spoken as "n choose k". It represents the number of ways of choosing k objects from a set of n, when either:

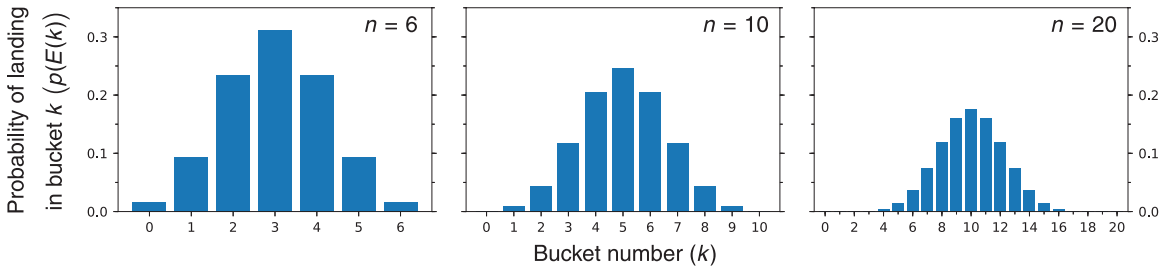
- Only a single order is valid (*i.e.,* the plinko) or
- The order doesn't matter at all (k unlabeled beans placed in n cups, with only one bean per cup allowed).

The total number of paths through an n-row plinko is calculated by multiplying the number of alternatives from the single peg in row 1 (2) by the number of alternatives from a peg in row 2 (2), and then multiplying by the number of alternatives from a peg in row 3 (2), and so on. Thus, the number of paths is 2^n. If, at each peg, the probabilities of turning right or left are equal, then all of the paths will have equal probabilities, equal to 2^{-n}.

The probability of landing in bucket k, that is event $E(k)$, is the sum of the probabilities for all of the paths leading to the bucket. If all of these paths have the same probability, 2^{-n}, then the probability of landing in bucket k is:

$$p(E(k)) = \frac{n!}{k!(n-k)!}2^{-n}$$

The probabilities for plinkos with 6, 10 and 20 rows are shown as bar graphs in the figure below:



Some things to note about these graphs are

- Each graph has the familiar "bell-curve" shape that arises frequently in a variety of contexts.
- As the number of rows, n, increases the maximum probability decreases, as the balls are spread out into more buckets.
- Also as n increases, it becomes increasingly unlikely that a ball will land in one of the buckets near the left or right end. As a fraction of the total number of buckets, the distribution of balls becomes more concentrated towards the center.

We will consider all of these features in more detail as we see the same type of distribution arise in different contexts.

2.5 Biased plinkos

So far, we have assumed that the probability of a ball falling to the left or right at any peg is equal. This assumption leads to the conclusion that all of the paths through the plinko have

equal probabilities, and that the different probabilities for landing in the different buckets are only due to the different *numbers* of paths leading to the different buckets. But, things get more interesting when we consider that the probabilities of left and right turns might be unequal for some or all of the pegs.

Suppose that all of the pegs are not quite round, so that the probability of turning to the right, p_{R}, is 0.6, and the probability of turning left, $p_{\mathrm{L}} = (1 - p_{\mathrm{R}})$, is 0.4. For now, we will consider a specific case of a 10-row plinko, starting with the paths leading to bucket 3. The fact that the pegs are biased doesn't change the number of paths leading to a specific bucket, which we calculate as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{10!}{3!(10-3)!} = 120$$

Now, consider the fact that right and left turns have different probabilities. For each of the 120 paths to bucket 3, there are 3 turns to the right and 7 turns to the left. This represents an "and" situation: The ball must take 3 right turns AND 7 left turns. So, to calculate the probability of each path, we have to multiply the probabilities for 3 right turns and 7 left turns.

$$p_{\mathrm{path}}(3) = p_{\mathrm{R}}^3 \cdot p_{\mathrm{L}}^7$$

Note that the placements of the 3 right turns and 7 left turns does not matter in this context, and this expression applies to all of the paths that lead to bucket 3. Since a ball can land in bucket 3 by any of the paths with equal probability (an "or" situation), the probability of landing in that bucket is the *sum* of all of the probabilities for the individual paths:

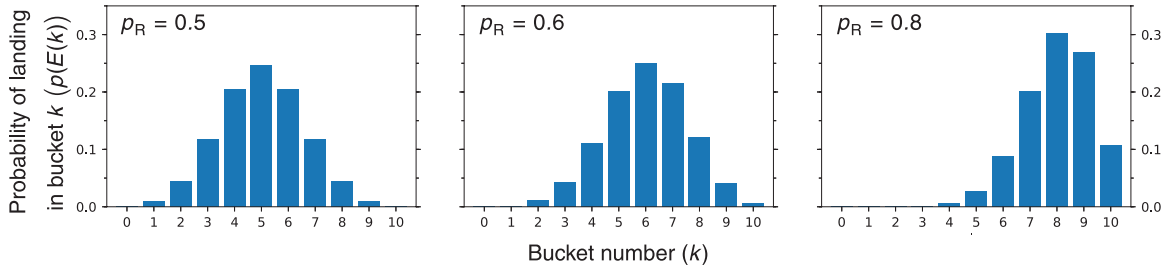
$$p(E(3)) = \frac{10!}{3!(10-3)!} p_{\mathrm{R}}^3 \cdot p_{\mathrm{L}}^7$$

For the case where $p_{\mathrm{L}} = 0.4$ and $p_{\mathrm{R}} = 0.6$, the probability of a ball landing in bucket 3 is 0.0425, compared to 0.117 for the unbiassed plinko. The bias of each turn towards right has moved the overall distribution of probabilities towards the right, reducing the probabilities of falling on the left-hand side of the plinko.

The more general expression, for bucket k in an n-row plinko is:

$$p(E(k)) = \frac{n!}{k!(n-k)!} p_{\mathrm{R}}^k \cdot p_{\mathrm{L}}^{n-k}$$

The bar graphs bellow show the effects of making the right turns progressively more favored, for the case of the 10-row plinko.



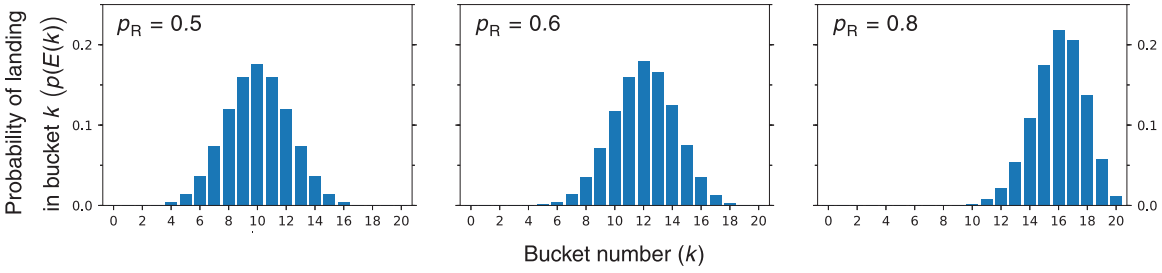
As the graph shows, increasing the probability of a turn to the right at each peg, leads to a progressive shift of the overall distribution to the right. Whereas the probability of a ball landing in bucket 10 is about 0.1% when there is no bias, this probability increases to about 10% if p_R is increased to 0.8.

When the probabilities of right and left turns are not equal, there are two competing factors that determine the distribution:

- A statistical factor favoring the central buckets, because there are more paths available toward these buckets than towards the buckets near the left and right hand edges of the plinko.
- A "forcing" factor that causes a systematic tendency towards one side of the plinko or the other.

The forcing factor can be adjusted by changing the relative values of p_R and p_L as shown in the graphs above. The statistical factor, on the other hand, can be modified by changing the number of rows in the plinko. For instance, with 10 rows, there are 252 paths to the central bucket, as compared to 1 for each of the buckets on the edge and 10 for the buckets one in form the edges. With 20 rows, there are 184,756 paths to the central bucket, as compared to 1 to each of the buckets on the edge and 20 to the buckets one in from the edges. Thus, the statistical bias towards the center is much greater for the 20-row plinko.

The graphs below show the effects of increasing biases to the right for a 20-row plinko.



As expected from the arguments above, the distribution is still shifted towards the right, but the buckets at the far right side are not nearly as favored as they are in the 10-row plinko, because the statistical "resistance" to the bias is greater.

One can imagine other ways in which the plinko could be biased, with only selected pegs with unequal values of p_R and p_L. Think of some examples of this kind and see if you can calculate the probabilities for these scenarios.

2.6 Binomial coefficients, Pascal's triangle and the binomial distribution function

I. Binomial coefficients in algebra

Recall that the general expression that we found for calculating the number of paths to bucket k in an n-row plinko is:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

This expression arises in a variety of contexts, and the values generated from it are most commonly called *binomial coefficients*. This term reflects their appearance in algebra in the expansion of binomials, which have the general form of:

$$(a+b)^n$$

where n is the order of the binomial. The results for expanding the binomial for $n = 0$ through 6 are shown below

$$(a+b)^0 = 1$$
$$(a+b)^1 = a+b$$
$$(a+b)^2 = a^2+2ab+b^2$$
$$(a+b)^3 = a^3+3a^2b+3ab^2+b^3$$
$$(a+b)^4 = a^4+4a^3b+6a^2b^2+4ab^3+b^4$$
$$(a+b)^5 = a^5+5a^4b+10a^3b^2+10a^2b^3+5ab^4+b^5$$
$$(a+b)^6 = a^6+6a^5b+15a^4b^2+20a^3b^3+15a^2b^4+6ab^5+b^6$$

If you examine the coefficients for the 6$^{\text{th}}$-order expanded binomial, you will find that they are exactly the same as the number of paths to the buckets in the 6-row plinko (page 38).

This may seem an odd coincidence, but there is an underlying connection. In the plinko, the number of paths to bucket k reflects the number of ways of combining k turns to the right with $n-k$ turns to the left, and the most paths are found at the center ($k = n/2$ when n is even, or $k = n/2-0.5$ and $k = n/2+0.5$ when n is odd). In a binomial expansion, the coefficients reflect the number of ways of multiplying together a k times and b $(n-k)$ times, to generate products of the form $a^k b^{n-k}$.

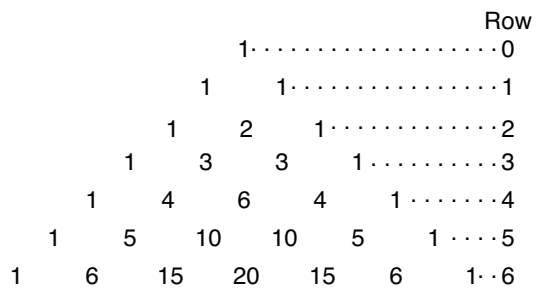
The binomial theorem, in its simplest form, is the equation:

$$(x+a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k}$$

where x and a are real numbers, and n is a positive integer. For a discussion of extension of the theorem to other classes of numbers, see: https://en.wikipedia.org/wiki/Binomial_theorem

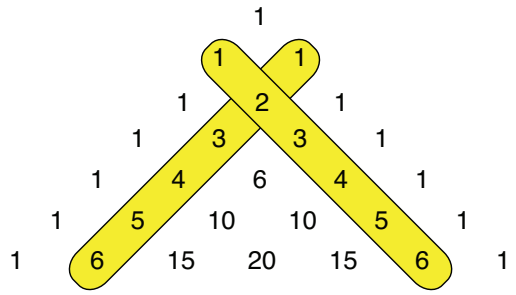
II. Pascal's triangle

The binomial coefficients for increasing values of n can be laid out in a triangle as shown below:

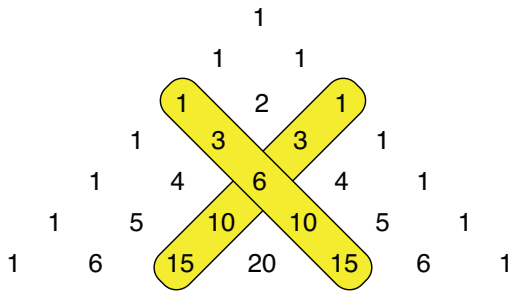


If the rows in the triangle are labeled from zero for the top row, row n contains the coefficients for the binomial expansion of order n. Although this representation was known before the 2nd century BC, it was brought to prominence in the western world by the French mathematician, Blaise Pascal, in a book published in 1665, two years after his death. It is now most commonly identified as *Pascal's Triangle*.

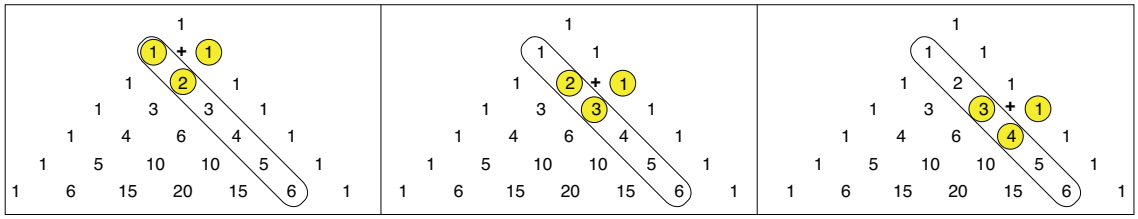
When the coefficients are laid out as a triangle, some interesting patterns become apparent. For instance, the two diagonals starting from the top of the triangle are composed entirely of 1s. The next diagonals down from the top are made up of the sequence of natural numbers: $1, 2, 3, \ldots$.



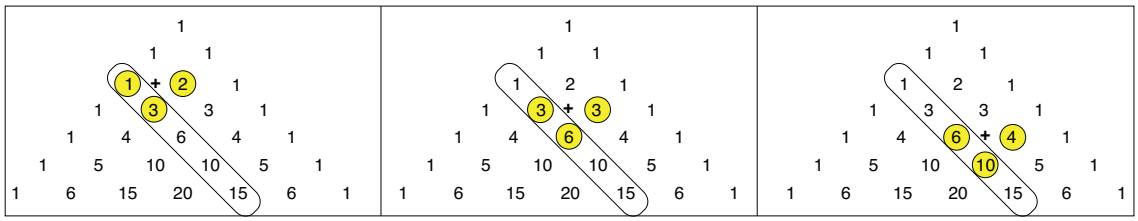
On the other hand, the pattern for the diagonals two down from the top may not seem to show an obvious pattern at first:



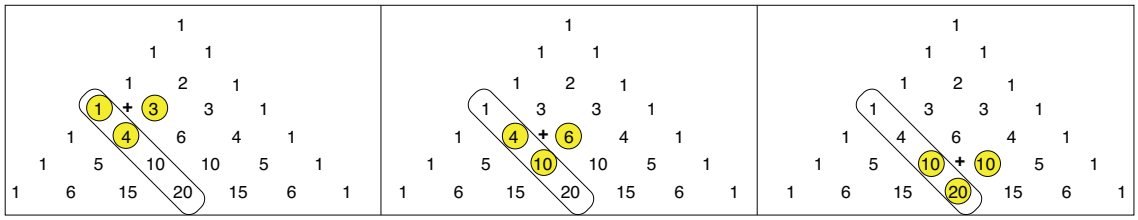
In this case the numbers along the diagonals increase as we move downward, and the increments themseles increase. The series of numbers on each of the diagonals do, in fact, follow a pattern, described by the *figurate numbers*, but that is a rather abstract bit of algebra that we will not pursue. There is, however, an easy way to calculate all of the elements in Pascal's triangle. If we go back to the second diagonal on the right-hand side, we find that each element of the diagonal is the sum of the numbers above and to the sides of it.



The same is true for the elements in the third diagonal on the left:



And, the fourth diagonal:



Indeed, all of the elements of Pascal's triangle can be calculated in this way. So, if we remember only that the outer edges of the triangle are composed entirely of 1s, the rest of the elements in the triangle, to any depth, can be calculated by taking the sum of the two elements above and to the sides of it.

III. The binomial probability distribution function: Trials and successes

The equation that we have derived for for calculating the plinko probabilities has considerably wider applications, and it is also representative of a general class of functions that we call *probability distribution functions*, or *pdf*s (not to be confused with *portable document format*). In particular, this is a class of pdfs described as *discrete probability distribution functions*, which, in general, are used to calculate the probabilities of outcomes or events, for the kinds of process for which the outcomes are discrete, such as tossing coins, rolling dice or the plinko. For many other processes, the outcomes are best described as continuous range of values, for which the probabilities are given by a *continuous probability distribution function*, which we will come to shortly.

Discrete probability distribution functions are commonly identified in the form:

$$p(k; a, b, \dots)$$

where k identifies a specific event, and $a, b \dots$ represent parameters of describing the process and probability function. For the binomial distribution function:

$$p(k; n, p_\mathrm{s}) = \frac{n!}{k!(n-k)!} p_\mathrm{s}^k (1 - p_\mathrm{s})^{n-k}$$

In the general formulation, this distribution represents the number of successes (k) in a series of n successive binary trials, with only two possible outcomes (success and failure), and p_s is the probability of success for an individual trial.

Some of the applications of the binomial distribution are:

- The plinko. Here each peg the ball hits represents a trial, with the "success" in this case defined as a turn to the right. In order to land in bucket k, there must be exactly k successes.
- The number of heads (or tails) in n successive coin tosses.
- The number of successes in prescribing a medication to a series of patients with the same condition, where p is the probability of success in any individual case.
- The probability of surviving n potentially deadly events. In this case, $k = n$, since we are only concerned with the case of surviving all n trials.

2.7 Random variables, expected value, variance and standard deviation

I. Playing for money

The origins of probability theory are closely tied to games of chance, and this context still offers one of the most vivid ways to start to think about the subject. To make the plinko more interesting, I might decide to let people drop a ball into a six-row plinko and promise to pay them $\$k$ if the ball lands in bucket k. Unless I only want to hand out money, I will need to charge people to play the game. So, I would like to know how much I need to charge if I want to at least not lose money. In other words how much, on average, should I expect to pay out to each player?

In this case, there is a relatively simple way to solve the problem:

- The probabilities of the ball landing in buckets 0 and 6 are equal. The average payout for these two buckets is $(\$0 + \$6)/2 = \$3$.
- The probabilities of the ball landing in buckets 1 and 5 are equal. The average payout for these two buckets is $(\$1 + \$5)/2 = \$3$.
- The probabilities of the ball landing in buckets 2 and 4 are equal. The average payout for these two buckets is $(\$2 + \$4)/2 = \$3$.
- The payout for bucket 3 is \$3.

So, the overall average payout must be \$3, and I should charge the players at least this much if I don't want to lose money.

The symmetry of the plinko makes the solution to this problem relatively easy to recognize, but in order to deal with more complicated problems, we will introduce some additional concepts, random variables and the expected value. Both of these concepts have been implicit in what we just did, but more formal definitions are called for when extending the ideas.

II. Random variables

A random variable is defined as a variable that is assigned a value for each possible outcome or event from a probabilistic process. Some examples include:

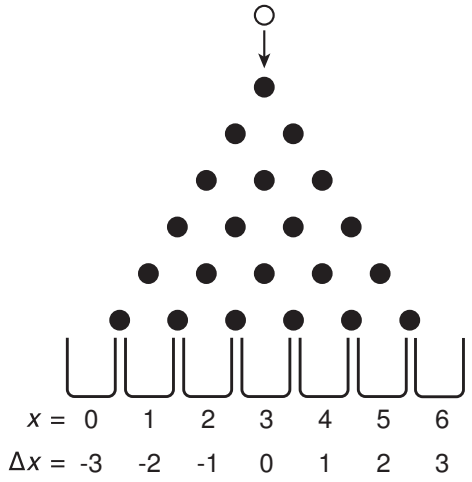
- For a coin toss, we could assign a random variable, x, the value of 1 for heads and 0 for tails. Or, just as arbitrarily, we could define $x = 0$ for heads and $x = 1$ for tails.
- For n successive coin tosses, we could define x to be the number of heads.
- For the plinko, we could define x to be the number associated with the bin that a ball falls in.

For all but the simplest process (*i.e.,* the coin toss), there are likely to be a variety of different random variables that could be defined. For instance, for a series of n coin tosses, we could define the following random variables:

- x_{even}: equal to 1 when the number of heads is even, and 0 when the number of heads is odd.
- x_0: equal to 1 when there are no heads and 0 otherwise.
- x_{run}: equal to the number of successive heads in the longest stretch of heads in the sequence.

Any of these random variables, or many others, could be used for a game of chance, and the trick lies in figuring out how to calculate the probabilities associated with different values for the random variable.

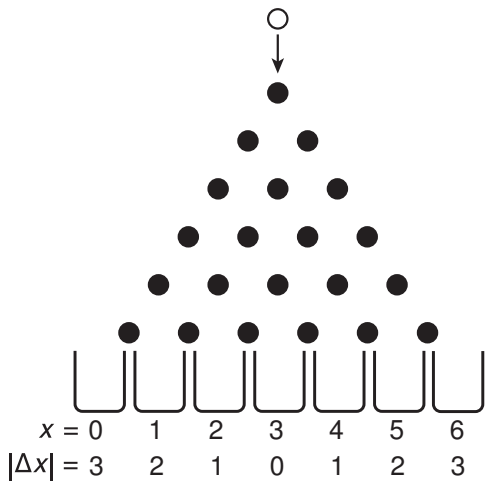
So far for the plinko, we have used the bin number, k as the random variable (without actually calling it that). Now, we will use the variable x to represent the bin number and define some additional random variables in terms of x. For instance, we can define another random variable, Δx, as the position of a bucket from the central bucket. For the six-row plinko the two random variables are related to one another as shown below:



For the six row plinko, the two random variables are related to one another in a simple way:

$$\Delta x = x - 3$$

Another random variable for the plinko is $|\Delta x|$, which represents the distance (in bucket numbers) from the central bucket, as illustrated below:



For the six-row plinko $|\Delta x|$ is related to x according to:

$$|\Delta x| = |x - 3|$$

We could use Δx or $|\Delta x|$ as the basis for gambling. For instance, I could offer to pay players $\$\Delta x$ or $\$|\Delta x|$ when the ball falls in bin x. These games will lead to very different transactions than when I just paid $\$x$. For instance, the negative values of Δx imply that the players will sometimes pay me. This means that I should reconsider how much to charge to play the game, and potential players should reconsider how much they are willing to play.

III. Expected value, or mean of a distribution

The expected value, or *expectation*, for a random variable, x, is the expected average value of x if the process is repeated a large number of times. More formally, consider a process that has m possible outcomes (or a complete set of n non-overlapping events). If the random variable, x, has values of x_i for outcomes $i = 1, 2, 3 \ldots m$, and the probabilities of the outcomes are $p(i)$, then the expected value is defined as:

$$E = \sum_{i=1}^{m} p(i) x_i$$

In essence, E is a weighted sum, in which each of the possible values of x are weighted by the probability of that value of x being the result of a single experiment. If the experiment is repeated a large number of times, we expect the average value of x to approach E. If the experiment represents a game of chance and x is the payout, then E is the expected average payout. This is what we need to calculate how much to charge to play the game.

For the six-row plinko, the possible values of x and the respective probabilities are listed below:

Bucket	x_i	$p(i)$	$p(i)x_i$
0	0	1/64	0
1	1	6/64	6/64
2	2	15/64	30/64
3	3	20/64	60/64
4	4	15/64	60/64
5	5	6/64	30/64
6	6	1/64	6/64
Total		1	192/64 =3

This confirms our earlier deduction that the average payout for the game should be 3.

For the random variable Δx the expected value is quite different, because there are both negative and positive values:

Bucket	Δx_i	$p(i)$	$p(i)\Delta x_i$
0	-3	1/64	-3/64
1	-2	6/64	-12/64
2	-1	15/64	-15/64
3	0	20/64	0
4	1	15/64	15/64
5	2	6/64	12/64
6	3	1/64	3/64
Total		1	0

This result makes sense, since we have, in effect, just moved the labels for the buckets to the right by 3. More generally, if x is a random variable and a is a constant, then:

$$E(x + a) = E(x) + a$$

Also,

$$E(ax) = aE(x)$$

If x and y are two random variables that describe the same set of events, like x and Δx, then

$$E(x + y) = E(x) + E(y)$$

More generally, if x and y are random variables describing the same set of events, and a and b are constants, then

$$E(ax + by) = aE(x) + bE(y)$$

On the other hand, the following relationship is *not* true:

$$E(xy) = E(x)E(y)$$

To help convince yourself of the validity of these relationships (except the last) it may be helpful to test them using x and Δx for the six-row plinko. But, this does not constitute a proof! For that you need to go back to the definition of the expected value.

Finally, consider the case for the random variable $|\Delta x|$

Bucket	$\lvert\Delta x\rvert_i$	$p(i)$	$p(i)\,\lvert\Delta x\rvert_i$
0	3	1/64	3/64
1	2	6/64	12/64
2	1	15/64	15/64
3	0	20/64	0
4	1	15/64	15/64
5	2	6/64	12/64
6	3	1/64	3/64
Total		1	15/16

This random variable has yet another expected value for the same experiment. Here, it is clear that a relationship that you might think would be true,

$$E(|\Delta X|) = |E(x)|$$

is not.

IV. The variance

Another important parameter that helps define a random variable is called the *variance*. For a discrete random variable, x, with m possible values, the variance, σ^2. is defined as:

$$\sigma^2 = \sum_{i=1}^{m} p(i)(x_i - \mu)^2$$

where μ represents the mean, or expected value, of the random variable. In brief, σ^2 is a measure of the breadth of the distribution. The closer all of the possible values of x are to the mean, the smaller the variance. Of equal importance, if the values of x close to μ are the more probable values, the smaller σ^2 will be. Notice that all of the differences in the sum are squared, which ensures that all of the terms are positive, and values above and below the mean are treated equally.

For the six-row plinko, with the random variable, x, defined as the original bucket number, the variance can be calculated as summarized in the table below. (Recall that the mean value of x is 3.)

Bucket	x_i	$p(i)$	$(x_i - \mu)^2$	$p(i)(x_i - \mu)^2$
0	0	1/64	9	9/64
1	1	6/64	4	24/64
2	2	15/64	1	15/64
3	3	20/64	0	0
4	4	15/64	1	15/64
5	5	6/64	4	24/64
6	6	1/64	9	9/64
Total		1	28	1.5

Thus, $\sigma^2 = 1.5$ for this particular random variable. It is rather difficult to compare directly the magnitude of the variance to values of the random variable, because the differences making up the variance have been squared. Although this particular random variable doesn't have units, random variables can have units in other cases. In such cases, the units of the variance are those of the random variable squared. For instance, if the random variable has units of meters, m, the the units of the variance will be m^2. For this reason, another parameter, the standard deviation, σ, is often used and is defined as:

$$\sigma = \sqrt{\sigma^2}$$

Defining things this way, may seem a bit convoluted, but it emphasizes the fact that the variance is defined in terms of a sum of squares, and is positive, and the standard deviation is derived from σ^2, and not the other way around. For the random variable x defined for the 6-row plinko, $\sigma = \sqrt{(1.5)} = 1.225$.

Next, consider the random variable Δx introduced earlier. for this random variable, the mean, μ, is 0.

Bucket	Δx_i	$p(i)$	$(\Delta x_i - \mu)^2$	$p(i)(x_i - \mu)^2$
0	-3	1/64	9	9/64
1	-2	6/64	4	24/64
2	-1	15/64	1	15/64
3	0	20/64	0	0
4	1	15/64	1	15/64
5	2	6/64	4	24/64
6	3	1/64	9	9/64
Total		1	28	1.5

Perhaps surprisingly, the variances for x and Δx are the same. While the means are different, the distribution of values *around* the respective means are the same.

For a binomial distribution defined previously as

$$p(k; n, p_\mathrm{s}) = \frac{n!}{k!(n-k)!} p_\mathrm{s}^k (1 - p_\mathrm{s})^{n-k}$$

where n is the number of trials, k is the number of successes, and p_s is the probability of a single succesful trial, the mean, variance and standard deviation for the random variable, $x = k$, are given by:

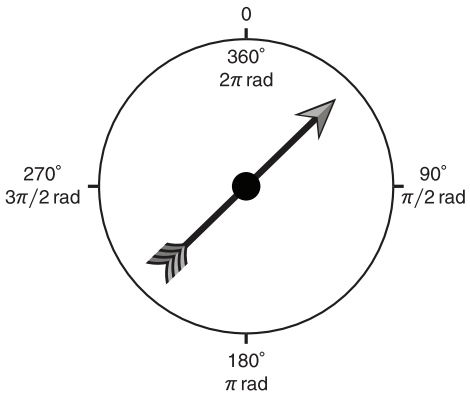
$$\mu = np_\mathrm{s}$$
$$\sigma^2 = np_\mathrm{s}(1 - p_\mathrm{s})$$
$$\sigma = \sqrt{np_\mathrm{s}(1 - p_\mathrm{s})}$$

Thus, the variance increases in proportion to the number of trials in the binomial experiment. This relationship also indicates that if p_s is closer to 1 or 0, the variance decreases. That is, the more biased the individual trials are, the narrower the distribution. For instance, for the distributions shown on page 45, for a 10-row plinko with $p_\mathrm{s} = 0.5$, 0.6 and 0.8, the corresponding values of the variance are 2.5, 2.4 and 1.6.

2.8 Continuous probability distribution functions

I. The spinner

So far, we have limited our discussion of probability to processes with discrete outcomes, such as coin tosses, dice or the plinko. But, many of the most interesting biological and physical processes give rise to a continuous range of possible outcomes. As a simple example of a process with continuous outcomes, consider a spinner, as used in some board games, consisting of a pointer mounted on a board with a bearing that allows it to spin freely after being given a sharp push, as with a flick of a finger, as illustrated below:



A short time after being pushed, the pointer slows down and stops, pointing in a particular direction. Assuming that the pointer is well balanced and the bearing is very smooth, the pointer should be equally likely to point in any directions. For the purposes of our discussion, we will define the direction of the pointer as the angle between the vertical, as drawn above, and the pointer.

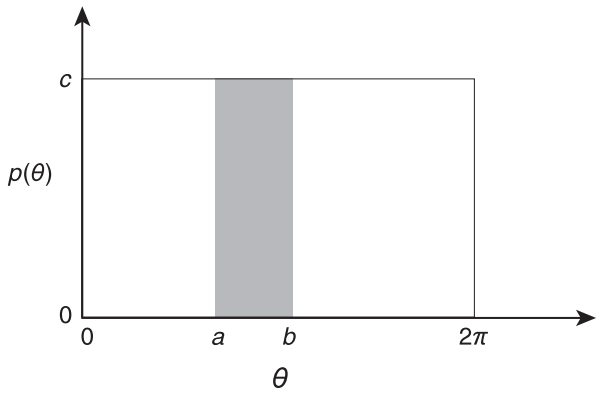
A spinner like this can be used to generate a variety of different random variables. For instance we could divide up the range of angles into two regions; from 0 to π radians and from π to 2π radians. We could then define a random variable so that it is 0 if the pointer lands in the 0–π region and 1 if it lands in the π–2π range. This would be equivalent to a coin toss with a fair coin. We could also divide the range into two non-equal ranges to simulate a biased coin toss. Alternatively, we could divide the range into six regions, to simulate a single six-sided die.

In principle, we can divide the range of angles into smaller and smaller angles, provided that we have a means to measure very small differences in position. This leads to the notion of a continuous random variable that represents the position of the pointer, as shown in the drawing above. We will call this random variable θ and express its value in radians, from 0 to 2π. Like other random variables we have discussed, every possible value θ has associated with it a probability, $p(\theta)$. Since the pointer is equally likely to point in any direction, the value of $p(\theta)$ must be equal for all values of θ. On the other hand, if θ can take on any value in a continuous range, which can be divided into infinitesimally small intervals, then the probability of pointing any single direction must be infinitesimally small, or essentially zero! To resolve this apparent paradox, we

interpret continuous probability distribution functions in terms of the probability that the random variable (θ in our case) lies between two defined values. Specifically, the probability that the random variable θ lies between a and b is given by the integral:

$$p(a \leq \theta \leq b) = \int_a^b p(\theta) d\theta$$

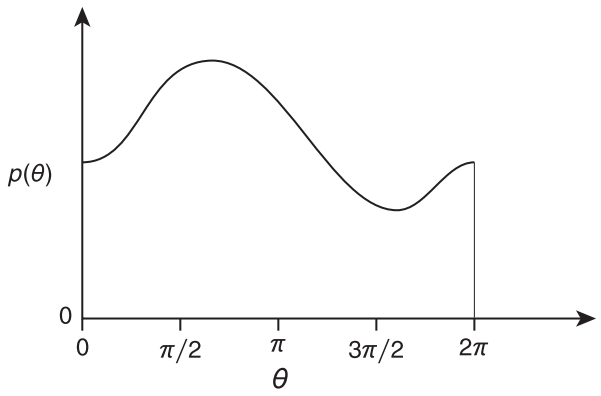
This relationship is illustrated graphically below, for the case of the spinner:



As argued above, the value of $p(\theta)$ must be equal for all values of θ between 0 and 2π. For now, we will call the value of $p(\theta)$ within this range c. Since the pointer must land within the range between 0 and 2π, $p(\theta)$ must be zero elsewhere. The probability that θ lies between a and b is then:

$$\begin{aligned} p(a \leq \theta \leq b) &= \int_a^b p(\theta) d\theta \\ &= \int_a^b c d\theta \end{aligned}$$

This integral corresponds to the area below the horizontal line segment representing $p(\theta) = c$ and bounded by $\theta = a$ and $\theta = b$, as indicated by the shaded box in the drawing above. Assuming that the probability function is, indeed, constant, then the probability is proportional to the difference between a and b, as we would intuitively expect if the spinner is fair. A continuous probability distribution function for which all possible values of the random variable are equal is called a *uniform* probability distribution. A more interesting probability distribution might arise if the pointer was more likely to land in some areas than others, as illustrated in the hypothetical graph below:



In this case, the probability distribution function indicates that the pointer is more likely to land in the region of $\theta \approx 3\pi/4$ than in the region of $3\pi/2$. If this spinner was used in a game of chance, a gambler with this information would be at a distinct advantage over one without it!

Since the spinner must point *somewhere* in the range between $\theta = 0$ and 2π (assuming that it doesn't break), the total probability must be 1. This is equivalent to the requirement that the probabilities of all of the possible outcomes must sum to 1 in a random process with discrete outcomes. For the uniform distribution function for the spinner, we can write this requirement as:

$$\begin{aligned} p(0 \leq \theta \leq 2\pi) = 1 &= \int_0^{2\pi} p(\theta) d\theta \\ &= \int_0^{2\pi} c d\theta \end{aligned}$$

where c is the constant introduced earlier, to which we can now assign a specific value, as follows:

$$\begin{aligned} 1 &= \int_0^{2\pi} c d\theta \\ 1 &= c\theta|_0^{2\pi} = c2\pi - c0 = c2\pi \\ c &= 1/(2\pi) \end{aligned}$$

We can then write the probability distribution function as:

$$p(\theta) = \frac{1}{2\pi}$$

This form of the function is said to be *normalized*, meaning that the integral over all possible values is equal to 1. This term is sometimes confused with the *normal probability function* which refers to a specific continuous probability function that we will discuss bellow and also goes by the name *Gaussian* distribution.

II. Expected value and variance for continuous random variables

Recall that for a discrete random variable, x, we defined the expected value, $E(x)$, as

$$E(x) = \sum_{k=1}^{n} p(x_k)x_k$$

where x_k represents the k^{th} value of x and $p(k)$ is the probability of x_k. For a continuous random variable, the sum above is replaced by an integral:

$$E(x) = \int p(x)x dx$$

where the integral is over all possible values of x. For the spinner random variable, θ, the expected value is calculated as follows (for an unbiased spinner):

$$\begin{aligned} E(\theta) &= \int_0^{2\pi} p(\theta)\theta d\theta \\ &= \int_0^{2\pi} \frac{1}{2\pi}\theta d\theta \\ &= \frac{1}{4\pi}\theta^2\Big|_0^{2\pi} = \frac{4\pi^2}{4\pi} - 0 \\ &= \pi \end{aligned}$$

Thus, the average value of θ, over a large number of trials, is expected to be π, that is the mid-point of the range of possible values. Keep in mind, however, that this outcome is no more likely than any other.

For a discrete random variable the variance is defined as the sum:

$$\sigma^2 = \sum_{k=1}^{n} p(k)(k-\mu)^2$$

where μ represents the mean, or expected value, of the random variable.

The equivalent relationship for a continuous random variable is the integral:

$$\sigma^2 = \int p(x)(x-\mu)^2 dx$$

For the random variable θ, the variance is calculated as:

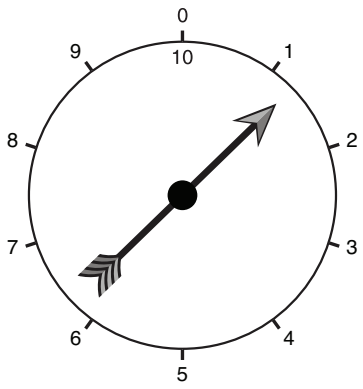
$$\begin{aligned}
\sigma^2 &= \int_0^{2\pi} p(\theta)(\theta - \pi)^2 d\theta \\
&= \int_0^{2\pi} \frac{1}{2\pi} \left(\theta^2 - 2\pi\theta + \pi^2\right) d\theta \\
&= \frac{1}{2\pi} \left(\frac{1}{3}\theta^3 - \pi\theta^2 + \pi^2\theta\right)\Bigg|_0^{2\pi} \\
&= \frac{1}{2\pi} \left(\frac{8}{3}\pi^3 - 4\pi^3 + 2\pi^3\right) \\
&= \frac{4}{3}\pi^2 - 2\pi^2 + \pi^2 \\
&= \frac{\pi^2}{3}
\end{aligned}$$

III. Some other random variables from the spinner

A variety of other random variables might be assigned to the spinner. For instance, we could create a game of chance where the payout ranges from zero to \$10 depending on the position of the pointer, with the payout increasing linearly with θ, from 0 to \$10. We will call this random variable x and define it in terms of θ according to:

$$x(\theta) = \frac{5}{\pi}\theta$$

This amounts to relabeling the spinner, as below:



Since the values of x are evenly distributed over the range of θ, which has a uniform probability distribution function, x should also have a uniform probability distribution. Following the approach used for θ, you should be able to show that $p(x) = 1/10$.

The expected value of x is calculated as

$$\begin{aligned} E(x) &= \int_0^{10} p(x)x dx \\ &= \int_0^{10} \frac{1}{10} x dx \\ &= \frac{1}{20}x^2\Big|_0^{10} = 5 \end{aligned}$$

This result can derived even more easily by using a relationship introduced on page 53:

$$E(ax) = aE(x)$$

where a is a constant. (Note that x in this equation is not the same as our $x(\theta)$.) Substituting θ for x and $5/\pi$ for a, we can write:

$$\begin{aligned} E(\frac{5}{\pi}\theta) &= \frac{5}{\pi}E(\theta) \\ &= \frac{5}{\pi}\pi = 5 \end{aligned}$$

Note that, just as for θ, the expected value for x is the midpoint in the range of possible values. This is a general property of a uniform probability distribution function, but not of all distribution functions.

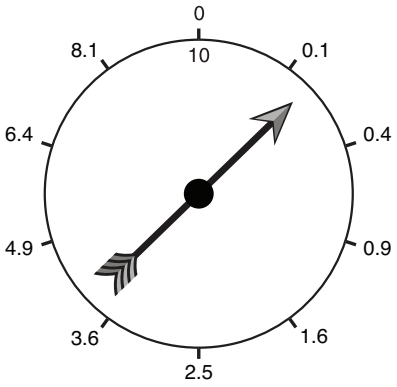
The variance of the new random variable is calculated as:

$$\begin{aligned} \sigma^2 &= \int_0^{10} p(x)(x-5)^2 dx \\ &= \int_0^{10} \frac{1}{10}\left(x^2 - 10x + 25\right) dx \\ &= \frac{1}{10}\left(\frac{1}{3}x^3 - 5x^2 + 25x\right)\Big|_0^{10} \\ &= \frac{1}{10}\left(\frac{1}{3}1000 - 500 + 250\right) \\ &= \frac{25}{3} \approx 8.333 \end{aligned}$$

As an example of a random variable that does not have a uniform distribution function, but is still based on the spinner, we can define y as:

$$y(\theta) = \frac{10}{4\pi^2}\theta^2$$

As in the previous example, a constant of multiplication has been introduced to make the range of possible values lie between 0 and 10. We can use this definition to relabel the spinner dial:



Now, we find that the values are not evenly distributed around the dial. For instance the range of y-values from 0 to 2.5 represents half of the dial, meaning that values in this range are expected to occur half the time, and values from 2.5–10 are expected the other half.

To derive the probability distribution function for y, we can use its relationship to θ, for which we do know the distribution function. For y, the expression $p(y)dy$ represents the probability that y lies within a small region, dy, of a specific value of y. Similarly, the expression $p(\theta)d\theta$ represents the probability that θ lies within $d\theta$ of θ. If y is $y(\theta)$ for a specific value of θ and dy is the small region of y corresponding to the small region of θ, $d\theta$, then the two probabilities must be equal, and we can write:

$$p(y)dy = p(\theta)d\theta$$

Taking some mathematical liberties, this can be rewritten in terms of the derivative, $d\theta/dy$:

$$p(y) = \frac{d\theta}{dy}p(\theta)$$

To find the derivative, $d\theta/dy$, we need the function $\theta(y)$, which can be obtained by rearranging the definition of y as a function of θ:

$$y = \frac{10}{4\pi^2}\theta^2$$

$$\theta^2 = \frac{4\pi^2}{10}y$$

$$\theta = \frac{2\pi}{\sqrt{10}}y^{1/2}$$

The derivative of θ with respect to y is:

$$\frac{d\theta}{dy} = \frac{2\pi}{\sqrt{10}}\frac{1}{2}y^{-1/2} = \frac{\pi}{\sqrt{10}}y^{-1/2}$$

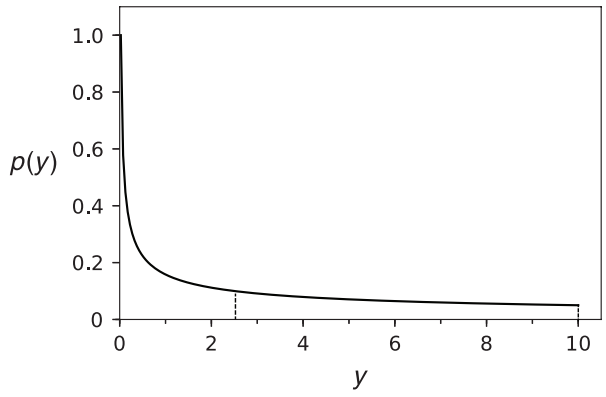
Recalling that $p(\theta) = 1/(2\pi)$, the desired probability function, $p(y)$, can now be written as:

$$\begin{aligned} p(y) &= \frac{d\theta}{dy} p(\theta) \\ &= \frac{1}{2\pi} \frac{\pi}{\sqrt{10}} y^{-1/2} = \frac{1}{2\sqrt{10}} y^{-1/2} \end{aligned}$$

Since $p(y)$ was derived from a normalized probability density function, $p(\theta)$, $p(y)$ should be normalized as well. To be sure, though, we can calculate the integral of $p(y)$ from 0 to 10.

$$\begin{aligned} \int_0^{10} p(y)dy &= \int_0^{10} \frac{1}{2\sqrt{10}} y^{-1/2} dy \\ &= \frac{1}{\sqrt{10}} y^{1/2} \Big|_0^{10} \\ &= \frac{1}{\sqrt{10}} \left(10^{1/2} - 0^{1/2}\right) = 1 \end{aligned}$$

Thus, $p(y)$ is, indeed normalized. A plot of $p(y)$ is shown below:



Note that $p(y)$ is not defined for $y = 0$, but it can be evaluated for any value of y arbitrarily close to 0. As expected from the relabeled spinner dial shown on page 63, the distribution favors smaller values of y. For instance, one half of the area under the curve lies between 0 and 2.5, and the other half lies between 2.5 and 10, as indicated by the vertical dashed lines.

Calculating the expected value of y is a bit more involved than in the previous examples,

because the probability function is not a simple constant:

$$\begin{aligned}
E(y) &= \int_0^{10} p(y) y dy \\
&= \int_0^{10} \frac{1}{2\sqrt{10}} y^{-1/2} y dy = \int_0^{10} \frac{1}{2\sqrt{10}} y^{1/2} dy \\
&= \frac{1}{2\sqrt{10}} \frac{2}{3} y^{3/2} \Big|_0^{10} \\
&= \frac{1}{3\sqrt{10}} \left(10^{3/2} - 0^{3/2}\right) \\
&= \frac{10}{3} \approx 3.333
\end{aligned}$$

Recall that the expected value of x, which also covers the range from 0 to 10, is 5. The lower value of $E(y)$ reflects the non-uniform probability distribution function for this variable, which more heavily favors lower values.

The variance is calculated as:

$$\begin{aligned}
\sigma^2 &= \int_0^{10} p(y)(y - 10/3)^2 dy \\
&= \int_0^{10} \frac{1}{2\sqrt{10}} y^{-1/2} \left(y^2 - \frac{20}{3} y + \frac{100}{9}\right) dx \\
&= \frac{1}{2\sqrt{10}} \int_0^{10} \left(y^{3/2} - \frac{20}{3} y^{1/2} + \frac{100}{9} y^{-1/2}\right) dy \\
&= \frac{1}{2\sqrt{10}} \left(\frac{2}{5} y^{5/2} - \frac{40}{9} y^{3/2} + \frac{200}{9} y^{1/2}\right) \Big|_0^{10} \\
&= \frac{1}{2\sqrt{10}} \left(\frac{2}{5} 10^{5/2} - \frac{40}{9} 10^{3/2} + \frac{200}{9} 10^{1/2}\right) \\
&= \frac{80}{9} \approx 8.888
\end{aligned}$$

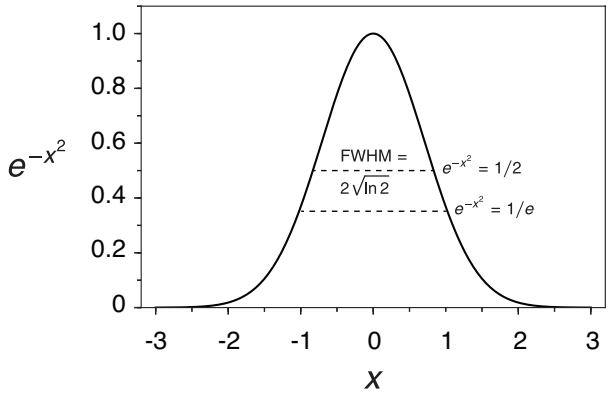
Notice that the variance is just a bit larger than for the random variable x, which has a uniform probability distribution function.

2.9 The Gaussian, or normal, probability distribution function

One of the most important continuous probability distribution functions is commonly referred to as a *Gaussian* or *normal* distribution function. The Gaussian function, in various forms, also arises in areas outstide of probability and statistics. At it's simplest, a Gaussian function has the form:

$$f(x) = e^{-x^2}$$

where $e \approx 2.71828$ is the base of the natural logarithms. A graph of the function has the familiar bell shape shown below:



The function has its maximum value, 1, when the exponent is 0 and decreases as x becomes either positive or negative. When $x = 1$ or -1, the function equals $1/e \approx 0.3679$. Another useful parameter for describing functions that describe peaks is the *full width at half maximum*, FWHM. For the simple Gaussian function, it is easy to show that FWHM $= 2\sqrt{\ln 2}$.

The simple Gaussian function, and forms derived from it, have the rather inconvenient property that its antiderivative cannot be written in terms of a finite number of simple functions. As a consequence, integrals over finite ranges of x cannot be evaluated exactly. However, the indefinite integral over all values of x can be shown to be:

$$\int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

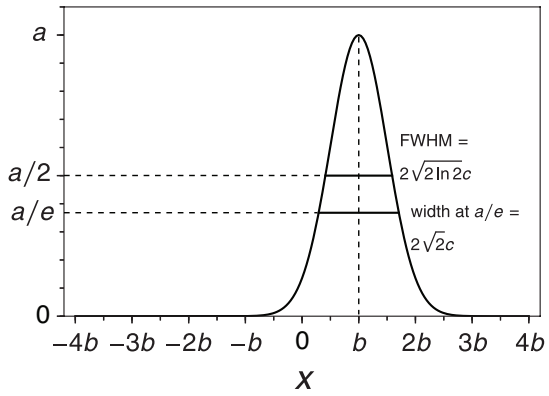
Thus, this form is not a properly normalized probability distribution function. It is also striking that integration of a function defined in terms of an important irrational number, e, is related to a second fundamental irrational number, π.

I. The general form of the Gaussian function

A more general form of the Gaussian function can be written as:

$$f(x) = ae^{-\frac{(x-b)^2}{2c^2}}$$

This form introduces three parameters, a, b and c, which affect the shape of the curve in different ways, as illustrated in the figure below.



The parameter a determines the value of the function at its maximum, where the exponent of e equals zero, which occurs when the value of x is equal to b. The width of the peak is determined by c^2: The larger the value of c^2, the more slowly the exponent decreases as x increases or decreases away from b, and the wider the peak is. As shown in the figure, both the full width at half maximum ($FWHM$) and the width at the maximum value divided by e (a/e) are proportional to c (which is assumed here to be positive).

The integral of the general form of the Gaussian function is:

$$\int_{-\infty}^{\infty} ae^{-\frac{(x-b)^2}{2c^2}}dx = a\sqrt{2c^2\pi}$$

If a is set so that it is equal to $1/\sqrt{2c^2\pi}$, then the value of the integral is equal to one:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2c^2\pi}}e^{-\frac{(x-b)^2}{2c^2}}dx = 1$$

The function thus has the required property of a normalized probability distribution function. However, the form that is usually used in probability and statistics uses the symbol μ in place of b and σ^2 in place of c^2, to give:

$$p(x) = \frac{1}{\sqrt{\sigma^2 2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

As with the other continuous probability distribution functions we have looked at, the (normalized) Gaussian distribution gives the probability that the variable x lies between two points, x_1 and x_2, when the function is integrated between the two points.

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\sigma^2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$$

In this context, μ is the mean value (or expectation value) of the distribution and σ^2 is the variance, as defined earlier.

II. Approximation of the binomial distribution by the Gaussian distribution

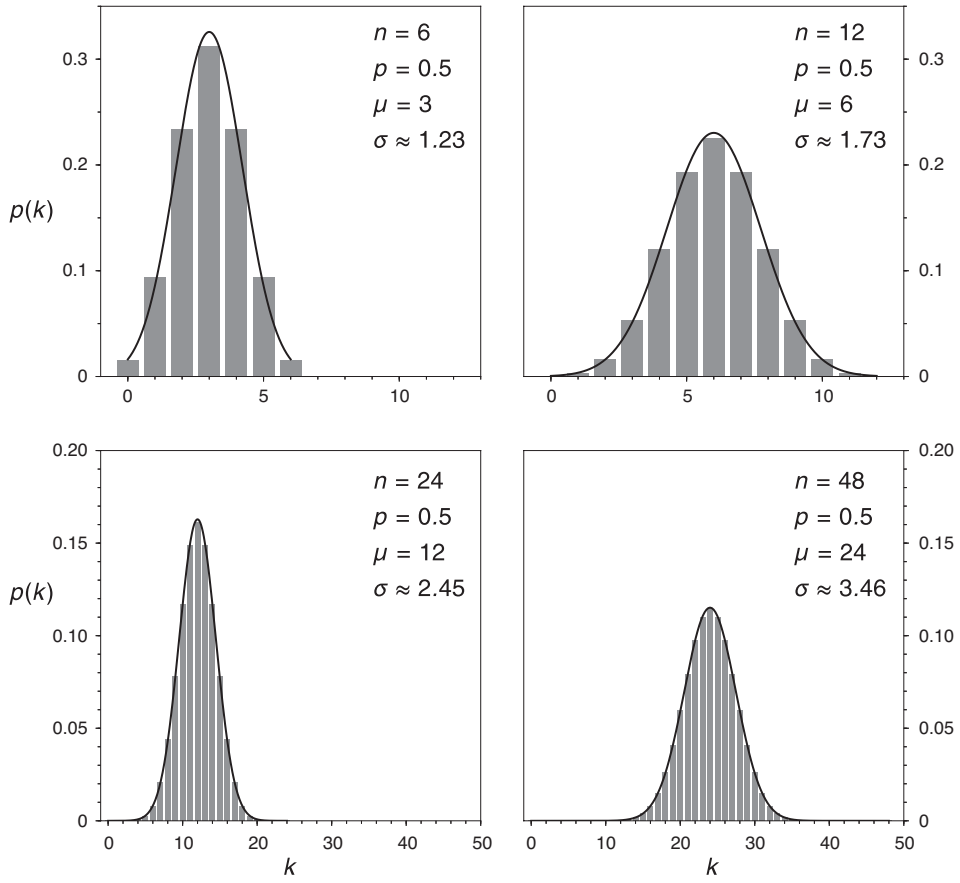
One important feature of the Gaussian (or normal) probability distribution function is that it represents the limiting case of the discrete binomial distribution when the number of trials (rows in the plinko, for instance) becomes large. Unfortunately, rigorously demonstrating this relationship is not simple. Instead, we will simply demonstrate how closely the two relationships match one another as the number of trials increases. Recall that the binomial distribution is given by

$$p(k; n, p_s) = \frac{n!}{k!(n-k)!} p_s^k (1-p_s)^{n-k}$$

where n is the number of trials, k is the number of successes, and p_s is the probability of a single succesful trial. Recall also that the expected value, or mean, of the binomial distribution is np_s, and the variance is $np_s(1-p_s)$. For given values of n and p_s, the Gaussian distribution with the same mean and variance can be written as:

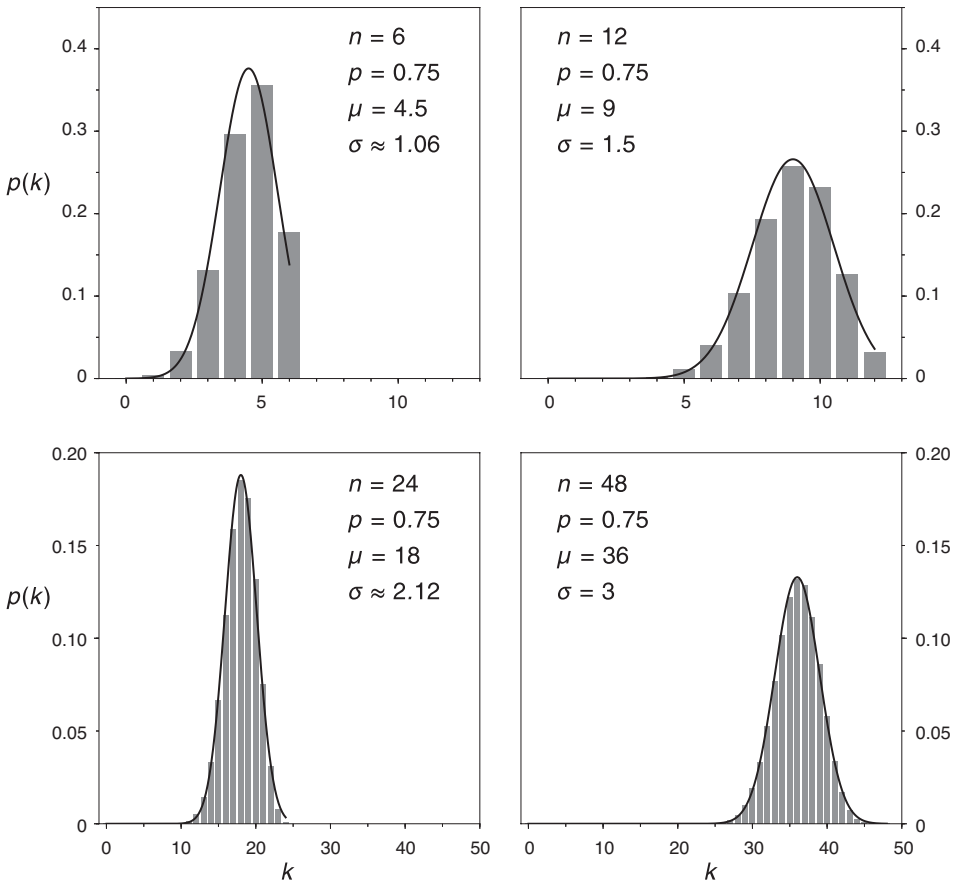
$$\begin{aligned} p(k) &= \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{np_s(1-p_s)2\pi}} e^{-\frac{(k-np_s)^2}{2np_s(1-p_s)}} \end{aligned}$$

The figure below shows direct comparisons between the binomial and matched Gaussian distribution for $p_s = 0.5$ and $n = 6$, 12, 24, and 48.



As you can see, the Gaussian distribution is a close match to the binomial distribution, even for n as small as 6. For even moderately large values of n, it is much easier to calculate values for the Gaussian distribution than for the binomial distribution, where the values of the coefficients quickly become very large.

The figure below shows the same comparisons between the binomial and Gaussian distributions, but now with $p_s = 0.75$, so that the distributions are shifted to the right. With the biased distributions, the match between the binomial and Gaussian distributions is not quite as close as when $p_s = 0.5$ The reason for this is that the Gaussian distribution is always symmetrical about the mean, even if the mean is shifted from the unbiased value. On the other hand, the binomial distribution becomes skewed when p_s is not equal to 0.5, especially for relatively small values of n. As n increases, the binomial distribution becomes more symmetrical, for a given value of p_s, and is better matched by the Gaussian distribution.



A general rule of thumb[1] states that the Gaussian distribution, with μ set to np_s and σ^2 set to $np_s(1 - ps)$ is a "good" approximation to the binomial if the following two

[1] https://en.wikipedia.org/wiki/Binomial_distribution#Normal_approximation

conditions are satisfied:

$$n > 9\frac{1 - p_s}{p_s}$$

and

$$n > 9\frac{p_s}{1 - p_s}$$

Do the examples shown above appear to confirm this rule of thumb?

2.10 Simulating randomness with a computer: (Pseudo) random numbers

One of the most interesting, and useful, applications of computers in science is the simulation of processes that have an underlying random, or unpredictable, character. Such processes include the diffusion of particles, mutations of genes and quantum-mechanical phenomena. The basic idea of such simulations is to figuratively flip a coin, throw dice or spin a roulette wheel to decide the outcome of specific events in a simulation. The simulation is usually repeated many times in order to describe the distribution of possible outcomes. The technique is usually attributed to two mathematicians, John Von Neumann and Stansilaw Ulam, who used it to study nuclear physics problems near the end of World War II and attached the name "Monte Carlo" to this kind of calculation. (Presumably, they thought that Monte Carlo sounded more glamorous than Wendover.)

Although they may not always seem it, computers are, by design, extremely predictable machines. So, the problem arises, how do we simulate a random event, like a coin flip? The answer is to generate a sequence of numbers using a completely predictable algorithm, that *appears* to have come from a random physical process. These numbers are properly called "pseudo-random numbers", but the shortened term "random number" is often used. For instance, a simulation of a series of coin tosses might be represented as a series of 1s and 0s. Over a long period, the number of 1s and 0s should be roughly, but not exactly, equal. But, the sequence should not be a simple alteration of 1s and 0s, since we know that "runs" of "heads" and "tails" are common. More generally, each number should be unpredictable from the previous one, *unless* one knows the algorithm. Very demanding tests for random number generators have been devised, and the development of improved generators is, itself, an ongoing endeavor.

A fairly simple approach to generating random numbers involves taking a "seed" value, applying some arithmetic operation to it, dividing the result by some other number and returning the remainder. The result is then used as the seed for calculating the next number in the series. One widely used algorithm uses the following equation to calculate number X_{n+1} from X_n:

$$X_{n+1} = (aX_n + b) \mod c$$

where a, b and c are integers, and the operator $\mod c$ indicates the remainder of dividing the quantity $(aX_n + b)$ by c. For instance:

$$5 \mod 4 = 1$$

The remainder has to be less than the value chosen for `c`, so this establishes a maximum number of unique numbers that can be generated. Eventually, a number will be returned a second time, and from that point on the series repeats exactly. How well this algorithm works depends critically on the choice of constants.

Many computer languages include built in functions for generating random numbers. For instance, the Python language includes a module, called `random`, that provides a much better random number generator than the one described above, as well as several nifty variations for special purposes. The function `random.random()` returns a number, x, such that $0 \leq x < 1$, ss illustrated below:

```
>>> random.random()
0.63876690995825403
>>> random.random()
0.98645481541390223
```

In general, the initial seed for a random number generator can be either set to a specific value derived from information provided by the computer or some outside source. A common way of setting the seed is to derive it from the time when the program is started, using the computer's clock.

The Python random module includes a function that allows the user to specify the seed. The listing below shows what happens if the same seed is used a second time:

```
>>> random.seed(12345)
>>> random.random()
0.41661987254534116
>>> random.random()
0.010169169457068361
>>> random.random()
0.82520650925374317
>>> random.seed(12345)
>>> random.random()
0.41661987254534116
>>> random.random()
0.010169169457068361
```

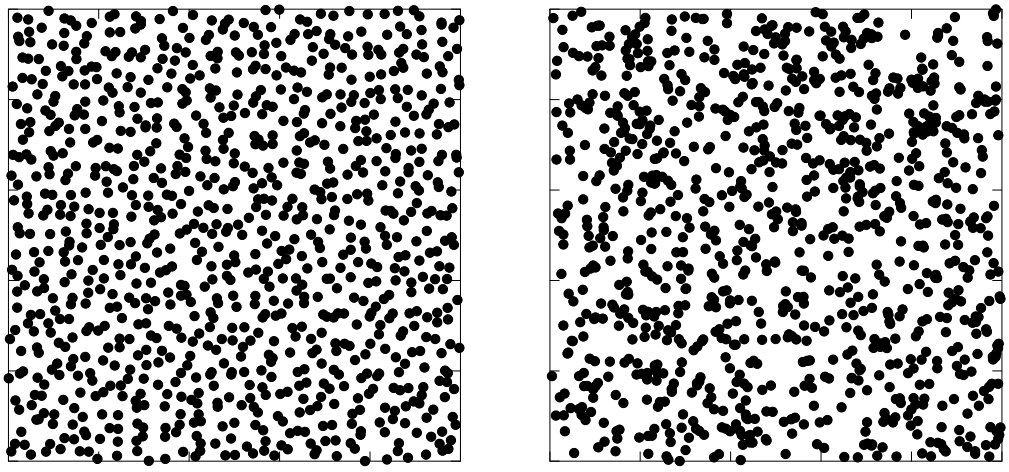
For a given seed, the same sequence of "random" numbers will be generated again. This is a clear demonstration that the numbers generated from most random number generators aren't truly "random." Sometimes, though, it is useful to be able to use the same set of random numbers multiple times, for instance in testing a computer program or algorithm.

In addition to pseudo-random number generators, there are ways to generate "true" random numbers from physical processes. One of these involves monitoring the decay of a radioactive element. Although the average number of decay events over a period of time can be well known, the intervals between successive events is random. A website that provides

"Hot bits" derived in this fashion is:
https://www.fourmilab.ch/hotbits/
The hardware used at this site can only generate numbers at a modest rate, but they can be used as the seeds for a pseudo-random number generator. At one time, there was a website that generated random numbers from images of a lava lamp, as described on Wikipedia:
https://en.wikipedia.org/wiki/Lavarand
Again, the idea was to generate true random numbers that could be used as seeds for pseudo-random number generators.

More recently, a variety of hardware devices have been developed that generate true random numbers from various forms of electronic noise or quantum mechanical phenomena. Some of these are relatively inexpensive USB dongles and some are built into newer computers, including those using newer Intel microprocessor chips. It may seem surprising that there would be so much demand for something random! But the reason for this demand is that random numbers play a central role in cryptography, including securing data that is transmitted over the internet. As the security problems of the modern age grow, random number generators are becoming increasingly critical and are coming under ever more careful scrutiny.

One of the useful things that we can do with a random number generator is to simulate some process and look at the distribution, just to get a sense of what a truly random process looks like. The figure below shows two distributions of points on a square:



For one of these figures, I choose random x and y values for 1,000 points and plotted them. For the other, I placed a similar number of points using another procedure.

Which one is the true random distribution? How could you decide?

Chapter 3

Random Walks

We have now spent a lot of time looking at "plinko probabilities", and you should have a good feel for how bell curves arise and how to calculate the probabilities of different outcomes in a binomial distribution. Now, we want to start talking about random walks and the ways in which they arise in physical and biological contexts.

Although the binomial distribution can, in principle, be used to describe a random walk in one dimension, actually using this function for large numbers of steps quickly becomes problematic. Calculating $n!$ requires $n-1$ multiplications, and the magnitudes of the numbers quickly become difficult to handle.. Also, we need to move beyond one dimension. So, we need some other mathematical approaches.

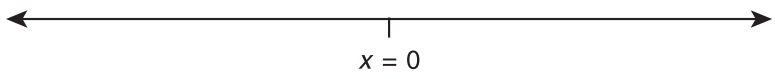
We will start, however, by considering a one-dimensional random walk.

3.1 Random walks in one dimension

In the simplest version of a random walk, we imagine an individual standing on a sidewalk and flipping a coin. If the coin lands heads-up, she turns in one direction and takes a step of length l. If the coin lands tails-up, she turns the opposite direction and takes a step of length l. She then repeats this process another $n-1$ times, for a total of n steps in the random walk.

I. The final position of the walker.

We will define the walker's position as x, which is zero at the beginning of the random walk. As the walker takes steps in the opposite directions, the value of x can take on positive or negative values, as illustrated by the single coordinate axis drawn below:



We will call the position after i steps, x_i, and the final position, after n steps, is x_n. At the outset, we can assume a few things abut the value of x_n, irrespective of whether the coin is fair or not:

- The maximum possible value of x_n is nl
- The minimum possible value of x_n is $-nl$
- Assuming that n is very large, the probability of a walk ending at either nl or $-nl$ is very small, since either outcome would require the coin to land the same way for each toss.

- If a large number of random walks are carried out the distribution of x_n should be related to a binomial distribution.

As a first step in analyzing the random walk in one dimension, we will calculate the expected value of x_n, that is the expected average value of x_n if a large number of random walks, each of n steps, is executed. Here and through out the discussion of random walks, we will call the number of steps in an individual random walk n, and the number of random walks, as used for calculating averages, N.

For each random walk, the final position is given by:

$$x_n = \sum_{i=1}^{n} \delta_i$$

where i is the step number, and δ_i is the change in x in step i. If the step is to the right, $\delta_i = l$, whereas if the step is to the left, $\delta_i = -l$. We will call the probability of an individual step to the right p_+ and the probability of a step to the left p_-.

The expected value of δ_i, for any individual step, is calculated as:

$$\begin{aligned} E(\delta_i) &= lp_+ - lp_- \\ &= lp_+ - l(1 - p_+) \\ &= lp_+ + lp_+ - l \\ &= 2lp_+ - l = l\big(2p_+ - 1\big) \end{aligned}$$

As a quick check, note that if the probability of left and right steps are equal, $p_+ = 0.5$, and the expected value of δ_i is 0.

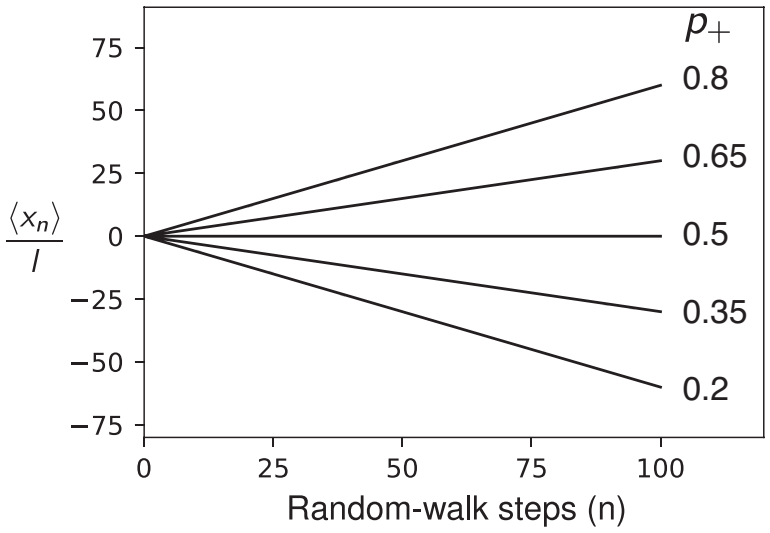
An important theorem from probability states that if x and y are two independent random variables, then the expected value of the sum of x and y is calculated as:

$$E(x + y) = E(x) + E(y)$$

Since x_n is simply the sum of δ_i for each step in the random walk, the expected value of x_n is calculated as:

$$\begin{aligned} E\big(x_n\big) &= \sum_{i=1}^{n} E(\delta_i) \\ &= \sum_{i=1}^{n} l\big(2p_+ - 1\big) \\ &= nl\big(2p_+ - 1\big) \end{aligned}$$

Note that if $p_+ = 0.5$, then the expected value of x_n is zero, that is the average final position is the starting point, irrespective of the number of steps. The plot below shows the expected value of x_n as a function of n for different probabilities of an individual forward step.



As one might expect, values of p_+ greater than 0.5 favor positive values of x_n, and values of p_+ less than 0.5 favor negative values of x_n. Notice also that, for any given value of p_+ (except 0.5), the expected value of x_n increases or decreases linearly with the number of steps.

If the number of individual random walks, N, is large, then the average value of x_n will approach the expected value. Note the distinction between the average value of x_n for N specific random walks and the expected value, $E\big(x_n\big)$ which is calculated from the probabilities of individual forward and reverse steps, as well as the number of steps. Mathematically, we would write the relationship between the two as:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} x_{n,j} = E\big(x_n\big)$$

where the index j indicates the individual random walks included in the average.

To write averages over multiple random walks in a more compact form, we will employ the practice of representing averages using pairs of angle brackets, $\langle\rangle$, as in the example:

$$\langle x_n \rangle = \frac{1}{N} \sum_{j=1}^{N} x_{n,j}$$

Though it is a bit sloppy, we will generally take averages represented in this way to mean that N is large enough that the average approaches the expected value, unless N is otherwise specified.

II. Other averages: The mean-square and root-mean-square

As shown above, it is quite easy to calculate the expected value of x_n for a one-dimensional random walk, even when the probabilities of turns in the two directions are not equal. However, this average provides only limited information. From our study of plinkos, we know that most of the balls don't actually land in the central

bucket (or central two buckets when the number of rows is odd), even though landing in the central bucket(s) is the most probable result.

What we need as a way to represent the distribution of final positions away from the average value of x_n. For this purpose, there are two other kinds of average, which are widely used in a variety of contexts. These are the mean-square and root-mean-square averages, which are defined below, using the angle brackets to represent averages.

The mean-square:

$$\langle x_n^2 \rangle = \frac{1}{N} \sum_{j=1}^{N} x_{n,j}^2$$

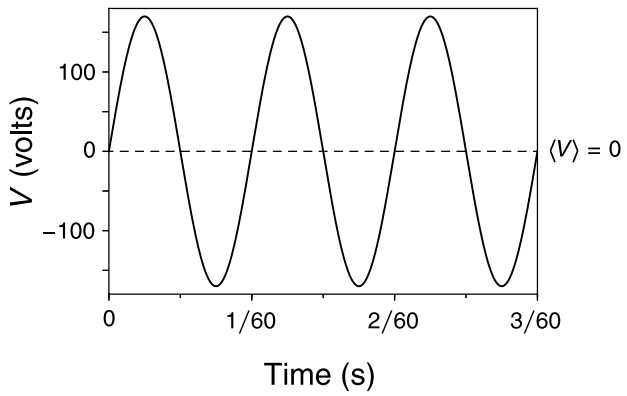
where, as before, the sum is over the N random walks.

The root-mean-square (RMS):

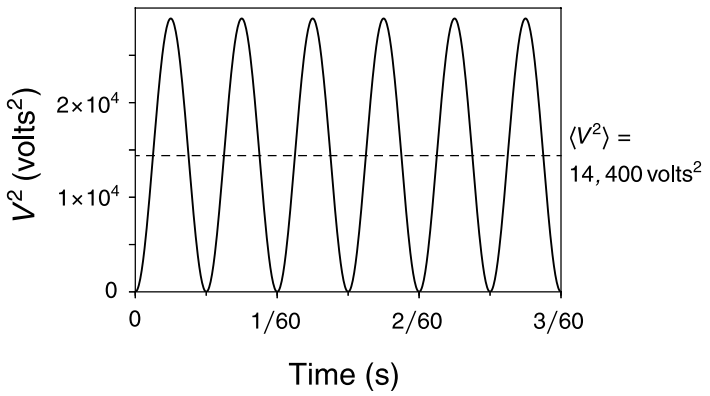
$$\text{RMS}(x_n) = \sqrt{\langle x_n^2 \rangle} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} x_{n,j}^2}$$

By summing over the squares of the final positions, both positive and negative values of x_n make a positive contribution to the averages, rather than canceling out, as when the simple average of x_n is calculated. This goal could be also be obtained by using the absolute value of x_n, but absolute values are more awkward when deriving general results, and the squared quantities have important statistical significance.

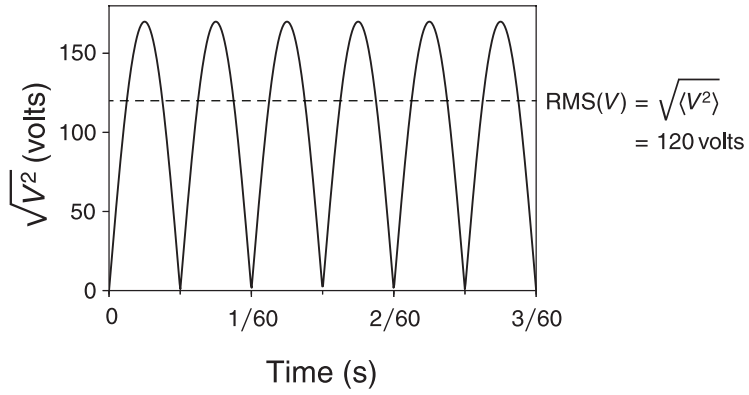
A common application of the mean-square and RMS averages is in electrical engineering, where they are used to treat alternating currents (AC). The graph below shows ideal behavior of voltage as a function of time, for the AC power used in US homes.



Note that the voltage oscillates between a maximum of 170 V and a minimum of −170 V, with the average over time being 0 V. For the US power system, each cycle takes 1/60 s, for a frequency of 60 Hz. Although mathematically correct, the simple average of the voltage over time obviously doesn't convey much information about the power available frome the current. To obtain a positive value, the instantaneous voltage can be squared to generate the plot below.



In this plot, both the maxima and minima in the original plots give rise to peaks of $28900\,\mathrm{V}^2$. The average square voltage over time is $14400\,\mathrm{V}^2$. Although this average is positive and reflects the magnitude of both the positive and negative voltage fluctuations, it has the disadvantage of being expressed in units of V^2, which are not so intuitively interpreted. This is the main reason for introducing the root-mean-square (RMS) average. In the plot below, the instantaneous voltage vales are squared, and the the square root is taken for each point.



Notice that the peaks in the plot have slightly different shapes than those in the plot of V^2, and the peaks have a height of 170 V. The RMS average over time is 120 V, and it is this average that is usually specified for AC circuits. Note that the RMS average is calculated as the square root of the mean-square average and *not* by averaging over the square root of the squares of the individual values, which would, in general, give a different result.

III. The mean-square and RMS end-to-end distance of a one-dimensional random walk

For the reasons discussed above, we would like to have an average that represents distance between the beginning and end of a random walk, calculated in a way that positive and negative steps don't cancel one another. We will begin with the mean-square distance, $\langle x_n^2 \rangle$, which is easier to work with. Once an expression for $\langle x_n^2 \rangle$ is derived, the root-mean-square is calculated by taking the square root.

We start with a definition of the mean-square distance for the random walk:

$$\langle x_n^2 \rangle = \frac{1}{N} \sum_{j=1}^{N} x_{n,j}^2$$

where N is the number of random walks, and the index j represents the individual random walks. For each of the random walks, the final position is given by:

$$x_n = \sum_{i=1}^{n} \delta_i$$

where δ_i is the change in position along the x-axis and can be either $+\delta$ or $-\delta$. Though the reason for doing so may not be obvious yet, we can also write x_n as:

$$x_n = x_{n-1} + \delta_n$$

where δ_n is the change in x in the very last step of the walk. Using this representation, the mean-square distance can be written as:

$$\begin{aligned}
\langle x_n^2 \rangle &= \frac{1}{N} \sum_{j=1}^{N} x_{n,j}^2 \\
&= \frac{1}{N} \sum_{j=1}^{N} \Big(x_{(n-1),j} + \delta_{j,n} \Big)^2 \\
&= \frac{1}{N} \sum_{j=1}^{N} \Big(x_{(n-1),j}^2 + 2 x_{(n-1),j} \delta_{j,n} + \delta_{j,n}^2 \Big)
\end{aligned}$$

where $\delta_{j,n}$ is the change in x for the last step in the j^{th} random walk. This can be broken down into individual sums and averages to give:

$$\begin{aligned}
\langle x_n^2 \rangle &= \frac{1}{N} \sum_{j=1}^{N} x_{(n-1),j}^2 + \frac{1}{N} \sum_{j=1}^{N} 2 x_{(n-1),j} \delta_{j,n} + \frac{1}{N} \sum_{j=1}^{N} \delta_{j,n}^2 \\
&= \langle x_{n-1}^2 \rangle + \langle 2 x_{n-1} \delta_n \rangle + \langle \delta_n^2 \rangle
\end{aligned}$$

As before the angle brackets represent averages over a large number of random walks. For each random walk, the final change in x will be either l or $-l$ and will be uncorrelated with the position, x_{n-1}. If we limit ourselves to the case where the probability of a forward or backward step is equal, the central term in the expression above, $\langle 2 x_{n-1} \delta_n \rangle$, will be zero. Thus, we can write:

$$\langle x_n^2 \rangle = \langle x_{n-1}^2 \rangle + \langle \delta_n^2 \rangle$$

Note that the average of δ_n^2 over all of the random walks is *not* expected to be zero.

Following the same arguments as above, the position of the walker after $n-1$ steps can ve written as:

$$x_{n-1} = x_{n-2} + \delta_{n-1}$$

and the average of x_{n-1} is:

$$\langle x_{n-1}^2 \rangle = \langle x_{n-2}^2 \rangle + \langle \delta_{n-1}^2 \rangle$$

The mean-square average of x_n can then be written as:

$$\begin{aligned} \langle x_n^2 \rangle &= \langle x_{n-1}^2 \rangle + \langle \delta_n^2 \rangle \\ &= \langle x_{n-2}^2 \rangle + \langle \delta_{n-1}^2 \rangle + \langle \delta_n^2 \rangle \end{aligned}$$

Since the individual steps in a random walk are uncorrelated, and the individual walks are uncorrelated, the average values of δ_{n-1}^2 and δ_n^2 should be the same, so that we have:

$$\langle x_n^2 \rangle = \langle x_{n-2}^2 \rangle + 2\langle \delta_i^2 \rangle$$

where $\langle \delta_i^2 \rangle$ is the mean-square average of the change in x, averaged over all of the steps in the random walks.

The same logic can be applied repeatedly:

$$\begin{aligned} \langle x_n^2 \rangle &= \langle x_{n-2}^2 \rangle + 2\langle \delta_i^2 \rangle \\ &= \langle x_{n-3}^2 \rangle + \langle \delta_{n-2}^2 \rangle + 2\langle \delta_i^2 \rangle \\ &= \langle x_{n-3}^2 \rangle + 3\langle \delta_i^2 \rangle \\ &= \langle x_{n-4}^2 \rangle + 4\langle \delta_i^2 \rangle \end{aligned}$$

and so on, until we have:

$$\begin{aligned} \langle x_n^2 \rangle &= \langle x(1)^2 \rangle + (n-1)\langle \delta_i^2 \rangle \\ &= \langle x(0)^2 \rangle + n\langle \delta_i^2 \rangle \\ &= n\langle \delta_i^2 \rangle \end{aligned}$$

This derivation does not depend on any assumptions about the value of $\langle \delta_i^2 \rangle$, though it does assume that $\langle \delta_i \rangle$ is zero. If we further assume that δ_i is either l or $-l$, with equal probability, then the average of δ_i^2 can be further specified from the expected value:

$$\begin{aligned} \langle \delta_i^2 \rangle = E(\delta_i^2) &= p_+ l^2 + p_-(-l)^2 \\ &= p_+ l^2 + p_- l^2 \\ &= l^2\big(p_+ + p_-\big) \\ &= l^2 \end{aligned}$$

We can then write $\langle x_n^2 \rangle$ in the terms defining the random walk, n, the number of steps and l, the length of each step:

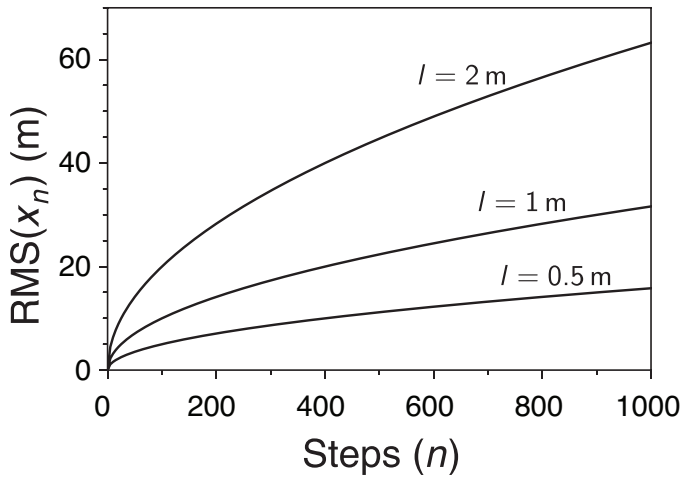
$$\langle x_n^2 \rangle = nl^2$$

The root-mean-square distance between the starting and ending positions is then given by:

$$\mathrm{RMS}(x_n) = \sqrt{nl^2} = \sqrt{n}l$$

Note that $\mathrm{RMS}(x_n)$ has the same dimensions, length, as the step length, l.

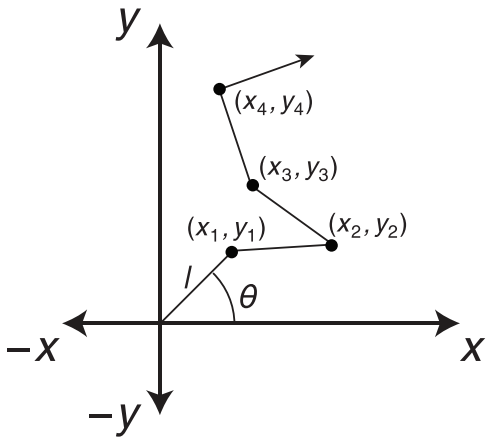
This is the key result: The RMS average distance increases with the square root of the number of steps. It doesn't increase linearly with the number of steps, because not every step moves the walker away from the starting point. But, the average distance isn't zero, even though the average position, $\langle x_n \rangle$ is zero. The relationship between the RMS end-to-end distance and the number of steps is shown in the figure below:



Note that for small values of n the RMS distance increases relatively rapidly with n. This is because, for a small number of coin flips, for instance, there is a relatively large probability that a significant majority will be either heads or tails. However, as the number of coin flips, n, increases, the likelihood of a significant deviation from the expected average decreases, and the RMS distance increases with n more slowly. For any given number of steps, $\mathrm{RMS}(x_n)$ is proportional to the step length.

3.2 Random walks in two dimensions

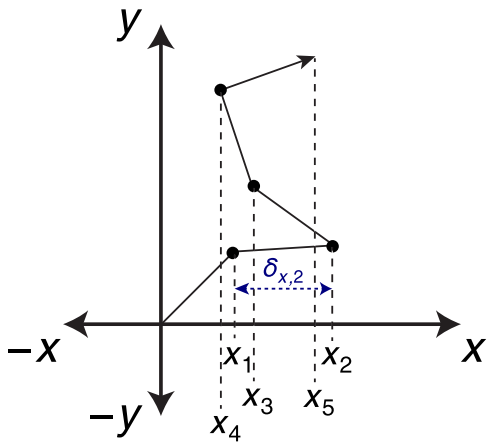
In the simplest form of a two-dimensional random walk, a walker begins at the origin of a two-dimensional coordinate system, where $x = 0$ and $y = 0$, and chooses at random an angle, θ, between 0 and 2π rad. The walker then takes a step, of length l in the direction defined by the angle θ with respect to the x-axis, as diagrammed below:



The process is then repeated $n-1$ times to generagte an n-step random walk.

I. The random walks along the x- and y-axes

As the walker generates a path in two dimensions, it also can be thought of as carrying out a walk along the x-axis. With each step, the projection of the current walker position onto the x-axis changes, as illustrated below:



For each step, the change in the x-coordinate is

$$\delta_{x,i} = x_i - x_{i-1}$$

Just as we did for the one-dimensional random walk along the x-axis, we can calculate the following averages for the walk defined by the projections along the x-axis:

$$\langle x_n \rangle = \frac{1}{N} \sum_{j=1}^{N} x_{n,j}$$

$$\langle x_n^2 \rangle = \frac{1}{N} \sum_{j=1}^{N} x_{n,j}^2$$

$$\text{RMS}(x_n) = \sqrt{\langle x_n^2 \rangle}$$

The central assumption that we will make at this point is that the turn angle at each step is equally likely to take on any value between 0 and 2π rad. This means that positive and negative changes in the x-coordinate are equally likely, leading to the result:

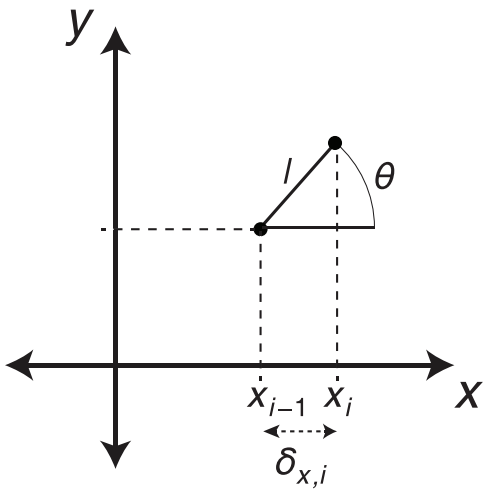
$$\langle x_n \rangle = 0$$

Recall that for the one-dimensional random walk we showed that:

$$\langle x_n^2 \rangle = n\langle \delta^2 \rangle$$

where $\langle \delta^2 \rangle$ is the mean-square average of the changes in position along the x-axis. This derivation assumed only that $\langle \delta_i \rangle = 0$, so it applies to the case of the x-projections in the two-dimensional random walk, as well.

For the one-dimensional random walk, we also argued that the individual changes in the x-position could only be l and $-l$ and, therefore, $\langle \delta_i^2 \rangle = l^2$. However, this argument does not apply to the changes in the x-projections in the two-dimensional random walk. To see why, consider the change in x-coordinate for a single step, as diagrammed below:



If the angle, θ is zero, then $\delta_{x,i} = l$, and $\delta_{x,i}^2 = \delta^2$. If θ is π, then $\delta_{x,i} = -\delta$, and $\delta_{x,i}^2 = l^2$. However, for most values of θ, $\delta_{x,i}$ lies between $-l$ and l and $\delta_{x,i}^2$ is less than l^2

To calculate the average value of $\delta_{x,i}^2$, we calculate the expected value for a continuous probability distribution function (see page 60):

$$\langle \delta_{x,i}^2 \rangle = E(\delta_{x,i}^2) = \int_{-\delta}^{\delta} \delta_x^2 p(\delta_x) d\delta_x$$

It's not so obvious what the probability distribution function, $p(\delta_x)$ is, but the random variable δ_x is related to the random variable θ according to:

$$\delta_x = l \cos\theta$$

From this relationship, the expected value of $\delta_{x,i}^2$, $\langle\delta_{x,i}^2\rangle$, can be calculated by integration with respect to θ:

$$\begin{aligned}\langle\delta_{x,i}^2\rangle &= \int_0^{2\pi} \delta_x(\theta)p(\theta)d\theta \\ &= \int_0^{2\pi} (l\cos\theta)^2 p(\theta)d\theta\end{aligned}$$

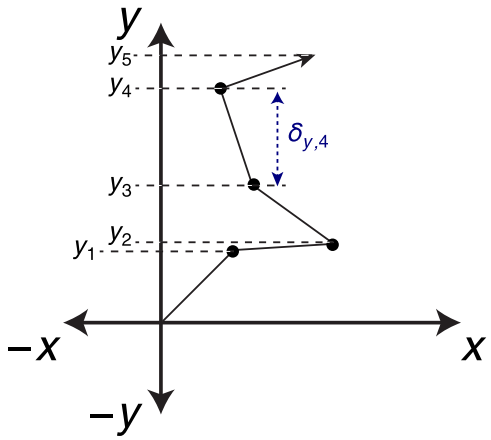
In Chapter 2 (page 59), it was shown that $p(\theta) = 1/(2\pi)$ for a uniform distribution of θ between 0 and 2π. The integral can then be evaluated as:

$$\begin{aligned}\langle\delta_x^2\rangle &= \frac{1}{2\pi}\int_0^{2\pi} (l\cos\theta)^2 d\theta \\ &= \frac{l^2}{4\pi}\left(\frac{\sin(2\theta)}{2} + \theta\right)\Bigg|_{\theta=0}^{2\pi} \\ &= \frac{l^2}{2}\end{aligned}$$

The mean-square projection along the x-axis of the endpoint after n steps is then calculated as:

$$\begin{aligned}\langle x^2\rangle &= n\langle\delta_x^2\rangle \\ &= \frac{nl^2}{2}\end{aligned}$$

The two-dimensional random walk can also be envisioned as creating a random walk along the y-axis:



There is nothing really special about either the x- or y-axis, or any other direction (though the *relationship* between the x- and y-axis is special, because they are perpendicular to each other). As a consequence, the results derived for the averages of the

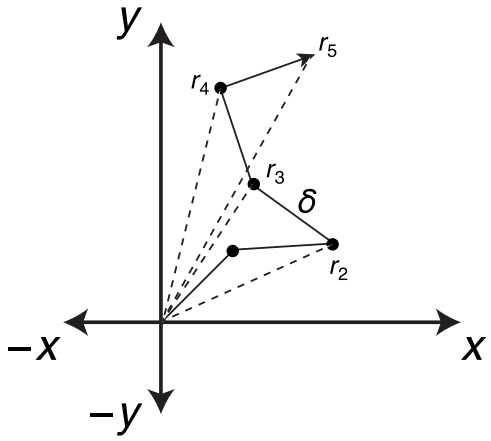
projections along the x-axis can be directly applied to the y-axis:

$$\langle y \rangle = 0$$

$$\langle y^2 \rangle = \frac{nl^2}{2}$$

II. The end-to-end distance

In addition to the projections along the x- and y-axis for a two-dimensional random walk, we can consider the distance between the starting and ending positions along the straight line connecting them, as opposed to the actual path of the walk. The diagram below shows how the distance of the walker from the starting position, r_i, changes as the number of steps in the random walk increases:



At the end of any specific random walk, the distance from the starting point, r_n, is related to the x- and y-projections according to:

$$r_n = \sqrt{x_n^2 + y_n^2}$$

and

$$r_n^2 = x_n^2 + y_n^2$$

To calculate the mean-square end-to-end distance, we again use the theorem for the expected value of a sum of two random variables. For two random variables, A and B, with expected values $E(A)$ and $E(B)$:

$$E(A + B) = E(A) + E(B)$$

The expected value for r^2 can thus be written:

$$E(r_n^2) = E(x_n^2) + E(y_n^2)$$

Assuming, as we have, that the number of random walks, N, over which the averages are taken is very large, this can be expressed in terms of the mean-square averages:

$$\langle r_n^2 \rangle = \langle x_n^2 \rangle + \langle y_n^2 \rangle$$

In the previous section, we showed that

$$\langle x_n^2 \rangle = \langle y_n^2 \rangle = nl^2/2$$

By substitution, we have:

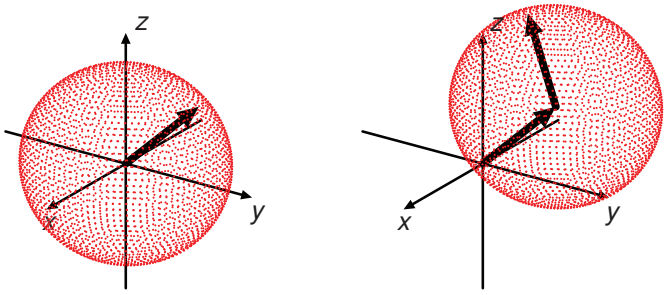
$$\begin{aligned} \langle r_n^2 \rangle &= \langle x_n^2 \rangle + \langle y_n^2 \rangle \\ &= nl^2/2 + nl^2/2 \\ &= nl^2 \end{aligned}$$

Thus, we have exactly the same result as for the one-dimensional random walk! The root-mean-square end-to-end distance is also the same as for the one-dimensional case:

$$\text{RMS}(r_n) = \sqrt{\langle r^2 \rangle} = \sqrt{n}l$$

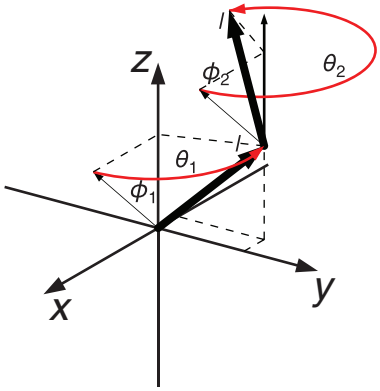
3.3 Three-dimensional Random Walks

A random walk in three-dimensions can be represented as a series of vectors in a three-dimensional coordinate system. The first step begins at the origin, as shown in the left-hand panel below, and ends on a random point on the surface of a sphere with its center at the origin and a radius equaling the step length.



The second step begins the end of the first and ends on a point on the sphere with its center at the starting point for the step, as illustrated in the right-hand panel.

To describe the changes in direction for each step, it is useful to use polar coordinates, as illustrated below:



In the polar-coordinate system, the position of the endpoint of a vector is described by the length of the vector and two angles. The vector is visualized as beginning initially aligned with the z-axis and then being rotated by an angle, ϕ, away from the z-axis in the plane of the x-z plane, and then rotated by an angle, θ, about the z-axis.

We can derive an expression for the mean-square end-to-end distance for a three-dimensional random walk by following the same general approach as for the two-dimensional case. For that case, we showed that

$$\langle r^2 \rangle = \langle x^2 \rangle + \langle y^2 \rangle$$

where $\langle x^2 \rangle$ and $\langle y^2 \rangle$ are the mean-square projections of the random-walk end-points onto the x- and $y-$ axes, respectively. For the three-dimensional case:

$$\langle r^2 \rangle = \langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle$$

In order to calculate $\langle x^2 \rangle$, $\langle y^2 \rangle$ and $\langle z^2 \rangle$, we need to consider the distributions of the projections of the individual steps onto the three axes and then calculate $\langle \delta_{x,i} \rangle$, $\langle \delta_{y,i} \rangle$ and $\langle \delta_{z,i} \rangle$.

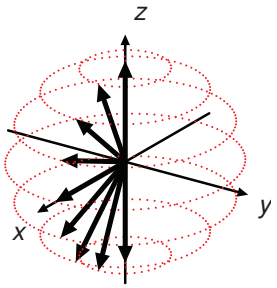
If the direction of each step is random with respect to the coordinate axis, the mean-square projection along any direction is the same as along any other direction. When using the polar coordinate system as defined above, the z-axis is the most convenient to work with, since the projection for a single step depends only on the step length, l, and the angle ϕ:

$$\delta_z = l \cos \phi$$

The mean-square projection of an individual step onto the z-axis is given by:

$$\langle \delta_z^2 \rangle = \int_0^\pi p(\phi) \delta_z^2 d\phi = \int_0^\pi p(\phi) (l \cos \phi)^2 d\phi$$

To evaluate this expression, we need to know the probability distribution function for the angle ϕ, $p(\phi)$. At first glance it might seem that all values of ϕ would be equally probable, so that $p(\phi)$ would be a simple constant. Recall, however, that the steps in the random walk were defined so that a step towards any point in the surrounding sphere is equally probable. The figure below represents the effect of rotating the vector by different values of ϕ from the z-axis and the points on the sphere that are accessible as the vector is rotated about the z-axis.



As indicated in the figure, the largest number of points on the sphere is associated with a rotation that places the vector in the x-y plane, corresponding to a value of ϕ equal to

$\pi/2$. At the other extreme, if $\phi = 0$ or π, the number of points is infinitesimally small. More generally, the number of points accessible for a given value of ϕ is proportional to the circumference of the circle swept out by the vector as it rotates about the z-axis. The circumference in turn is proportional to the radius, r_c, which is related to ϕ according to:

$$r_c = l \sin\phi$$

In order to satisfy the requirement that all directions in three dimensions be equally probable, the probability distribution function for ϕ must be proportional to $\sin\phi$.[1] The probability distribution function can thus be written in the form of:

$$p(\phi) = c \sin\phi$$

where c is a constant of proportionality. To evaluate this constant, we impose the requirement that the distribution function must be normalized:

$$\begin{aligned}\int_0^\pi p(\phi)d\phi &= \int_0^\pi c\sin\phi d\phi = 1 \\ &= \big(-c\cos\phi\big)\Big|_0^\pi \\ &= c - (-c) = 2c\end{aligned}$$

The constant c must then be equal to $1/2$ in order for the probability density function to be normalized:

$$p(\phi) = \frac{1}{2}\sin\phi$$

The average, or expected, value of the step-length projection onto the z-axis is then:

$$\begin{aligned}\langle\delta_z^2\rangle &= \int_0^\pi p(\phi)\delta_z^2 d\phi = \int_0^\pi p(\phi)(l\cos\phi)^2 d\phi \\ &= \int_0^\pi \frac{1}{2}\sin\phi(l\cos\phi)^2 d\phi \\ &= \frac{l^2}{2}\int_0^\pi \sin\phi\cos^2\phi d\phi\end{aligned}$$

The integral can be evaluated using a table of integrals or a computer program such as Mathematica, Maple or Maxima, to give:

$$\begin{aligned}\langle\delta_{z,i}^2\rangle &= \frac{l^2}{2}\int_0^\pi \sin\phi\cos^2\phi d\phi \\ &= \frac{l^2}{2}\left(-\frac{1}{3}\cos^3\phi\right)\Big|_0^\pi \\ &= \frac{l^2}{2}\left(\frac{1}{3}+\frac{1}{3}\right) = \frac{l^2}{3}\end{aligned}$$

[1] One could define the probability distribution function for ϕ as a constant, but this would give a different distribution of directions, in which those aligned more closely with the z-axis would be disproportionately favored.

We can now calculate the mean-square projection onto the z-axis of the end-to-end distances for a large number of n-step random walks:

$$\langle z_n^2 \rangle = n \langle \delta_{z,i}^2 \rangle = n \frac{l^2}{3}$$

Since the z-axis is not special (except for the definitions of ϕ and θ, which are arbitrary) the mean-square end-to-end projections onto the x- and $y-$ axis are also equal to $nl^2/3$, and the mean-square end-to-end distance is given by:

$$\begin{aligned} \langle r^2 \rangle &= \langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle \\ &= n\frac{l^2}{3} + n\frac{l^2}{3} + n\frac{l^2}{3} \\ &= nl^2 \end{aligned}$$

Thus, we have exactly the same result as for one and two dimensions. In fact, the same result applies to random walks in any number of dimensions, though it may be hard to visualize the ones in more than three dimensions.

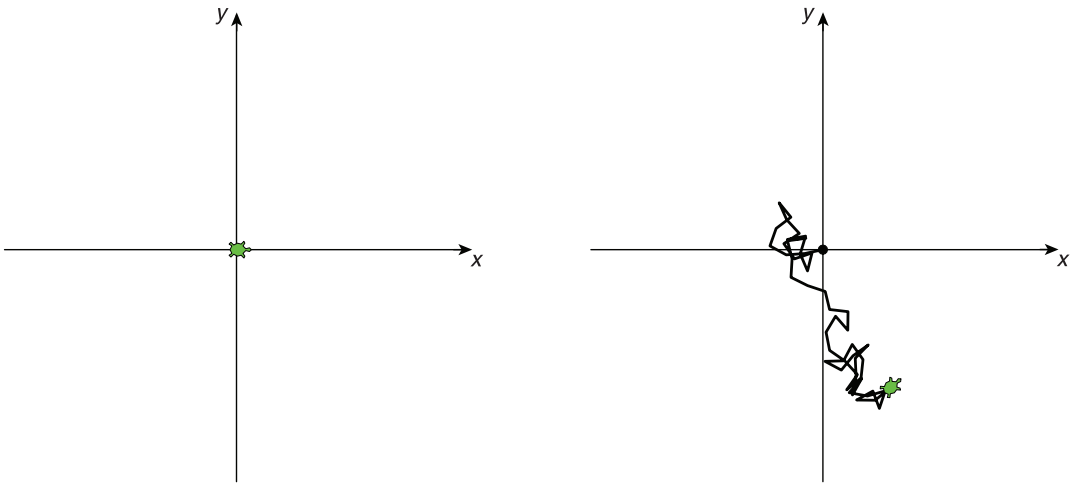
3.4 Computer Simulations of Random Walks

For many random processes, computer simulations can provide a useful complement to theoretical treatment, such as those derived in the previous sections. For random walks, simulations can provide snap shots of individual random walks and illustrate the distribution of properties, such as the end-point position, rather than just the averages that we have calculated so far. In addition, the process of writing the computer code for a simulation can help clarify one's thinking about the description of a random process.

I. Turtle graphics to illustrate two-dimensional random walks

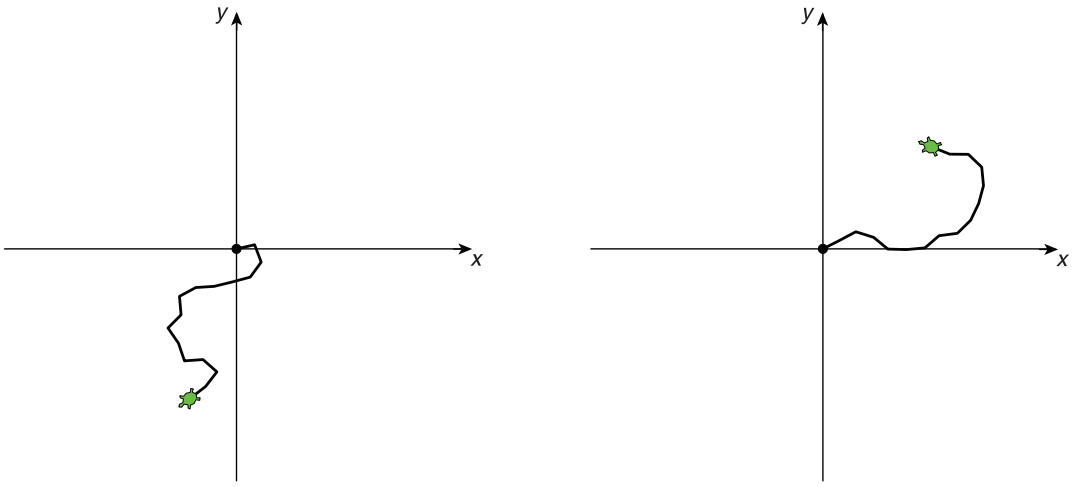
   Generating the high-quality computer graphics that we are now so familiar with takes a great deal of computer programming. But, there is a simple system that can be used for generating a graphical representation of two-dimensional random walks, called *turtle graphics*. Turtle graphics was introduced as a feature of a computer programming language, Logo, that was developed in the 1960s as a means for introducing children to computers and programming. Although there was much more to Logo than turtle graphics, that is probably the feature that it is best known for, and it has been adopted in other languages as well. The basic idea is that we imagine a turtle placed on a floor covered with a big piece of paper, and the turtle caries a pen. The turtle is given simple commands, such as "move forward by 10 units", or "turn right by 45°". These commands can be incorporated into programs where they are repeated numerous times, with variations, to generate a wide variety of patterns. Since a two-dimensional random walk consists of repeated steps and turns, turtle graphics is an ideal way to represent individual walks.

   The figures below show turtle-graphics representations of a turtle at its starting point, at the origin of the x-y coordinate system, (on the left) and after taking a taking a 200-step random walk (on the right).



The path of the turtle in this example illustrate an important general feature of random walks that is not readily apparent from the mathematical treatments of the previous two sections: The movement of the walker tends to be concentrated in small areas for a number of steps, followed by a series of steps in approximately the same direction, leading to a substantial excursion. The excursions are analogous to a series of coin tosses for which all, or nearly all, land heads. Although such series are relatively rare, they do occur on occasion, and have a significant effect when they do.

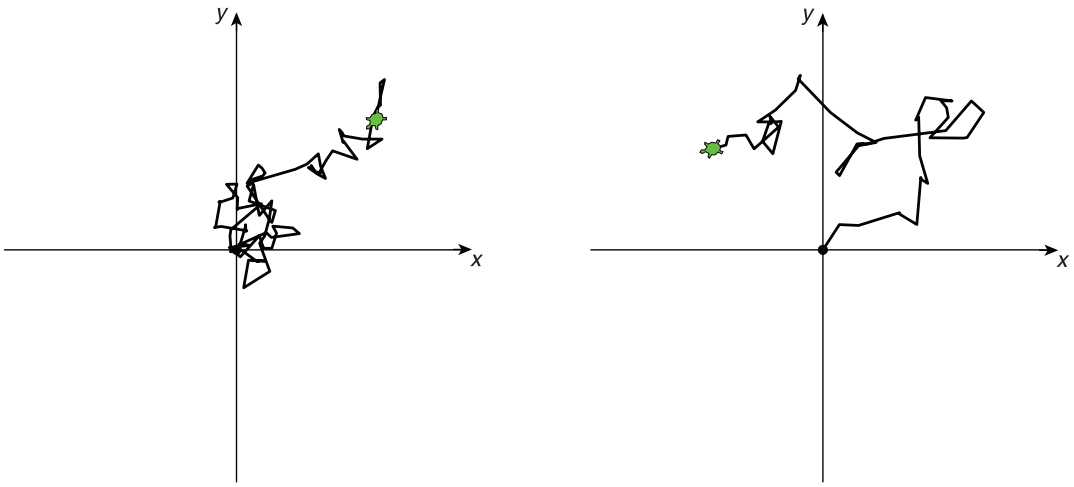
Another way in which simulations of this type are useful is that they let us explore the effects of changing the rules defining a random walk. For instance, the figures below represent random walks in which the turn angle at each step is constrained relative to the direction of the previous step. In the walk represented on the left, the turn from the previous was restricted to $\pm 90°$. In the right-hand figure, the turn was restricted to $\pm 46°$.



The obvious effects of restricting the turn angle in this way are to expand the random walk and reduce the number of times the path crosses itself. Note that these random walks consisted of only 20 steps each, yet the distance from the starting point is about

the same as for the 200-step random walk shown above, in which the turn angles were unrestricted. A random walk of this type is sometimes referred to as a *correlated random walk*.

We can also change the random walk by allowing the step length, as well as the turn angle, to vary randomly. The examples shown below were generated by sampling the step length from a Gaussian distribution centered at zero. For the diagram on the left, the standard deviation of the length distribution was 20 pixels, whereas the standard deviation was 30 pixels for the example on the right.

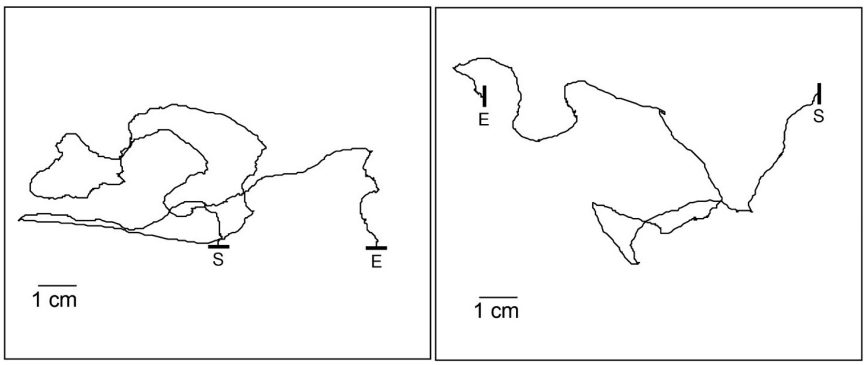


In both cases, short step lengths are the most probable, but longer ones are not uncommon. For the example on the left, with the narrower distribution, the random walk included 100 steps. To keep the walk to roughly the same extension, the number of steps was reduced to 50 for the walk shown on the right.

One of the important ways in which simulations of nearly any process can be used is to compare them to experimental observations. The extent to which the simulations matches the observed results can help support a theoretical model or indicate the ways in which the model might be improved. As an example the theory of random walks can be applied to analyzing the paths that animals follow when foraging for food.

The figures below represent the paths of individual ants, of two different species, as they foraged for food, as studied by Prof. Donald Feener and his colleagues at the University of Utah.[2]

[2]Pearce-Duvet, J. M. C., Elemens, C. P. H. and Feener, D. H. (2011) Walking the line: search behavior and foraging success in ant species. *Behavioral Ecology* 22, 501–509.



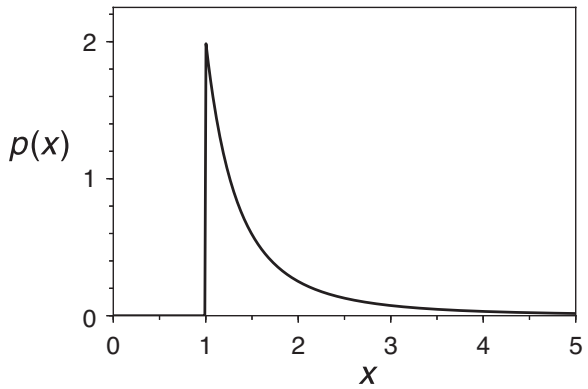
Brachymyrmex depilis (25 s) *Dorymyrmex insanus* (21 s)

The paths of the ants show some distinct similarities with those generated in the simulations shown above. In particular, there are frequent turns with a wide range of angles, and relatively straight segments of varying lengths. Of course, an ant has to take a large number of tiny steps to cover any significant distance, but it is reasonable to define random walk steps that correspond to the relatively straight segments in the path.

Particularly for *Brachymyrmex deplis*, the range of step lengths is extremely wide. A random walk model that has been used to describe walks with occasional steps that are very long is called a *Lévy flight*. This type of walk is characterized by a *long-tailed distribution* of step sizes, meaning that long steps are favored much more than by a Gaussian distribution. One such distribution is called the Pareto function, which has the general form:

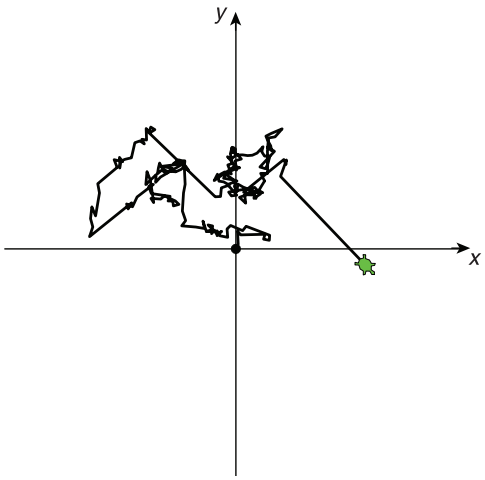
$$p(x) = \begin{cases} \frac{\alpha x_{\mathrm{m}}^{\alpha}}{x^{\alpha+1}} & \text{if } x \geq x_{\mathrm{m}} \\ 0 & \text{if } x < x_{\mathrm{m}} \end{cases}$$

where x_{m} is the minimum value for which the probability is greater than zero. The parameter α determines the rate at which the probability falls as x increases and lies between 0 and 2. An example of a Pareto distribution, with $x_{\mathrm{m}} = 1$ and $\alpha = 2$ is plotted below.



As written above, the Pareto function is a normalized probability distribution function. An interesting property of this function is that for $\alpha \leq 2$, its variance is infinite. (More properly, the integral representing the variance increases without bound as x increases.)

The figure below shows a simulated Lévy flight based on the Pareto function with the step lengths determined by a Pareto function with $x_{\mathrm{m}} = 10$ and $\alpha = 2$.

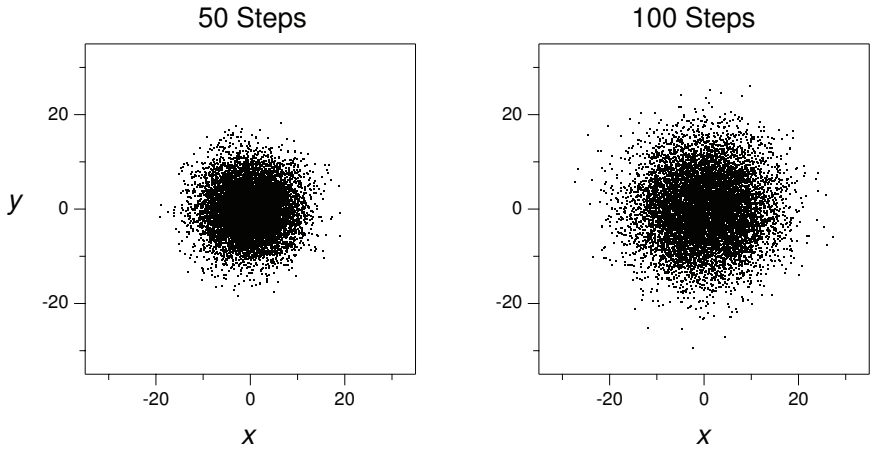


Compared to the random walks shown earlier, this one is characterized by a few very long steps, separated by much more localized random steps. In this respect, it seems to be a better model for the behavior of the ants shown above, and there is some evidence that this type of random walk is appropriate model for the foraging behavior of many species.

It should be noted that each of the turtle graphics representations shown above is just a single random walk and was chosen without complete objectivity. In particular the examples were chosen to highlight particular features, and for the fact that they didn't exceed the arbitrary boundaries of the axes. None the less, the features highlighted are quite real and can be found when large samples are examined.

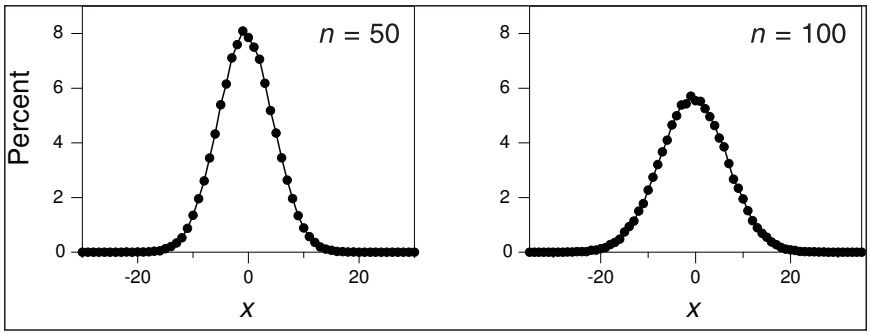
II. Simulating large samples of random walks

Another way in which simulations can be used is to generate large samples of random walks, from which general statistical insights can be gained. The figure below shows the endpoints of 10,000 two-dimensional random walks of 50 and 100 steps, on the left and right respectively. The steps in these random walks have a length of 1, in arbitrary units.



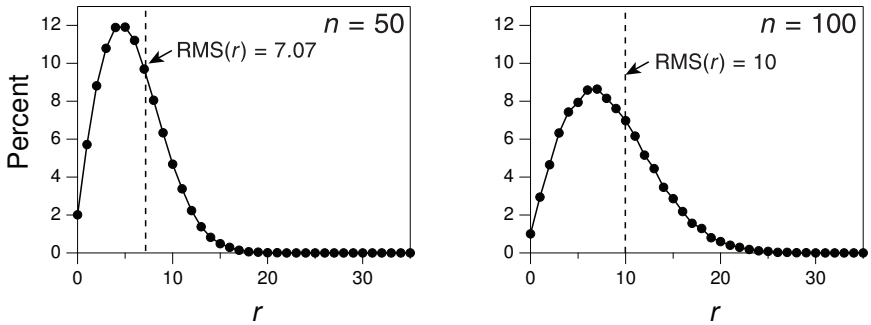
Note that the maximum projection along the x- or y-axes is 50 or 100, for the left and right panels, respectively. However, even among 10,000 random walks, distances greater than 20 are rare in either case.

The graphs below show the relative probabilities of the x-projections of the endpoints lying within intervals 1 step-length wide.

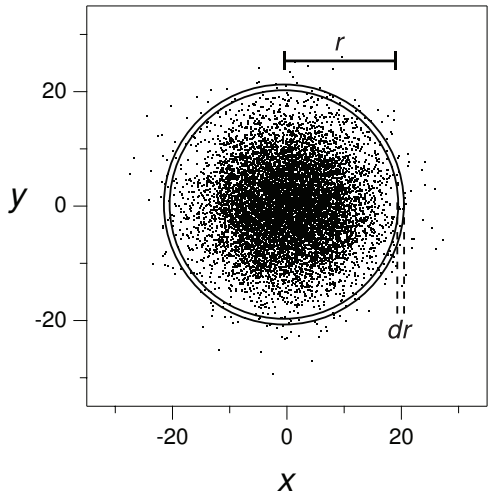


To produce these relatively smooth curves, 100,000 simulated random walks were generated. As expected, the distributions are bell shaped and centered at $x = 0$. As the number of steps increases, the breadth of the distribution increases.

The next set of graphs, below, show the distribution of distances between the starting and end points.



It may seem surprising that the peaks of these distributions are not at $r = 0$, since the highest density of endpoints in near the starting point. To understand this apparent paradox, it is important to keep in mind the meaning of a probability distribution function. If the probability distribution function is $p(r)$, then the probability that the distance lies with in a small interval of r-values is the product $p(r)dr$, as represented in the figure below:



As shown in the figure, the probability distribution function, $p(r)$, represents the probability that the endpoint of the random walk lies within an annulus dr thick at a distance r from the starting point. This probability depends on both the density of points at distance r *and* the area of the annulus. This area is calculated as:

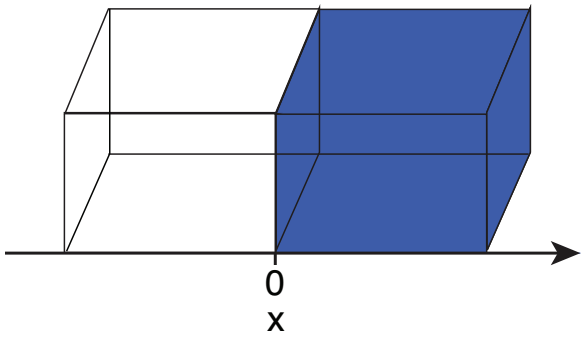
$$A = 2\pi r dr$$

To help visualize this relationship, imagine a thin metal ring, with radius r and thickness dr. If you were to cut this ring and flatten it out, the cross-sectional area (viewed along the thin edge) would be the length, $2\pi r$, times the thickness dr. Thus, the area of the annulus increases as r increases, while the density of endpoints decreases with r. The product of the area and the density is small when either term is small, and reaches a maximum at an intermediate value of r, as shown in the distribution function.

Chapter 4

Diffusion

Now, we go back to the problem introduced when we first began discussing probability, Brownian motion and diffusion. We would like some real numbers about how far a molecule or particle will diffuse in a given time, and we would like to know what factors determine this. Traditionally most experimental measurements of diffusion have been based on measuring changes in bulk concentration. An example that we will consider in some detail involves setting up two adjacent volumes, one containing a molecule of interest and the other without. At the beginning of the experiment, there is a sharp boundary between the two volumes, as diagrammed below.



Setting up an arrangement like this is technically challenging, but not impossible. Typically, the apparatus is set up vertically, but it is shown horizontally here, because we will define the $x$axis as the axis of diffusion, as indicated

With time, we expect the molecules to diffuse and the concentration to become more even. The rates at which the concentration changes at different points along the x-axis will depend on the rates at which the molecules move, and we should be able to deduce the parameters of the random walk from the rate of change in concentration.

What we are trying to do here is to extend the treatment of individual random walks to the bulk behavior of molecules that lead to concentration changes. This will require thinking about things a little differently.

4.1 Flux: Fick's First Law

I. The derivation

We will look at diffusion along a single dimension, x. Diffusion depends on Brownian motion, which can be described as a random walk. In each step of the walk, the direction is random, in three dimensions. If the mean-square length of the steps is $\langle l^2 \rangle$,

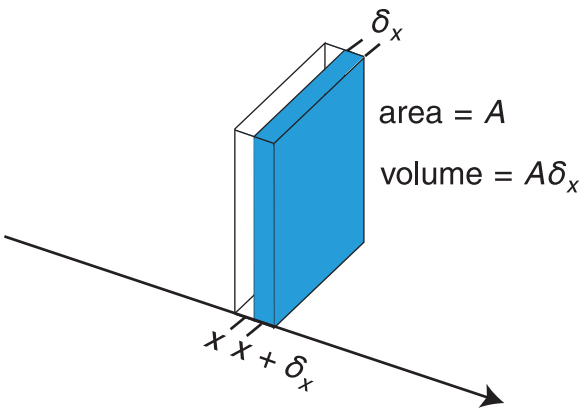
then the mean-square displacement of the step along the x-axis will be $\langle \delta_x^2 \rangle = \langle l^2 \rangle / 3$. For convenience, we will refer to the step size along the x-axis as $\delta_x = \mathrm{RMS}(\delta_x)$.

We will also define the average time interval between steps as τ. After a time period t, the number of steps will be $n = t/\tau$. The mean-square displacement along the x-axis will be:

$$\langle x^2 \rangle = n\delta_x^2 = t\delta_x^2/\tau$$

Notice that the average (RMS) distance will increase with the square root of time.

Now, let's consider two thin slices of the volume diagrammed in in the previous figure, cut perpendicular to the direction of the concentration difference:



The thickness of each volume is set to be δ_x, the RMS length of a random-walk step along the x-axis, and the cross-section along the x-axis has an area of A. Therefore, the volume of each slice is $A\delta_x$. During the interval τ, one half of the molecules within a slice will move to the left and half will move to the right. This will happen in both slices. If the number of molecules is the same in the two slices, the number of molecules that cross in each direction will be the same. But, if there are more molecules at position $x + \delta_x$ than at position x, then there will be a net movement of molecules to the left.

Call the number of molecules in the slice centered at x N_x, and the number of molecules in the slice at $x + \delta_x$ $N_{x+\delta_x}$. The net number of molecules going to the right in time t will be:

$$\begin{aligned} dN &= \frac{1}{2}N_x - \frac{1}{2}N_{x+\delta_x} \\ &= -\frac{1}{2}(N_{x+\delta_x} - N_x) \end{aligned}$$

Notice that we have defined things so that if the number of molecules to the right is larger than to the left, the flow of molecules to the right will be negative.

The flux across a given surface area is expressed as the number of molecules (or moles) per unit time per unit area. The expression above is divided by A and the time interval,

τ, to give the flux, J:

$$J = -\frac{1}{A\tau}\frac{1}{2}(N_{x+\delta_x} - N_x)$$

We can express the number of molecules in each slice in terms of the concentrations and the volumes of each slice:

$$N_x = C_x A\delta_x$$

$$N_{x+\delta_x} = C_{x+\delta_x} A\delta_x$$

We can then re-write the flux equation as:

$$J = -\frac{1}{A\tau}\frac{1}{2}(N_{x+\delta_x} - N_x)$$

$$= -\frac{1}{A\tau}\frac{1}{2}(C_{x+\delta_x} A\delta_x - C_x A\delta_x)$$

$$= -\frac{\delta_x}{\tau}\frac{1}{2}(C_{x+\delta_x} - C_x)$$

Now, we can write the difference in concentrations at the two positions in terms of a derivative with respect to x

$$\frac{dC}{dx} = \lim_{\delta_x \to 0} \frac{C_{x+\delta_x} - C_x}{\delta_x}$$

In the limit of small δ_x:

$$C_{x+\delta_x} - C_x = \delta_x \frac{dC}{dx}$$

So, we now have:

$$J = -\frac{\delta_x^2}{2\tau}\frac{dC}{dx}$$

Consider the quantity $\delta_x^2/(2\tau)$. Both parameters in this ratio are properties of the diffusing particle under a particular set of conditions. For now, we won't worry abut it's particular significance, but we will replace it with a new parameter, which we call the diffusion constant, D. Thus:

$$J = -D\frac{dC}{dx}$$

This equation is known as Fick's first law. (Adolf Eugen Fick, German physiologist, 1829–1901.)

Important conclusion: The *net* flux of molecules per unit area and per unit time is determined by the difference in concentration, and the diffusion coefficient, which reflects the steps in a random walk.

Why do molecules "move down a concentration gradient"? It's not because they can sense concentration! All of the molecules move randomly, but the probability for moving from a high concentration to a lower concentration is higher than the reverse simply because there are more molecules in the high-concentration region.

Let's also look at the units in this equation:

- The flux, J, has dimensions of molecules per cross-sectional area per unit time, or, in SI units: molecules $\cdot\, \mathrm{m}^{-2}\mathrm{s}^{-1}$. Alternatively, it can be expressed in terms of moles.
- The diffusion constant, $D = \delta_x^2/(2\tau)$, has units: $\mathrm{m}^2\mathrm{s}^{-1}$.
- The derivative of concentration with respect to x has dimensions of molecules per volume per length. In SI basic units this is: molecules $\cdot\, \mathrm{m}^{-3}\mathrm{m}^{-1}$ = molecules $\cdot\, \mathrm{m}^{-4}$
- Combining these:

$$J = -D\frac{dC}{dx}$$

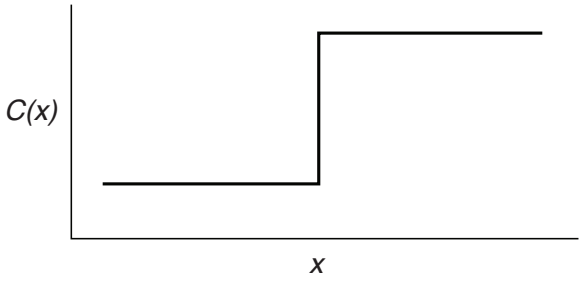
$$\text{molecules} \cdot \mathrm{m}^{-2}\mathrm{s}^{-1} = \mathrm{m}^2\mathrm{s}^{-1} \times \text{molecules} \cdot \mathrm{m}^{-4}$$

$$\text{molecules} \cdot \mathrm{m}^{-2}\mathrm{s}^{-1} = \text{molecules} \cdot \mathrm{m}^{-2}\mathrm{s}^{-1}$$

Looks good!

II. The distribution of molecules diffusing from a single position.

Consider the case described earlier, where there is initially a sharp boundary between an area where $C = 0$ and an area with $C = 1$, in arbitrary concentration units



We might ask: For a molecule at any initial position along the x-axis, what is its most likely position after a given period of time? If we treat this as a random-walk problem, we conclude that the most likely position is the starting position, even though the random walk will have taken the molecule to many other positions during the time period. But, this must be true for all of the molecules in the sample. So, how does net diffusion ever take place?

The solution to the paradox lies in the nature of the probability distribution function. Recall that the Gaussian distribution for a random walk is given by:

$$p(x) = \sqrt{\frac{1}{2\pi\langle x^2\rangle}}e^{-x^2/(2\langle x^2\rangle)}$$

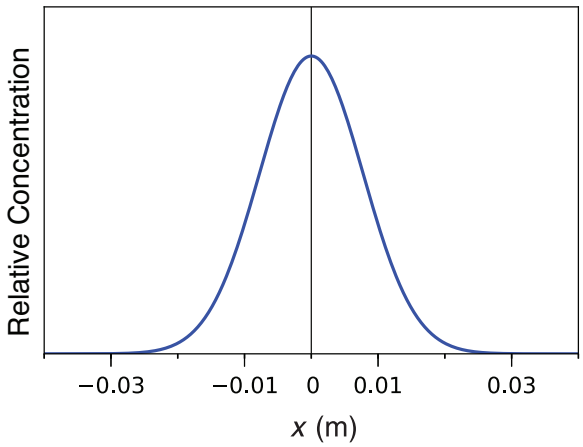
where x is the position of the endpoint and $\langle x^2 \rangle$ is the mean-squared value of x. For diffusion, we defined the random walk parameters in terms of δ_x (the step size), τ (the time interval between steps), t (the total time) and $D = \delta_x^2/(2\tau)$, so that $\langle x^2 \rangle$ is given by:

$$\begin{aligned}\langle x^2 \rangle &= n\delta_x^2 \\ &= \frac{t\delta_x^2}{\tau} \\ &= 2Dt\end{aligned}$$

So, the probability function can be expressed in terms of D and t:

$$p(x) = \sqrt{\frac{1}{4\pi Dt}} e^{-x^2/(4\pi Dt)}$$

A plot of the function looks like:



For this plot, $D = 3 \times 10^{-10}\,\mathrm{m}^2/\mathrm{s}$, and $t = 10^5\,\mathrm{s}$.

Remember that a continuous probability distribution function is interpreted in terms of its integral. In this case, the probability that the molecule lies between two positions, a and b is given by the integral:

$$\int_a^b \sqrt{\frac{1}{4\pi Dt}} e^{-x^2/(4Dt)} dx$$

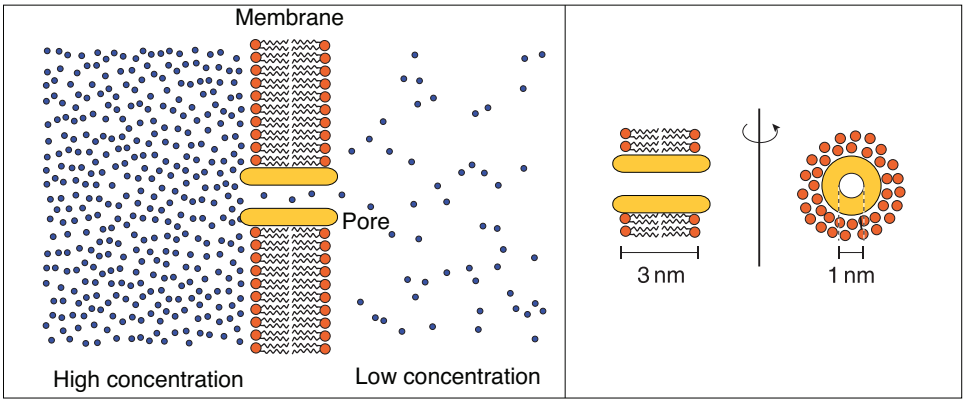
If we divide up the range of x values into thin slices, the slice with the largest probability is the one centered at $x = 0$. *But* the *total* probability that the molecule will be in one of the other slices is much larger than the probability that x will be close to zero.

So, we really do expect virtually all of the molecules to be somewhere else after the time period. The lesson here is that it is not enough to ask what the most likely outcome is! The most likely may represent only a tiny fraction of the total.

III. A (simplified) biological example

In biology, we don't really have examples where we start with a perfectly sharp boundary, with nothing separating the two sides. But, we do have lots of cases where there is a sharp change in concentration across a membrane. Biological membranes are composed of bilayers of phospholipids, along with proteins that are embedded in the bilayer. For now, the structural details are not very important, except that most molecules diffuse across lipid bilayers extremely slowly, and proteins can act as pores that allow much faster selective diffusion of molecules.

Suppose that we have a membrane with a thickness of about 3 nm, a typical value, and a small-molecule compound that has a concentration of 50 mM on one side of the membrane and 5 mM on the other.



Across the width of the membrane, we can estimate the concentration gradient as:

$$\frac{dC}{dx} = \frac{50\,\text{mM} - 5\,\text{mM}}{3\,\text{nm}} = \frac{0.045\,\text{M}}{3 \times 10^{-9}\,\text{m}}$$
$$= 1.5 \times 10^{7}\,\text{M/m}$$

To go further, we need to convert the concentration gradient to units with consistent units of length:

$$\frac{dC}{dx} = 1.5 \times 10^{7}\,\text{M/m}$$
$$= 1.5 \times 10^{7} \frac{\text{moles}}{\text{L} \times \text{m}} \times \frac{1\,\text{L}}{10^{-3}\,\text{m}^3}$$
$$= 1.5 \times 10^{10}\,\text{moles/m}^4$$

From Fick's first law, we can calculate the flux, J:

$$J = -D\frac{dC}{dx}$$

A typical diffusion coefficient for small molecules is $10^{-10}\,\mathrm{m}^2/\mathrm{s}$.

$$\begin{aligned} J &= -10^{-10}\,\mathrm{m}^2/\mathrm{s} \times 1.5 \times 10^{10}\,\mathrm{moles/m}^4 \\ &= -1.5\,\mathrm{moles}/(\mathrm{m}^2\mathrm{s}) \end{aligned}$$

This looks like a lot of molecules per second, but remember that the flux is expressed per unit of area, in this case $1\,\mathrm{m}^2$. The negative sign simply indicates that the flux is in the opposite direction of the concentration gradient.

The pores in membranes vary greatly is size and shape, and many of them have very small and specialized structures for which a general treatment of diffusion is probably not appropriate. But, there are examples of pores with diameters of a few nm. For a pore diameter of 1 nm, the area is:

$$A = \pi r^2 = \pi(0.5 \times 10^{-9}\,\mathrm{m})^2 = 7.8 \times 10^{-19}\,\mathrm{m}^2$$

The flow through this pore is then:

$$1.5\,\mathrm{moles}/(\mathrm{m}^2\mathrm{s}) \times 7.8 \times 10^{-19}\,\mathrm{m}^2 = 1.2 \times 10^{-18}\,\mathrm{moles/s}$$

The number of molecules per second is:

$$1.2 \times 10^{-18}\,\mathrm{moles/s} \times 6.02 \times 10^{23}\,\mathrm{molecules/mole} \approx 7 \times 10^5\,\mathrm{molecules/s}$$

How many molecules would be in the volume of the pore at any instant? The volume is:

$$\begin{aligned} V = A \times L &= \pi r^2 \times L = \pi(0.5 \times 10^{-9}\,\mathrm{m})^2 \times 3 \times 10^{-9}\,\mathrm{m} \\ &= 7.8 \times 10^{-19}\,\mathrm{m}^2 \times 3 \times 10^{-9}\,\mathrm{m} = 2.4 \times 10^{-27}\,\mathrm{m}^3 \\ &= 2.4 \times 10^{-24}\,\mathrm{L} \end{aligned}$$

If the concentration within the pore is the average of that on the two sides of the membrane, 20 mM, the average number of molecules in the pore is calculated as:

$$2.4 \times 10^{-24}\,\mathrm{L} \times 0.02\,\mathrm{moles/L} = 5 \times 10^{-26}\,\mathrm{moles} \approx 0.03\,\mathrm{molecules}$$

This result means that the pore is empty nearly all of the time, even though about 10^6 molecules are passing through every second. So, each is there for a very short time.

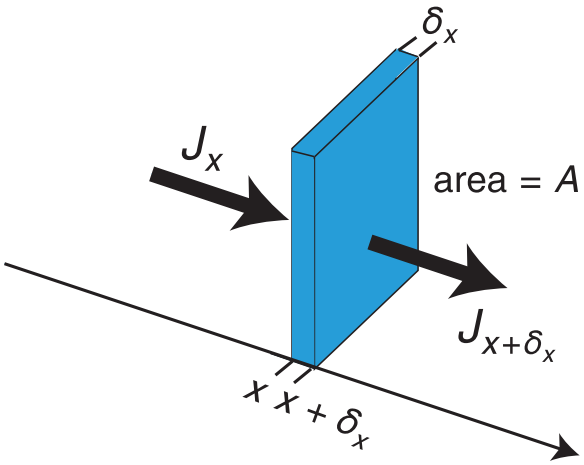
Can a flow of 10^6 molecules/s be detected through a singe pore? It can be if the molecule is charged, by measuring electric current. A current of 1 ampere (A) corresponds to 1 coulomb per s, or about $6{\times}10^{18}$ charges per s. So 10^6 charges/s corresponds to about $2 \times 10^{-13}\,\mathrm{A}$, or 0.2 pA. This is a very small current, but currents of this magnitude are routinely measured by electrophysiologists studying single channels in membranes in (or removed from) neurons and muscle cells.

4.2 Fick's second law

As soon as there is a net flux between regions of a sample, the concentrations will change. Fick's second law describes the change in concentration with time.

I. The derivation

Consider, again, a thin slice cut perpendicular to the x-axis, with area A and thickness δ_x:



The net number of molecules moving to the right at the two sides of the slice during an interval dt will be:

$$N_x = AJ_x dt$$

$$N_{x+\delta_x} = AJ_{x+\delta_x} dt$$

where J_x and $J_{x+\delta_x}$ are the values of the flux at positions x and $x + \delta_x$ respectively. The change in the number of molecules in the slice will be:

$$\begin{aligned} dN &= AJ_x dt - AJ_{x+\delta_x} dt \\ &= Adt(J_x - J_{x+\delta_x}) \end{aligned}$$

The change in concentration will be:

$$\begin{aligned} dC &= \frac{dN}{A\delta_x} \\ &= \frac{Adt(J_x - J_{x+\delta_x})}{A\delta_x} \\ &= -dt\frac{J_{x+\delta_x} - J_x}{\delta_x} \end{aligned}$$

In the limit of small dt and small δ_x:

$$\frac{dC}{dt} = -\frac{J_{x+\delta_x} - J_x}{\delta_x} = -\frac{dJ}{dx}$$

From Fick's first law, we know how J depends on the change of C with respect to x:

$$J = -D\frac{dC}{dx}$$

Differentiating J with respect to x gives:

$$\frac{dJ}{dx} = -D\frac{d^2C}{dx^2}$$

Substituting:

$$\frac{dC}{dt} = D\frac{d^2C}{dx^2}$$

This is the usual form of Fick's second law. It is also referred to as the "diffusion equation", and it provides the basis for calculating how concentration will change with time, provided that we know how concentration depends on position, x, which, of course, changes continuously.

The good news is that diffusion is described by this very simple equation. The bad news is that it's not at all simple to solve this equation for real problems. What is required is to find a function, C, of both x and t, that satisfied this differential equation *and* describes the particular physical arrangement at when $t = 0$.

Historically, problems of this type were first solved in the context of heat flow through materials, which follows the same mathematical laws as diffusion. Consideration of problems of this type led the French mathematician Joseph Fourier to develop the methods now known by his name, Fourier analysis. This can be, and is, the subject of entire courses.

The two laws of Fick describe different, but closely related, features of the dynamics:

1. Fick's first law states that the net flux of diffusing molecules is proportional to the change in concentration with respect to distance, *i.e.,* the "concentration gradient".
2. Fick's second law states that the change in concentration with respect to time, at a given position, is proportional to the derivative of the concentration gradient, *i.e.,* how rapidly the concentration gradient changes with position.

To see how these relationships work, we will look at the solution to the case of diffusion from a sharp boundary.

4.3 Diffusion from a Sharp Boundary

The case introduced earlier, diffusion from a sharp boundary, is one for which a solution to the diffusion can be found relatively easily. Though this case is highly simplified and restricted to diffusion in only a single dimension, its solution reflects general properties of diffusion and provides considerable insight.

I. A solution to the diffusion equation

In looking for a solution to the diffusion equation, we are looking for a function, $C(x,t)$, that describes the concentration of the diffusing molecules as a function of both x and t such that the derivatives of the function satisfy the differential equation:

$$\frac{dC}{dt} = D\frac{d^2C}{dx^2}$$

For each value of x, the concentration at time t will represent all of the molecules that have diffused to that point, from all of the points at which molecules were initially present (including x itself). For the case of diffusion from a sharp boundary, we can say the following:

- For $x < 0$, the initial concentration is 0.
- For $x >= 0$ the concentration is initially the same, and we can call this value 1, in arbitrary units.

These constitute the boundary conditions for the problem. Any solution must satisfy these conditions, as well as the differential equation.

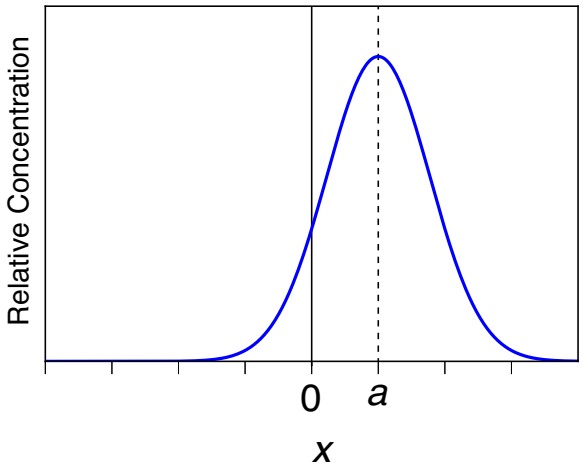
Assuming that there are initially a very large number of molecules in the vicinity of each value of $x >= 0$, we know that the final distribution of molecules *from that position* will be described by a Gaussian probability distribution function that is centered at the initial position. Recall that the Gaussian function, for molecules beginning at $x = 0$, can be written as:

$$p(x) = \sqrt{\frac{1}{4\pi Dt}}e^{-x^2/(4Dt)}$$

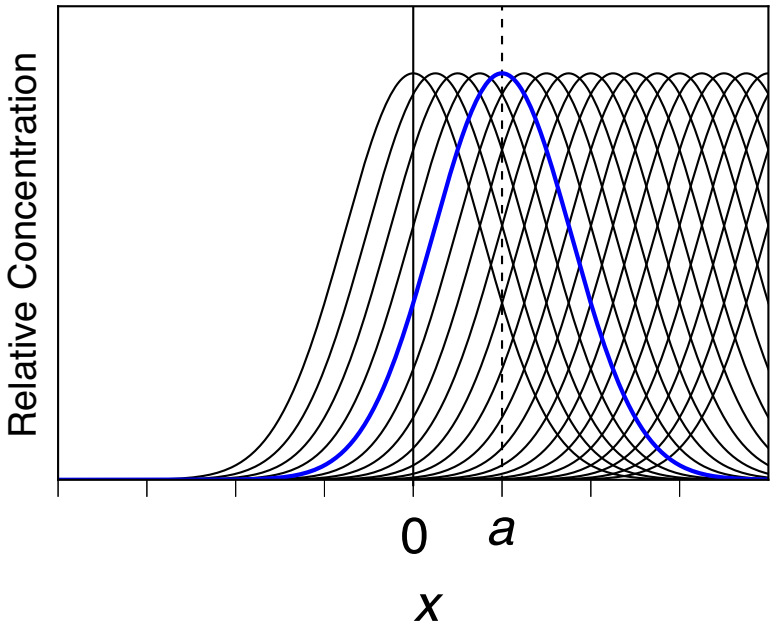
More generally, if we consider molecules that begin at position $x = a$, we can replace x with $(a - x)$ in the distribution to give

$$P(x,a) = \sqrt{\frac{1}{4\pi Dt}}e^{-(a-x)^2/(4Dt)}$$

A plot of the function looks like:



For a given position x, the final concentration will be the total of molecules from all values of $a > 0$. So, we add together the value of all of the probability functions for all values of $a > 0$, as represented in the figure below:



In other words, we integrate:

$$C(x,t) = \int_0^\infty \sqrt{\frac{1}{4\pi Dt}} e^{-(a-x)^2/(4Dt)} da$$

Notice that the variable a doesn't appear in the final result, it simply represents all of the possible starting positions of the molecules. This integral cannot be evaluated analytically, but it can be estimated numerically.

In most textbooks, this result is presented somewhat differently, as:

$$C(x,t) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{4Dt}}\right)\right]$$

where erf is called the "error function" (because it arises in the statistical analysis of measurement errors) and is defined as:

$$\text{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-u^2} du$$

We can check that our solution satisfies Fick's second law by calculating the appropriate derivatives. For this purpose, it is convenient to use the form using the error function, substituting $x/\sqrt{4Dt}$ for z:

$$C(x,t) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{x/(4Dt)} e^{-u^2} du$$

Since the function $C(x,t)$ is an integral of the Gaussian function, it should not be surprising that the derivatives of $C(x,t)$ are Gaussian functions. The fundamental theorem of calculus stipulates that if a function $F(X)$ is defined as:

$$F(X) = \int_a^X f(x)dx$$

then the derivative of $F(X)$ is simply $f(X)$.

Using this relationship and some substitutions, differentiating $C(x,t)$ with respect to x gives:

$$\frac{dC}{dx} = \frac{d}{dx}\left[\frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{x/(4Dt)} e^{-u^2} du\right]$$

$$= \sqrt{\frac{1}{4\pi Dt}} e^{-x^2/(4Dt)}$$

The second derivative of C with respect to x is:

$$\frac{d^2C}{x^2} = -\frac{x}{4\sqrt{\pi D^3 t^3}} e^{-x^2/(4Dt)}$$

With some effort (or the assistance of a computer program such as Mathematica or Maxima), the derivative of C with respect to t can be shown to be:

$$\frac{dC}{dt} = -\frac{x}{4\sqrt{\pi D^3 t^3}} e^{-x^2/(4Dt)}$$

Thus, Fick's second law is satisfied by this solution:

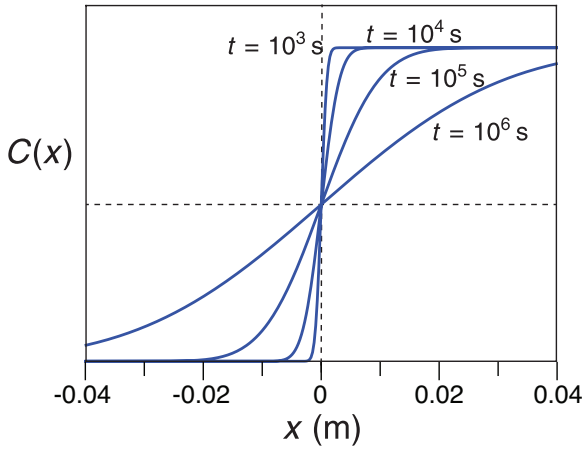
$$\frac{dC}{dt} = D\frac{d^2C}{dx^2}$$

$$-\frac{x}{4\sqrt{\pi Dt^3}} e^{-x^2/(4Dt)} = -D\frac{x}{4\sqrt{\pi D^3 t^3}} e^{-x^2/(4Dt)}$$

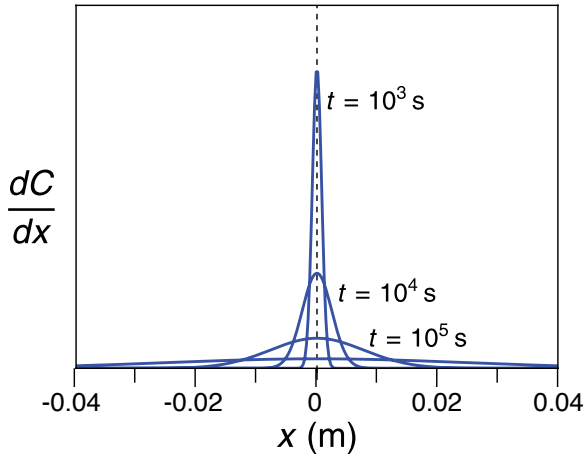
II. Graphical representations of the solution

The best way to get a feel for all of this is to look at graphs representing the solution and its derivatives.

The profiles predicted by the solution to the diffusion equation (for this particular starting state) are shown in the graph below, for the case where $D = 3 \times 10^{-10}\,\mathrm{m}^2/\mathrm{s}$ and the time after creation of the sharp boundary, t, is 10^3, 10^4, 10^5 or $10^6\,\mathrm{s}$, as indicated. (Note that the function is not defined for $t = 0$.)



The first derivative of C with respect to x is plotted below for the same value of D and the indicated times.

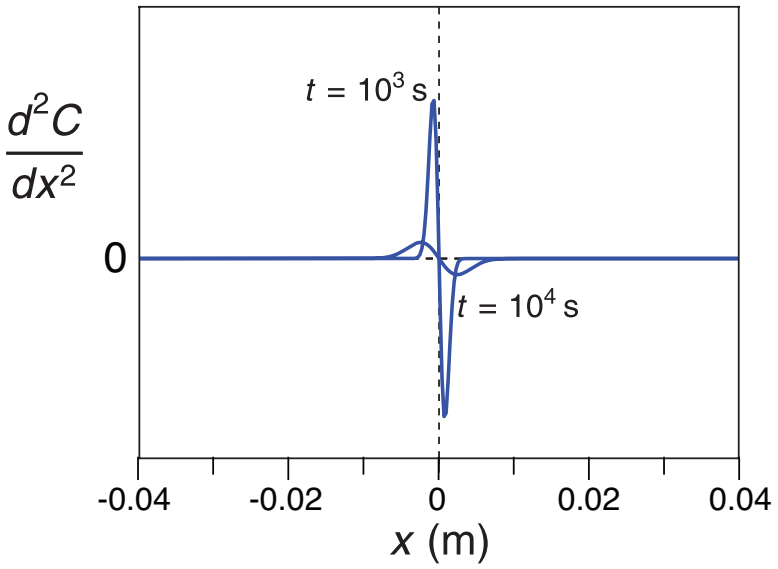


Notice that the derivative has the form of a Gaussian function, which reflects the fact that $C(x,t)$ has the form of an integral of the Gaussian. The peak of the concentration gradient remains at $x = 0$, but decreases in steepness with time as the gradient covers a wider range of x values.

The *net* rate at which molecules pass a particular point is proportional to the concentration gradient. Thus, the flux, J, is always maximal at $x = 0$ but decreases with

time at this point. At other points, however, the flux increases and then decreases with time.

Finally, we look at the second derivative of C with respect to x:



Notice that there are two peaks, one positive and one negative. At $t = 0$, these are extremely sharp and represent the two edges of the sharp concentration gradient. With time, these peaks move apart and become less pronounced as the concentration gradient becomes more gentle.

The positive peak on the left represents the region where the concentration is beginning to increase and where the gradient increases most rapidly. This is where the concentration is increasing most rapidly. But, it's not where the flux, J, is maximal! The flux is always maximal at $x = 0$.

Why is the region where the flux is greatest *not* where the concentration changes most rapidly?

At $x = 0$, the flux is maximal, but the molecules are constantly being replaced by the "reservoir" to the right and are being drawn off to the left. So, the concentration stays constant.

Where the second derivative is maximal, the flux changes most rapidly with position. This means that the flux going into a volume element is greater than that leaving, so that the concentration changes most rapidly.

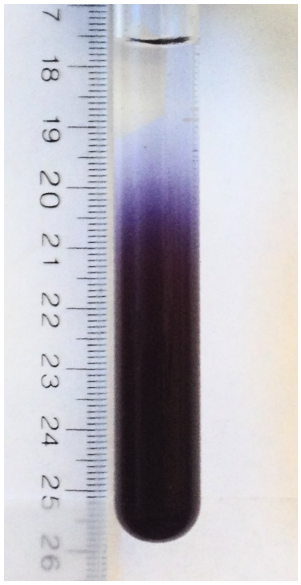
As time increases, the absolute values of both the first and second derivatives decrease, so that both the flux and the rate of change in concentration decrease.

Notice, also that the concentration at positions close to $x = 0$ change very rapidly, but even a millimeter away, the change is quite slow.

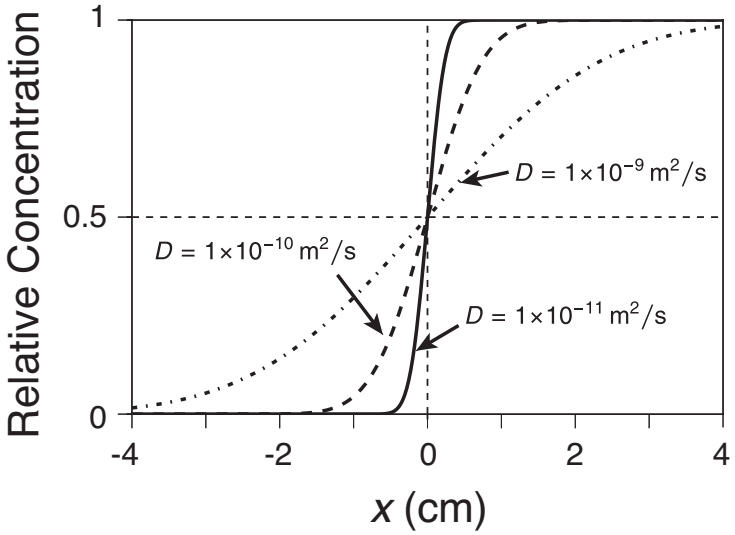
4.4 Estimating a Diffusion Constant from a Simple Experiment

An approximation to the ideal sharp boundary experiment can be realized in practice by overlaying two solutions, one containing a dye. In order to form the (relatively) sharp boundary easily, it is also necessary to make the lower solution slightly more dense than the upper one, for instance by adding a few percent of glycerol. The increased viscosity of the lower solution slightly complicates the situation, but not so much as to obscure the essential features of the experiment

The photograph below shows the result of such an experiment 48 h after establishing the boundary.



Even after 48 h, very little, if any, of the dye has reached the top of the solution, about 2 cm from the initial boundary. We can use the result of this experiment to make a rough estimate of the diffusion coefficient, D, by comparing the observation to those predicted by the solution for diffusion from a boundary. The plot below shows the expect concentration, after 48 h, as a function of position, x, with different values assumed for the diffusion coefficient, as indicated.



As a rough estimate, it appears that the experimental result lies somewhere between the curves calculated for $D = 10^{-10}$ and $10^{-9}\,\mathrm{m}^2/\mathrm{s}$. Later, we will see how to calculate the diffusion coefficient from the size of a molecule and the viscosity of the solution, and we'll see how close the two estimates match.

4.5 Molecular Motion and Kinetic Energy

So far, we have considered diffusion from a macroscopic point of view, focusing on the net flux of molecules and changes in concentration, without considering the nature of the molecular motions. To take a more microscopic view, we need to consider the motions of individual molecules, which reflect their kinetic energy.

I. Kinetic energy

We can begin this discussion by asking, in the most general way, what is energy? The standard textbook definition is that energy is the ability do do work. But what, then, is work?

For mechanical motion, work, and therefore energy, represents an integral of force with respect to distance:

$$w = \int_a^b F dx$$

The units of energy are those of force times distance. In SI units N·m = joule (J).

The unit of force, N, is defined from Newton's second law:

$$F = m \cdot a = \mathrm{kg} \cdot \mathrm{m} \cdot \mathrm{s}^{-2}$$

Therefore, in the basic SI units, energy has units of $1\,\mathrm{J}{=}1\,\mathrm{kg} \cdot \mathrm{m}^2{\cdot}\mathrm{s}^{-2}$.

The kinetic energy of an object moving along a given direction, x, from classical mechanics is:

$$E_{\mathrm{k},x} = m \cdot v^2/2$$

This represents the work required to accelerate an object of mass m from rest to velocity v, in the absence of friction. It doesn't matter how rapidly or slowly the acceleration is, the final energy depends only on mass and velocity. If the object slows down, it looses some of that energy.

Note that kinetic energy has the correct units, $= \mathrm{kg} \cdot \mathrm{m}^2{\cdot}\mathrm{s}^{-2}$

Also note that doubling the velocity increases the kinetic energy four fold. This is why car accidents become so much more dangerous at higher speeds.

II. Thermal energy

Fundamentally, temperature is a measurement of the motion of molecules. The simplest thermometer is a container of gas that expands or contracts when the temperature changes (at constant pressure). Alternatively, we can measure temperature by measuring the pressure of a gas at constant volume. Though this was not always understood, we now know that the pressure represents the collisions of molecules against the side of the container.

For an ideal gas, we have the relationship:

$$PV = nRT$$

where P is pressure, V is volume, n is the number of moles of gas, T is temperature and R is the gas constant. An ideal gas is one that is made up of particles that do not interact at all with one another. At moderate temperatures and pressures, real gasses are well approximated by the ideal gas law.

What are the units of the gas constant?

Pressure has the units of force per unit area, or $\mathrm{N} \cdot \mathrm{m}^{-2}$, and volume has the units of m^3. T has units of K, and n has units of moles. Therefore, R has units of:

$$R = \frac{PV}{nT} = \frac{\mathrm{N} \cdot \mathrm{m}^{-2}\mathrm{m}^3}{\mathrm{K} \cdot \mathrm{mol}} = \mathrm{N} \cdot \mathrm{m}/(\mathrm{K} \cdot \mathrm{mol})$$

In the basic SI units, R has the units of:

$$\mathrm{N} \cdot \mathrm{m}/(\mathrm{K} \cdot \mathrm{mol}) = \mathrm{kg} \cdot \mathrm{m}^2\mathrm{s}^{-2}\mathrm{K}^{-1}\mathrm{mol}^{-1}$$

Notice, though, that we just showed that $\mathrm{kg} \cdot \mathrm{m}^2 \cdot \mathrm{s}^{-2}$ is a unit of energy, the Joule. So, R can be expressed in units of J/K, and the product RT must, in some sense be a measure of the energy that one mole of moving molecules have at a given temperature, irrespective of pressure and volume.

To discuss the energy of individual molecules, it is convenient to divide the gas constant by Avogadro's number. ($\approx 6.02 \times 10^{23}$). This is the Boltzmann constant, which, in SI units, is:

$$k = 1.3806 \times 10^{-23}\,\mathrm{kg} \cdot \mathrm{m}^2 \mathrm{s}^{-2} \mathrm{K}^{-1} = 1.3806 \times 10^{-23}\,\mathrm{J/K}$$

In a given volume of gas, not all of the molecules will have the same energy (or velocity) at a given instant. In fact, they will have a broad distribution of energies as they collide with one another and the walls and exchange energy. So, we have to express the kinetic energy as an average. Without going through the derivation, the RMS translational kinetic energy in one direction is:

$$\mathrm{RMS}(E_{k,x}) = kT/2$$

And the total translational energy, summed over all three directions is:

$$\mathrm{RMS}(E_k) = 3kT/2$$

Remember, though that the kinetic energy is also expressed in terms of the velocity and mass of a particle:

$$E_k = m \cdot v^2/2$$

Combining these equations gives:

$$v = \sqrt{kT/m}$$

where v is understood to be an RMS average velocity. With this equation, we can calculate the RMS velocity knowing only the mass of a molecule and the temperature.

III. Steps in the random walk

The equations for kinetic energy in a gas also apply to molecules in a liquid, at the same temperature. The instantaneous velocities are the same, it's just that the molecules collide with one another much more frequently in a liquid, and we can now calculate just how frequent that is.

For a single (average) step in the random walk, the displacement is δ_x and the time is τ. Therefore, the velocity during this period is:

$$v = \delta_x/\tau$$

If we have measured the diffusion coefficient, D, we can now calculate δ_x:

$$\begin{aligned} D &= \frac{\delta_x^2}{2\tau} \\ &= \frac{\delta_x}{2} v \\ &= \frac{\delta_x}{2} \sqrt{kT/m} \\ \delta_x &= \frac{2D}{\sqrt{kT/m}} \end{aligned}$$

The average time between collisions is given by:

$$\begin{aligned} \tau &= \delta_x / v \\ &= \frac{2D}{\sqrt{kT/m}} \frac{1}{\sqrt{kT/m}} \\ &= \frac{2D}{kT/m} \end{aligned}$$

What is implied by these equations?

- The average velocity of a molecule depends on the temperature and mass, irrespective of the surrounding environment.
- But, the average distance that a molecule goes before colliding and bouncing off in different directions does depend on the environment.
- All molecules at a given temperature have the same average kinetic energy. (This is implied by the ideal gas law.) However, the average velocity is inversely related to the square root of the molecular mass. Big molecules move more slowly.

IV. Some typical values of D, δ_x and τ.

From the simple diffusion experiment, we estimated that the diffusion coefficient for the bromophenol blue dye is about $10^{-10}\,\mathrm{m}^2/\mathrm{s}$. The molecular weight of this dye is $670\,\mathrm{g/mol}$. Thus, the mass of a single molecule is $1.1{\times}10^{-21}$ g, or $1.1{\times}10^{-24}$ kg. At room temperature ($\approx$300 K), the expected velocity of the molecule is then:

$$\begin{aligned} v &= \sqrt{kT/m} \\ &= \sqrt{\frac{1.38 \times 10^{-23}\,\mathrm{kg \cdot m^2 \cdot s^{-2} \cdot K^{-1}} \cdot 300\,\mathrm{K}}{1.1 \times 10^{-24}\,\mathrm{kg}}} \\ &\approx 60\,\mathrm{m/s} \end{aligned}$$

This seems very fast, especially considering how slowly the molecules diffused. But, we know that they only move in a given direction for a short time before colliding with another molecule in the solution. From the relationships derived earlier, and the estimate of the diffusion coefficient, we can calculate the distance between collisions as:

$$\begin{aligned}\delta_x &= 2D/v \\ &\approx 2 \times 10^{-10}\,\mathrm{m^2/s} \div 60\,\mathrm{m/s} \\ &\approx 3 \times 10^{-12}\,\mathrm{m}\end{aligned}$$

Thus, the average displacement is extremely small: A hydrogen atom is about $10^{-10}\,\mathrm{m}$ in diameter. The time interval between collisions is correspondingly small:

$$\begin{aligned}\tau &= \delta_x/v \\ &\approx 3 \times 10^{-12}\,\mathrm{m} \div 60\,\mathrm{m/s} \\ &\approx 5 \times 10^{-14}\,\mathrm{s}\end{aligned}$$

The RMS displacement along one axis as a function of time is given by:

$$\begin{aligned}\mathrm{RMS}(x) &= \sqrt{2Dt} \\ &\approx \sqrt{2 \times 10^{-10}\,\mathrm{m^2/s} \cdot t(\mathrm{s})} \\ &\approx 1.4 \times 10^{-5}\,\mathrm{m}\sqrt{t(\mathrm{s})}\end{aligned}$$

V. The relationship between molecular size and diffusion coefficient

In general, we expect the diffusion coefficient to depend on the molecule and its environment. More specifically, we might expect D to depend on the size of the molecule, the temperature and the viscosity of the solution. Indeed, this is the key relationship that Einstein formulated in his classic 1905 paper on Brownian motion:

$$D = \frac{kT}{6\pi\eta r}$$

where η is the viscosity of the solution and r is the radius of a spherical particle. This is usually referred to as the Stokes-Einstein equation, showing that even Einstein built on the work of others! Strictly, this applies only to spherical particles, but it is common to refer to an "effective radius" for particles that are less symmetrical.

We will ignore the question of how viscosity is defined and measured except to note that it is most commonly expressed in units of centipoise, which is equivalent to $10^{-3}\,\mathrm{N \cdot s \cdot m^{-2}}$. The factor of 10^{-3} has obscure historical origins, but the unit of centipoise has been retained, probably because water at room temperature has a viscosity very close to 1 centipoise. It is left to the student to demonstrate that the units in the Stokes-Einstein equation are consistent.

Here are a few examples of particles with a range of sizes and calculated diffusion coefficients:

- Small molecule (1 nm): $2 \times 10^{-10}\,\mathrm{m^2 s^{-1}}$
- Protein (10 nm): $2 \times 10^{-11}\,\mathrm{m^2 s^{-1}}$
- Bacterium (1 μm): $2 \times 10^{-13}\,\mathrm{m^2 s^{-1}}$
- 1 mm sphere: $2 \times 10^{-16}\,\mathrm{m^2 s^{-1}}$

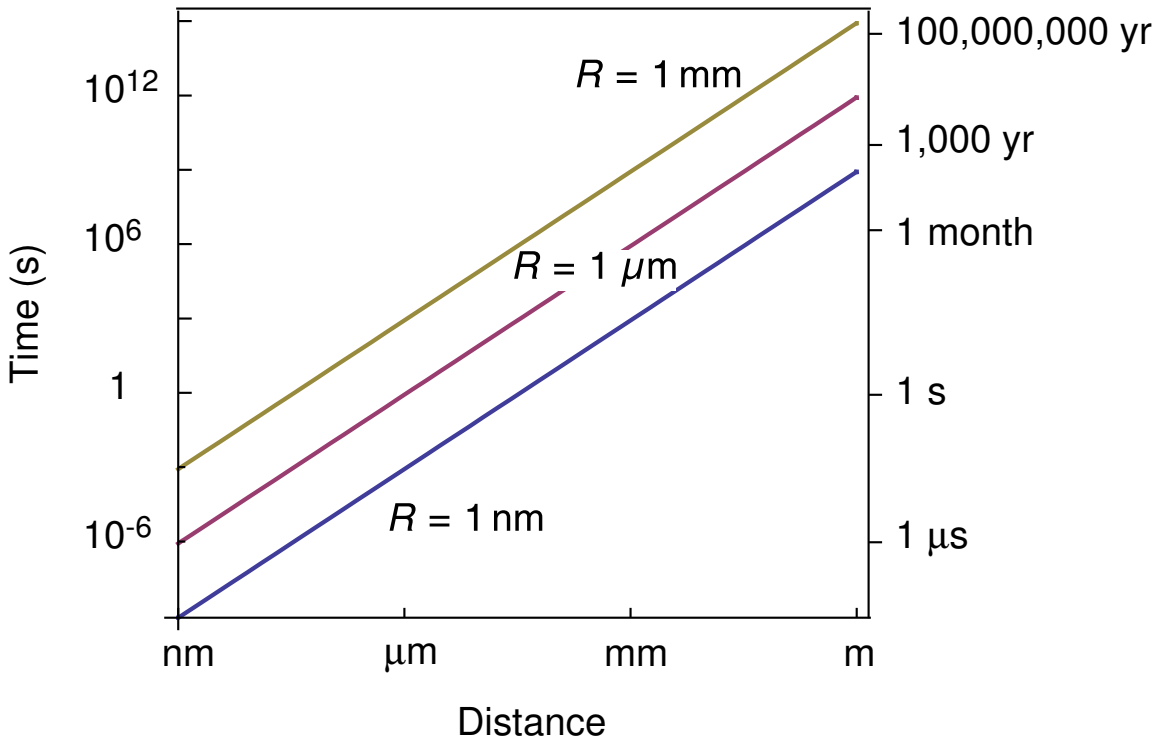
To place these diffusion coefficients into some context, it is helpful to calculate the time required for a particle to diffuse until the RMS distance from the starting point reaches a given value. Earlier, the relationship between the RMS distance and time was given as:

$$\mathrm{RMS}(x) = \sqrt{2Dt}$$

Rearranging this equation gives:

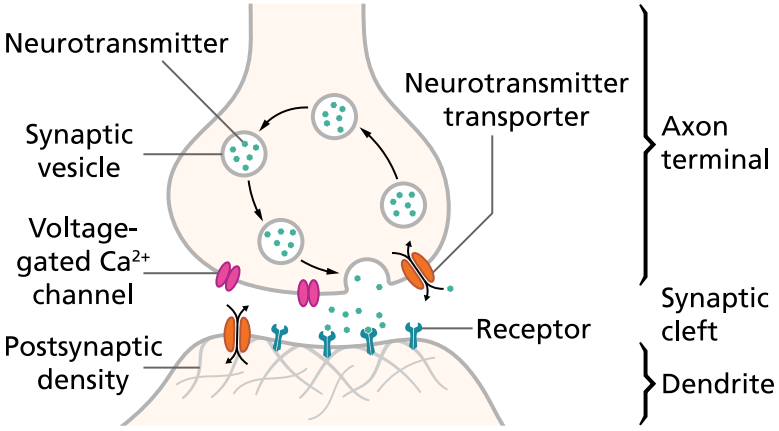
$$t = \frac{\mathrm{RMS}(x)^2}{2D}$$

The graph below shows how the time required to diffuse a given RMS distance depends on time for different size molecules:



This graph shows that diffusion leads to quite fast net movements of molecules over short distances, but the times required for movement over longer distances can greatly exceed what is necessary in many biological contexts.

A biological example highlighting the differences in diffusion times over different distances is neural transmission. Individual neurons communicate with adjacent neurons and muscle cells via chemical synapses, as diagrammed below[1]:



When stimulated, the axon terminus of a neuron (the pre-synaptic cell) releases neurotransmitter molecules, such as acetylcholine, glutamate or dopamine. These small molecules diffuse across the synaptic cleft, which has a width of about 20 nm, and bind to receptors on the adjacent neuron or muscle cell (the post-synaptic cell). Binding to these receptors then generates a signal within the post-synaptic cell. The time required for the transmitter to diffuse an RMS distance of 20 nm is calculated (assuming a diffusion coefficient of $2 \times 10^{-10}\,\mathrm{m^2 s^{-1}}$) as:

$$\begin{aligned} t &= \frac{\mathrm{RMS}(x)^2}{2D} \\ &= \frac{(2 \times 10^{-8}\,\mathrm{m})^2}{2 \times 2 \times 10^{-10}\,\mathrm{m^2 s^{-1}}} \\ &= 10^{-6}\,\mathrm{s} = 1\,\mu\mathrm{s} \end{aligned}$$

Signals also must travel along the lengths of neurons, which can be up meters in length. To travel by diffusion a distance this long, would take:

$$\begin{aligned} t &= \frac{(1\,\mathrm{m})^2}{2 \times 2 \times 10^{-10}\,\mathrm{m^2 s^{-1}}} \\ &= 2.5 \times 10^{10}\,\mathrm{s} \approx 80\,\mathrm{yr} \end{aligned}$$

The obvious conclusion from this calculation is that some other mechanism must be employed to transmit signals over the length of the neuron, and this mechanism is the propagation of an electrical potential across the membrane. Within neurons and other eukaryotic cells, the components of various structures must be transported over distances that are similarly too long for diffusion to be effective. Molecular motors that move along structural tracks in the cell facilitate this kind of motion.

[1]Figure from https://en.wikipedia.org/wiki/Chemical_synapse, Thomas Splettstoesser (www.scistyle.com).

4.6 A Plant Faces Diffusion

Diffusion plays a role in the physiology of all organisms, as they exchange nutrients and other compounds with their environments. Here we consider diffusion at the surface of plant leaves, a process that dictates many aspects of plant physiology, structure and ecology.

I. A plant's demand for CO_2

Consider the growth of a seed to a plant. Where does all of the mass, especially the carbon, come from? Nearly all of the carbon comes, literally, from thin air. The net chemical reaction is:

$$6CO_2 + 6H_2O \longrightarrow C_6H_{12}O_6 + 6\,O_2$$

This is an extremely unfavorable chemical reaction, except when it is coupled to the absorption of energy provided by sunlight.

A back of the envelope calculation: How much CO_2 must cross the leaf surface per second to support a plant's growth? Suppose that a plant incorporates 1 kg of carbon a year. How much leaf area does such a plant have? A rough estimate might be that the plant has 1,000 leaves with an area of $1\,cm^2$ each, for $0.1\,m^2$ total.

Total moles of carbon per year:

$$1\,kg \div 12\,g/mol \approx 80\,mol$$

Total seconds per year:

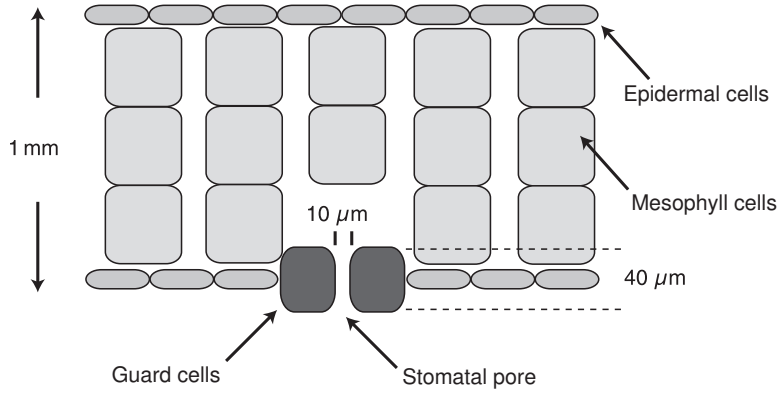
$$1\,yr \times 365\,days/yr \times 24\,hr/day \times 60\,min/hr \times 60\,s/min \approx 3 \times 10^7\,s$$

But, CO_2 is incorporated only during daylight, so the total time available is only about half of this. The number of moles per second is:

$$80\,mol \div 1.5 \times 10^7\,s \approx 5 \times 10^{-6}\,mol/s$$

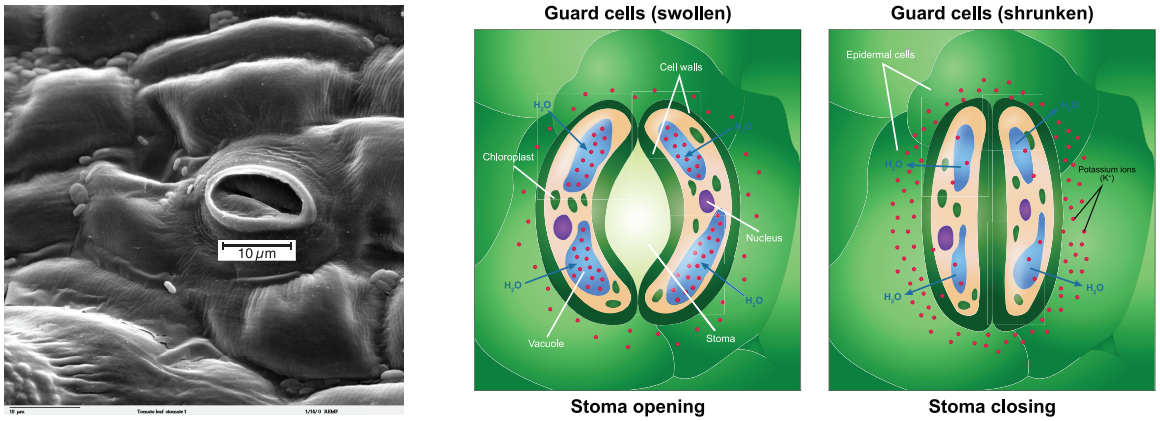
II. Leaf structure and stomata

CO_2 enters leaves only through special openings, called stomata, which can be regulated depending on physiological state. A rough cross-sectional drawing of a typical plant leaf is shown below.



The cells that carry out photosynthesis, the mesophylls, are enclosed in a space bounded by the leaf epidermis on both sides. In some plants, the stomata are found only on the lower leaf surface, but in others they are on both surfaces. Photosynthesis takes place within the chloroplasts of the mesophylls, and the CO_2 is converted into sugars by the enzyme ribulose-1,5-bisphosphate carboxylase (Rubisco). This is the most abundant enzyme on earth, and is arguably the most important for life. Essentially every carbon atom in our bodies passes through this enzyme.

The left-most panel in the figure[2] below is a scanning electron micrograph of a tomato-plant leaf. The large gaping opening is a single open stoma, with a diameter of about $10\,\mu$m and a depth of about $40\,\mu$m.



The two other panels in the figure are diagrams of a stoma in the open and closed states. The opening is controlled by two large cells, called guard cells on either side, which expand and change shape when their water content increases to open the stoma. When the water content decreases, the guard cells contract, and the opening closes.

III. Diffusion of CO_2 through stomata

In the air, CO_2 is a trace gas, making up about 580 parts per million of the atmosphere by mass. This number has increased with the burning of fossil fuels, which is, of course, a very important issue right now. At sea level, the concentration of CO_2 is about $1.5 \times 10^{-2}\,\mathrm{mol \cdot m^{-3}} = 15\,\mu$M.

Within the leaf, the consumption of CO_2 by the chloroplasts depletes the concentration in the leaf airspace. The flux through the leaf involves many steps and concentration gradients, but the most significant barrier to diffusion is in the stomata. In the airspace, the CO_2 concentration is about half of what it is in the atmosphere, *i.e.,* about $7.5 \times 10^{-3}\,\mathrm{mol \cdot m^{-3}}$. So, the concentration difference across the stomata is about $7.5 \times 10^{-3}\,\mathrm{mol \cdot m^{-3}}$.

Recall Fick's first law:

$$J = -D\frac{dC}{dx}$$

[2]Scanning electron micrograph of tomato-leaf surface by Louisa Howard, http://remf.dartmouth.edu/images/botanicalLeafSEM/source/16.html
Diagrams of open and closed stomata by Ali Zifa https://en.wikipedia.org/wiki/Stoma

where J is the flux, D is the diffusion coefficient, C is concentration and x is distance along the direction of diffusion.

For the stomata:

$$\frac{dC}{dx} = \frac{1.5 \times 10^{-2}\,\mathrm{mol \cdot m^{-3}} - 0.75 \times 10^{-2}\,\mathrm{mol \cdot m^{-3}}}{40 \times 10^{-6}\,\mathrm{m}} = 190\,\mathrm{mol \cdot m^{-4}}$$

The diffusion coefficient for CO_2 at atmospheric pressure is $1.5 \times 10^{-5}\,\mathrm{m^2 s^{-1}}$. This is much larger than the numbers we discussed for molecules in water. Why?

The flux, per unit of surface area, is then:

$$\begin{aligned} J = -D\frac{dC}{dx} &= -1.5 \times 10^{-5}\,\mathrm{m^2 s^{-1}} \times 190\,\mathrm{mol \cdot m^{-4}} \\ &= -2.8 \times 10^{-3}\,\mathrm{mol \cdot m^{-2} s^{-1}} \end{aligned}$$

How much surface area do we need in order to fix 1 kg of CO_2 per year?

$$5 \times 10^{-6}\,\mathrm{mol \cdot s^{-1}} \div 2.8 \times 10^{-3}\,\mathrm{mol \cdot m^{-2} s^{-1}} \approx 0.002\,\mathrm{m^2}$$

We can then calculate the minimal number of stomata required to allow this transfer of CO_2 into the leaves. The area of each stomatal pore is:

$$\pi(5 \times 10^{-6}\,\mathrm{m})^2 \approx 10^{-10}\,\mathrm{m^2}$$

and the number of pores required is:

$$0.002\,\mathrm{m^2} \div 10^{-10}\,\mathrm{m^2} = 2 \times 10^7$$

If the plant has a total of $0.1\,\mathrm{m^2}$ of leaf area, then about 2% of that must be devoted to stomatal pores. A plant that grows by 1 kg/yr might have about 1,000 leaves of $1\,\mathrm{cm^2} = 10^{-4}\,\mathrm{m^2}$ each, so that there would be about 20,000 stomata per leaf, or 200 stomata per $\mathrm{mm^2}$. Actual densities of stomata on plant leaves range from 100 to 1,000 per $\mathrm{mm^2}$, depending on plant species and environmental conditions.

IV. The big problem: Water diffusion

Because the stomata are just open holes in the leaf surface, other gasses can also diffuse in and out of the airspace. The diffusion of water out of leaves, *transpiration*, is a major factor that limits the ability of plants to fix CO_2.

Within the airspace of the leaf, water reaches saturation concentration, *i.e.,* close to 100% humidity. For a leaf at 25 °C, this is about $1.3\,\mathrm{mol \cdot m^{-3}}$. Outside the leaf, the water vapor concentration is about half this.

So, we can calculate the H_2O vapor gradient across the stomata:

$$\frac{dC}{dx} = \frac{0.6\,\mathrm{mol \cdot m^{-3}}}{40 \times 10^{-6}\,\mathrm{m}} = 1.5 \times 10^4\,\mathrm{mol \cdot m^{-4}}$$

The diffusion coefficient is slightly larger for H_2O than for CO_2. Why? D $=2.4 \times 10^{-5}\,m^2s^{-1}$.

The flux per unit area of stomata is:

$$\begin{aligned} J = -D\frac{dC}{dx} &= -2.4 \times 10^{-5}\,m^2s^{-1} \times 1.5 \times 10^4\,mol \cdot m^{-4} \\ &= -0.36\,mol \cdot m^{-2}s^{-1} \end{aligned}$$

For our plant with $0.1\,m^2$ of leaf surface area and $0.002\,m^2$ of stomatal surface area:

$$0.36\,mol \cdot m^{-2}s^{-1} \times 0.002\,m^2 = 7 \times 10^{-4}\,mol/s$$

In one year:

$$\begin{aligned} &1.5 \times 10^7\,s \times 7 \times 10^{-4}\,mol/s = 10^4\,mol \\ &10^4\,mol \times 18\,g/mol = 18 \times 10^4\,g \\ &= 180\,kg \\ &\approx 45\,gal \end{aligned}$$

The plant needs about 180 kg of water for each kg of carbon it fixes, just because of evaporation from the leaves.

Note: This is a very rough approximation!

This dwarfs the amount of H_2O directly used in the photosynthesis reactions ($\approx 80\,mol \approx 1.4\,kg$).

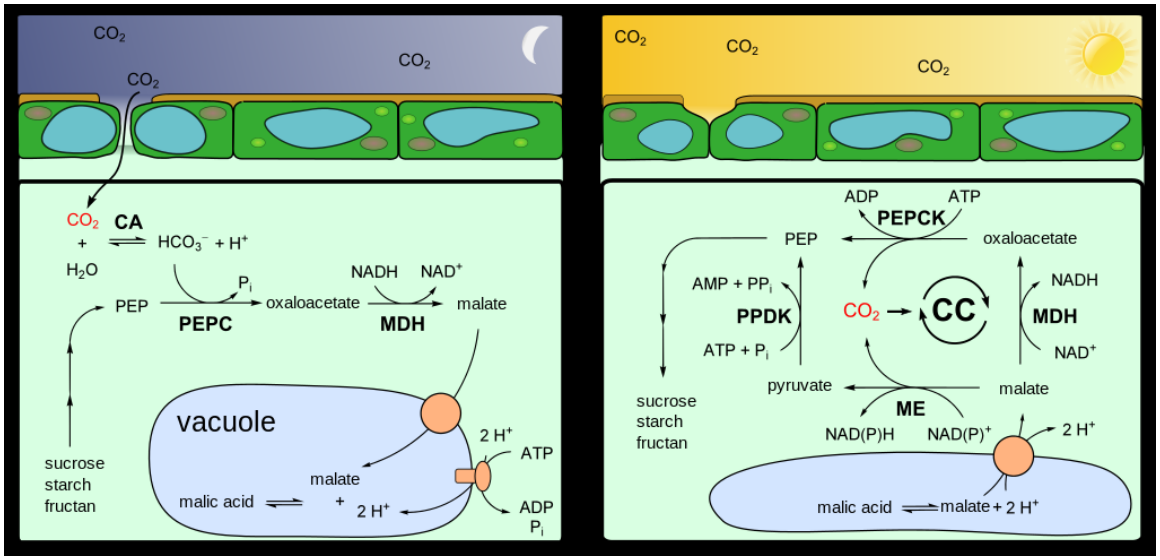
Consequences of huge water losses:

- Stomata are closed when photosynthesis rates are low (*e.g.,* at night). This is probably why plants evolved stomata, rather than allowing diffusion across the entire leaf area.
- The tradeoff between photosynthesis and water loss is the major physiological challenge to plants. Plants in different environments evolve to optimize this tradeoff.
- All of this water has to pass through roots and stems of the plant.
- For tall trees, there is a huge pressure difference from the bottom to the top of the trunk of the tree. Conduction depends on unbroken flow of liquid. If bubbles form, conduction stops. Tree trunks have small parallel tubes to conduct water, the xylem. If one develops a cavity, it is sealed off.
- For trees, each year there is a discontinuity in the flow, and new tissue has to be grown, leading to rings.

V. The Crassulacean Acid Metabolism Cycle

In some plant lineages, special adaptations have evolved to minimize the loss of water through stomata, particularly in species that live in arid environments. One of

these adaptations is based on a metabolic pathway, the crassulacean acid metabolism (CAM) cycle, which allows CO_2 to be captured at night and then incorporated into carbohydrates during the day. The name, crassulacean, comes from the plant family *Crassulacae* (which includes the pineapple and jade plants), where the pathway was first studied in detail. The overall cycle is illustrated in the figure[3] below:



During the night, when water transpiration is minimal because of lower temperatures, the stomata are open to allow CO_2 into the leaves. Because, there is no sunlight, however, the plants are not able to convert the CO_2 into carbohydrate via photosynthesis. Instead, the CO_2 is used to convert phosphoenolpyruvate (PEP) into oxaloacetate, which is then converted to malate. The resulting malate is then stored in vacuoles. During the daytime, the stomata are closed, to prevent transpiration, and the malate that accumulated overnight is metabolized to regenerate CO_2, which is used for photosynthesis.

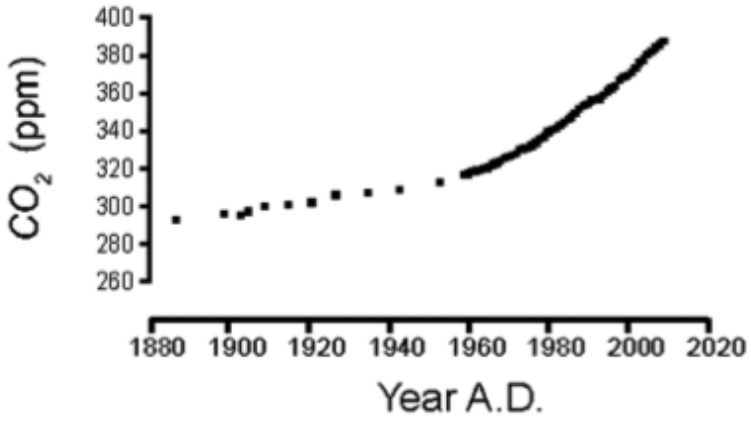
By temporally separating the processes of CO_2 diffusion and photosynthesis, plants using this cycle minimize the loss of water. But, this does come with a cost in metabolic energy, including the hydrolysis of ATP to drive the reformation of PEP and for transport of malate into vacuoles.

VI. Changes in atmospheric CO_2 concentration

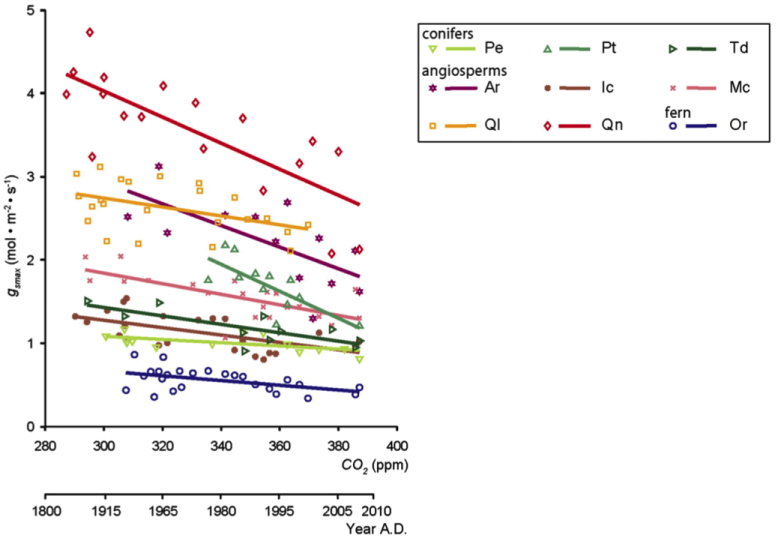
It is now well established that the atmospheric concentration of CO_2 has increased markedly over the past century, as shown in the graph below[4]:

[3] Figure from https://en.wikipedia.org/wiki/Crassulacean_acid_metabolism

[4] The figures in this section are from Lammersma, E. I., de Boer, H. J., Dekker, S. C., Ditcher, D. L., Lotter, A. F. & Wagner-Cremer, F. (2011). Global CO_2 rise leads to reduced maximum stomatal conductance in Florida vegetation. *Proc. Natl. Acad. Sci., USA*, 108, 4035–4040. http://dx.doi.org/10.1073/pnas.1100371108



In molar units, the increase in CO_2 concentration from 300 to 400 ppm is an increase from 12 to 16 μM. This increase in CO_2 concentration could, in principle, be beneficial for plants, both by increasing the efficiency of CO_2 fixation and by minimizing the loss of water through stomata. It appears that plants have responded to the relatively recent increase in CO_2 concentration by reducing the total area of open stomata on their leaves. The results of a study examining the nature of this change in nine plant species found in Florida is shown below:



The quantity plotted on the vertical axis of this plot is termed the *anatomical maximal stomatal conductance to water*, g_{smax} and has the units of $\mathrm{mol} \cdot \mathrm{m}^{-2}\mathrm{s}^{-1}$. Note that these are the same units as used for flux, J, but g_{smax} is made up of several terms, as defined by the equation:

$$g_{\mathrm{smax}} = \frac{d_{\mathrm{wat}} D \cdot a_{\mathrm{max}}/v}{l + \sqrt{\pi a_{\mathrm{x}}}/2}$$

where d_{wat} is the diffusion coefficient of water, D is the density of stomata on a leaf, a_{max} is the maximum open area of an individual stoma, l is the length of the stomatal

pore and v is the molar volume of air (the inverse of molar concentration, with units m^3/mol). g_{smax} represents the maximum conductivity of water per unit of leaf area and reflects both the dimensions of the stomata and their density on the leaves of a plant.

The important message from the figure shown above is that conductance per unit of leaf area has decreased over the time period when the atmospheric CO_2 concentration has increased. Analogous changes have been demonstrated over geological timescales, when atmospheric CO_2 concentrations have both increased or decreased, and stomatal conductance has decreased or increased, respectively.

In different plant species, the reduction in overal stomatal surface area is due to a reduction in stomatal density or in pore size, or both. For the species of Florida plants examined in this study, the major change appears to be in the density of stomata on the leaf surface, rather than changes in the size of the stomata.

4.7 Bacterial Chemotaxis: Overcoming the Limits of Diffusion

Microorganisms are subject to Brownian motion and can diffuse over short distances, but the requirements for traveling longer distances has led to the evolution of special mechanisms, referred to as chemotaxis. Among the best studied examples of chemotactic microorganisms are the closely related gram-negative bacteria *Escherichia coli* and *Salmonella enterica*[5].

I. Diffusion from a bacterial perspective

As an approximation, we will treat a bacterial cell as a sphere of $1\,\mu$m radius. First, we calculate the diffusion coefficient from the Stokes-Einstein equation:

$$D = \frac{kT}{6\pi\eta r}$$

η is viscosity and r is the radius of the particle.

$$k = 1.38 \times 10^{-23}\,\mathrm{kg \cdot m^2 s^{-2} K^{-1}}$$

$$T = 300\,\mathrm{K}$$

$$kT = 4.1 \times 10^{-21}\,\mathrm{kg \cdot m^2 s^{-2}}$$

$$\eta = 10^{-3}\,\mathrm{N \cdot s \cdot m^{-2}} = 10^{-3}\,\mathrm{kg \cdot m^{-1} s^{-1}}$$

$$D = \frac{4.1 \times 10^{-21}\,\mathrm{kg \cdot m^2 s^{-2}}}{6\pi 10^{-3}\,\mathrm{kg \cdot m^{-1} s^{-1}} \cdot 10^{-6}\,\mathrm{m}}$$

$$= 2.2 \times 10^{-13}\,\mathrm{m^2 s^{-1}}$$

[5]The species *S. enterica* is classified into smaller groups, called serovars, and the serovar that has been most extensively studied with respect to chemotaxis is Typhimurium. Until recently this serovar was identified as a species, *Salmonella typhimurium*. It's very confusing.

This is about 1/1,000 of the value for a small molecule.

A short cut: Remember a few key facts:

- A "small molecule" (≈ 100 Daltons) has a radius of $r \approx 1\,\text{nm}$.
- A molecule of this size has a diffusion coefficient of about $10^{-10}\,\text{m}^2\text{s}^{-1}$.
- The diffusion coefficient is inversely proportional to the radius of a particle.

A bacterium with a radius of $1\,\mu\text{m}$ should have a diffusion coefficient of about 1,000th that for a small molecule, or about $10^{-13}\,\text{m}^2\text{s}^{-1}$.

Next, we calculate the velocity of the bacterial cell during its random-walk steps, using the relationship:

$$\text{RMS}(v) = \sqrt{kT/m}$$

We need to know the mass. Bacteria (and the great majority of all organisms) have a density that is about the same as water. (Because they are mostly made up of water!) The density is about $1\,\text{g/mL} = 1\,\text{kg/L}$. So if we know the volume we should be able to make a reasonable estimate of the mass.

$$\begin{aligned} V &= \frac{4}{3}\pi r^3 = \frac{4}{3}\pi(10^{-6}\,\text{m})^3 \\ &= 4.2 \times 10^{-18}\,\text{m}^3 \end{aligned}$$

$1\,\text{m}^3 = 10^3\,\text{L}$, so we can calculate the mass as:

$$\begin{aligned} m &= 4.2 \times 10^{-18}\,\text{m}^3 \times \frac{10^3\,\text{L}}{1\,\text{m}^3} \times \frac{1\,\text{kg}}{1\,\text{L}} \\ &= 4.2 \times 10^{-15}\,\text{kg} \end{aligned}$$

The average velocity is:

$$\begin{aligned} \text{RMS}(v) = \sqrt{kT/m} &= \sqrt{\frac{4.1 \times 10^{-21}\,\text{kg}\cdot\text{m}^2\text{s}^{-2}}{4.2 \times 10^{-15}\,\text{kg}}} \\ &= \sqrt{10^{-6}\,\text{m}^2\text{s}^{-2}} \\ &= 10^{-3}\,\text{m/s} \end{aligned}$$

About $1\,\text{mm/s}$: Much slower than the small molecules, which have a velocity of about $100\,\text{m/s}$.

The random-walk step size is then calculated as:

$$D = \frac{\delta_x^2}{2\tau} = \frac{v}{2}\delta_x$$

$$\delta_x = \frac{2D}{v}$$

$$\delta_x = \frac{2 \times 2.2 \times 10^{-13}\,\mathrm{m^2 s^{-1}}}{10^{-3}\,\mathrm{m/s}}$$

$$\delta_x = 4.4 \times 10^{-10}\,\mathrm{m}$$

Compare this with $3 \times 10^{-12}\,\mathrm{m}$ for a small molecule. The average step size increases with particle size, but the velocity decreases much more rapidly

For a three-dimensional random walk, the mean-square end-to-end distance is calculated as:

$$\langle r^2 \rangle = 6Dt$$

where r is the distance, in three-dimensions, from the start to end of a walk, and t is time.

Note that this expressions differs from the one for one-dimensional diffusion, with the factor of 6 replacing 2. The reason for this has to do with the definition of the diffusion coefficient in terms of the random-walk step size, as given above. The parameter δ_x is the average projection of the steps onto the x-axis (or any arbitrary axis, for that matter). If we only consider the net diffusion in one-dimension, the δ_x corresponds to the average step size in that direction. But, if we are considering the diffusion away from the starting point and are calculating the distance through three dimensions, then the projections along all three axes contribute to the distance:

$$\langle r^2 \rangle = \langle x^2 \rangle + \langle y^2 \rangle + \langle z^2 \rangle = 6Dt$$

How long does it take for an average walk to reach 1 mm?

$$\langle r^2 \rangle = (10^{-3}\,\mathrm{m})^2 = 6Dt$$

$$t = \frac{(10^{-3}\,\mathrm{m})^2}{6 \times 2.2 \times 10^{-13}\,\mathrm{m^2 s^{-1}}}$$

$$t = 7.6 \times 10^5\,\mathrm{s}$$

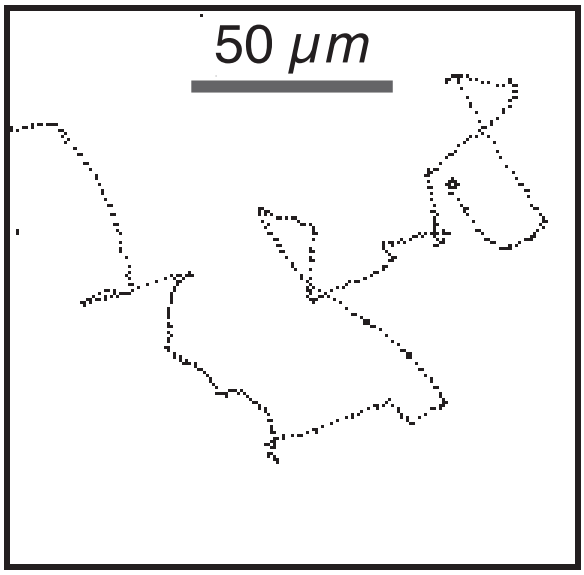
$$t \approx 9\,\mathrm{days}$$

Since an *E. coli* bacterium can divide in as little as 20 min, this is obviously a long time for such an organism!

II. Bacteria under the microscope

When bacterias such as *E. Coli* are examined under a microscope it is often observed that they move *much* faster than the calculated rates of brownian motion. For many of these motile bacteria, the motion appears to be a random walk, but with much larger steps than expected for Brownian motion.

A pioneer in the biophysical study of bacterial swimming is Prof. Howard Berg. In 1972 he built a very fancy microscope, especially for the time, that could track the motion of individual bacteria in three dimensions[6]. The figure below shows an example of one of the paths, projected onto two dimensions, as traced by Berg and his colleagues:



This looks like a random walk with variable step length, and detailed analyses showed that the typical parameters for the random walks were:

- Velocity $\approx 2 \times 10^{-5}\,\mathrm{ms}^{-1}$
- Average time of forward motion $\approx 3\,\mathrm{s}$
- Average step length $\approx 6 \times 10^{-5}\,\mathrm{m}$

Note that the velocity is much lower than the instantaneous velocity from thermal motion. But, the length of the steps is vastly longer, about 60 bacterial body lengths.

The number of steps is $n = t/(3\,\mathrm{s/step})$

[6]Berg, H. C. & Brown, D. A. (1972). Chemotaxis in *Esherichia coli* analyzed by three-dimensional tracking. *Nature*, 239, 500–504. http://dx.doi.org/10.1038/239500a0

What is the average time to move 1 mm? First calculate the number of steps:

$$n = \frac{\langle r^2 \rangle}{\delta^2}$$

$$= \frac{(10^{-3}\,\mathrm{m})^2}{(6 \times 10^{-5}\,\mathrm{m})^2}$$

$$\approx 280\,\mathrm{steps}$$

The total time, then, is:

$$t = 280\,\mathrm{steps} \times 3\,\mathrm{s/step}$$
$$\approx 840\,\mathrm{s} \approx 15\,\mathrm{min}$$

This is almost 1,000 times shorter than the time required for diffusion over the same distance.

This is an important feature of random walks: For a given period of time, the *average* distance from the starting point will be larger if the steps are longer, even if there are fewer of them.

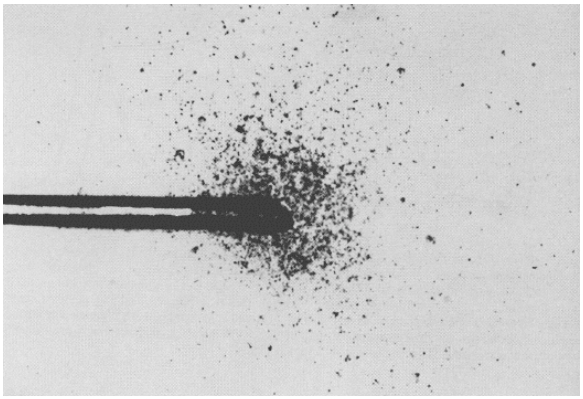
But, why doesn't the bacterium take longer steps? Because it wouldn't help! Notice the tracks in the microscope. They are curved, because Brownian motion is moving them off of a straight path. After a few seconds, the bacterium essentially forgets what direction it is going.

But, it is suddenly changing direction. Why do this if it is going to be randomly altering direction anyway?

Because, it is doing something much smarter!

III. Chemotaxis: Movement to or from specific chemicals

The ability of bacteria to systematically move towards or away from certain compounds was first demonstrated by Wilhelm Pfeffer in 1884 in a very simple experiment. Pfeffer placed a solution of sugar in a small capillary tube and then placed the end of this tube in a liquid culture of bacteria, as illustrated in the figure below[7]:



[7]Figure from: Adler, J. (1969). Chemoreceptors in bacteria. *Science*, 166, 1588–1597. http://dx.doi.org/10.1126/science.166.3913.1588

As shown in the microphotograph, the bacteria quickly cluster around the open end of the capillary tube. This simple experiment demonstrated that bacteria have the capability to detect specific compounds and move in a directed fashion.

The first question these observation raise is, how do they know which way to go? Somehow, they need to detect a concentration gradient and then move in the direction of increased concentration, if they want to use the compound as a nutrient, or decreased concentration, if the compound is toxic. One might imagine that the bacteria could somehow sense concentrations at the two ends of the cell and compare these to determine the concentration gradient. But, our earlier calculation show that the time required for a small molecule to diffuse over the the length of a typical bacterial cell, about 1-2 μm, is less than a second, so that concentration gradients are insignificant over these distance.

Instead, bacteria use a modified random walk strategy that involves the following steps:

1. Choose a random direction.
2. Swim for a while.
3. Is life getting better? (more food, less poison)
    - Yes: keep going.
    - No: Stop and choose a new *random* direction.

Steps in the good direction are still limited to a few seconds, but steps in the wrong direction can be much shorter.

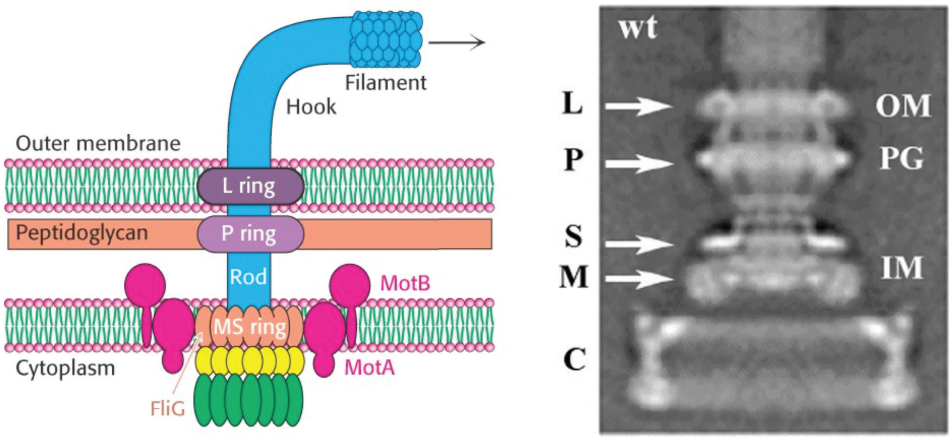
This requires that the bacteria have a concentration sensor and a "memory", so that they can compare concentrations as they move in a particular direction. How do the bacteria actually do all of this?

- The bacteria swim using a propeller and a rotary motor.
- Bacteria change direction by stopping the flagella and tumbling for an instant.
- Sensors on cell surface detect concentration changes and transmit that information to the rotary motor.

IV. The rotary motor

E. coli and many other bacterial species are propelled through liquid media by long helical flagella. Each cell contains multiple flagella, each with a rotary motor embedded in the cell membranes and cell wall, as shown in the diagram in the left panel below[8]. The right panel shows an image of the motor reconstructed from electron micrographs.

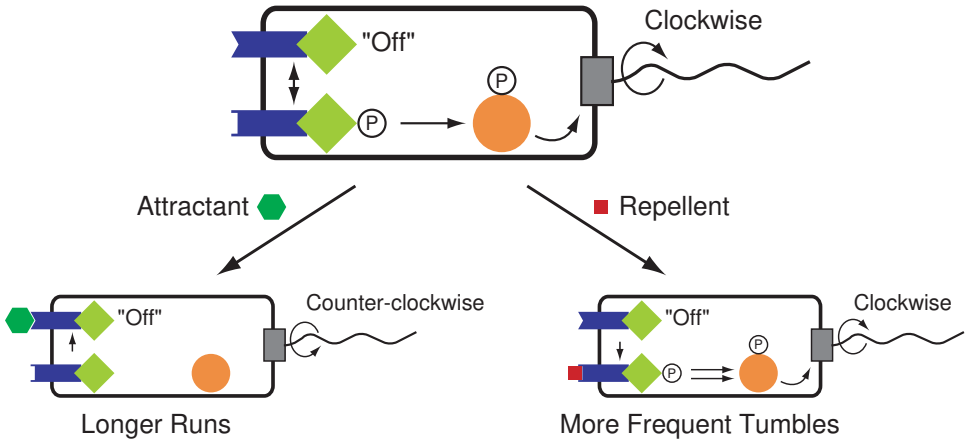
[8]The diagram of the bacterial motor is from Berg, J. M., Tymoczko, J. L. & Stryer, L. (2002). *Biochemistry*. W. H. Freeman, 5^{th} edition. https://www.ncbi.nlm.nih.gov/books/NBK22489/
The electron microscopy reconstruction is from Thomas, D., Morgan, D. G. & DeRosier, D. J. (2001). Structures of bacterial flagellar motors from two FliF-FliG gene fusion mutants. *J. Bacteriol.*, 183, 6404–6412. http://dx.doi.org/10.1128/JB.183.21.6404-6412.2001



The motor, which we will discuss in more detail later in the course, is driven by the flow of H^+ ions from outside of the cell inward and rotates the flagella at up to 10000 RPM. When the motor rotates in the counter-clockwise direction, the individual flagella of a cell bundle together and act as a propeller. The bacterium then swims in a (relatively) straight direction. When the motors reverse direction, the flagella unbundle, and the bacterium tumbles randomly. After a few seconds, the motor reverses again, and the bacterium swims in a new direction. By this process, the bacteria carry out their random walk. The reversals of the motor are controlled by a signalling system that detects changes in the concentrations of specific compounds in the surrounding liquid.

V. The sensory and signalling system

The system for sensing specific molecules and signalling the rotary motor is quite complex, but a simplified diagram is shown below:



Molecules in the extracellular environment are detected when they bind to receptors that cross the cellular membrane. These receptors are bound to an enzyme and can exist in two conformations. In one conformation ("on"), the enzyme, a kinase, is activated and phosphorylates another protein, which, in turns binds to the rotary motor. When this phosphorylated protein is bound to the motor, the clockwise rotation is favored, leading to tumbling. When the receptors are in the other conformation ("off"),

the kinase is inactivated, the signalling molecule is less likely to be phosphorylated, and counter-clockwise rotation is favored.

The equilibrium between the two conformations of the receptors is controlled by multiple factors, including the presence of attractant and repellant molecules. When attractants are bound to the receptors, the off conformation is favored, promoting counter-clockwise rotation and forward swimming. In the absence of attractants or the presence of repellants, the on conformation is favored, leading to tumbling and shorter steps in the random walk.

In addition to sensing the concentrations of attractants and repellants, the bacterium has to do one other very important thing: It has to remember what the concentrations were a short time ago! In order to bias the random walk in the direction of a concentration gradient, the bacterium has to compare the concentrations at different times as it swims. This memory is established by another set of enzymes that covalently modify the receptors by methylating specific glutamate residues. These modifications are reversible and adjust the sensitivity of the receptors to attractants or repellents. As the concentration of an attractant increases, the receptors are modified so that higher concentrations of attractant are required to keep the receptor in the off conformation. In a real sense, the cell becomes addicted to the attractant and require more of it to keep swimming in the same direction. In this way, the random walk is biased in a way that leads it up the concentration gradient. The system adjusts to repellent concentrations in the opposite way, to favor moving down the concentration gradient.

Chapter 5

Thermodynamics

Thermodynamics is the branch of chemistry and physics that is concerned with the interconversion of different forms of energy. The laws of thermodynamics place absolute constraints on how much work can be obtained from different forms of energy and whether or not specific processes will occur spontaneously. Nobody gets to break these laws! Thermodynamics does not tell us how a process will occur, but only whether or not the process will occur spontaneously.

Thermodynamics is definitely one of the most challenging subjects for nearly all students, for several reasons:

- The ideas are fundamentally abstract and subtle. Even people who have thought about thermodynamics for many years easily get tripped up.
- It uses math! Even if many of the ideas seem to be conceptually simple, a deep and useful understanding depends on mathematics.
- The language can be confusing. Different disciplines sometimes use different terms. In addition, the language has historical origins, and the history is convoluted.
- Historical confusion. Unlike classical mechanics, which almost all came from one person, Newton, over a short period of time, thermodynamics was developed over a long period of time by several generations of scientists, and there were periods of profound confusion.

But, it's worth it! The current issues regarding energy and climate change offer a dramatic examples of how important the interconversion of energy forms can be. In the context of our course, we will use these principles to understand how biological systems become organized and do amazing things.

5.1 Energy, Work and Heat

I. Units of energy.

Before going on, it is worthwhile to review again the definition of energy and the units we use. A general definition of energy is that it is the "ability to do work". We define mechanical work as the integral of force applied over distance:

$$w = \int_{x_1}^{x_2} F dx$$

If the force is constant, then

$$w = F(x_2 - x_1)$$

The units of work must be force times distance. The SI unit of force is the newton, which is the force required to accelerate a mass of 1 kg by 1 m/s per s. The SI unit of work or energy is then the Nm, or joule:

$$1\,\text{joule} = 1\,\text{newton} \cdot \text{meter} = 1\,\text{kg} \cdot \text{m}^2\text{s}^{-2}$$

Another commonly used unit of energy is the calorie, which is defined as the amount of energy required to raise the temperature of 1 g of water by 1 °C, or 1 K. One problem with defining the calorie in this way is that the heat required to raise the temperature of water depends on the starting temperature. 4 °C and 15 °C have been used to define the unit, sometimes with a subscript indicating the temperature. The standard definition now is in terms of the Joule, as defined above:

$$1\,\text{cal} = 4.184\,\text{J}$$

Another source of confusion is that the calorie has been defined at different times in terms of the energy to heat either 1 g of water or 1 kg. The current convention is to use a lower case "c" to designate the "gram calorie" and an uppercase "C" to designate the "kilogram calorie". The energy content of foods is expressed in kilogram calories. This is one reason that the idea of loosing weight by eating ice cream (which should cool the body and result in negative calorie intake) is doomed to failure!

II. An important distinction: Temperature versus heat.

Before considering some simple examples, it is important to be sure that we are using language carefully. One common cause of confusion is the difference between temperature and heat.

- Temperature has a relatively straight-forward definition. It is a property of a given mass of matter that is directly related to average kinetic energy of the molecules making up the matter. For an ideal gas, the temperature determines the relationship between volume and gas according to the ideal gas law:

  $$PV = nRT$$

  For an ideal gas, temperature is related to the average kinetic energy of the ideal gas molecules according to:

  $$E = 3kT/2$$

- Heat is often defined as "a form of energy". More specifically, heat is the flow of energy from a warm object to a cooler one, with the result that the temperatures become more equal.

At one time, heat was thought to be a massless substance, often called "caloric", that flowed invisibly between objects. We now understand that heat is not something that objects contain, but the language still seems to imply this. This is one of the historical origins of some of the confusion in thermodynamics.

It is possible to formulate the laws of thermodynamics without invoking "heat", but that seems rather artificial to me and only shifts the difficulties to other words.

- In summary: temperature is a property of matter that can be directly measured, while heat is a flow of energy associated with temperature changes. The heat flow is always from a warm object to a cooler one.

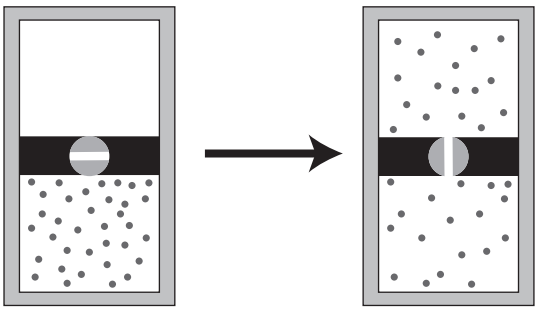
III. Some examples based on the expansion and compression of gasses.

Historically, one of the most important motivations for the development of thermodynamics was the invention of the steam engine. Once steam engines started to be used, there was great interest in getting the maximum amount of work from them. Since steam engines work through the expansion of a gas, many of the ideas in thermodynamics were formulated in this context.

At the time, though, the relationship between the pressure of a gas and the motion of molecules was not understood, and the theory was developed without any explicit model for what actually generated forces or work. This treatment is usually referred to as "classical thermodynamics". Later, a molecular interpretation was developed, what we now call "statistical mechanics". Both can stand on their own and are completely consistent with one another. But, I think that it is often easier to merge the two approaches when trying to understand things.

1. Gas expansion without work

   Suppose we build an apparatus as shown below:



   This apparatus should have the following features:

   - It is completely insulated from its surroundings, so that heat can't flow in or out of it.
   - There are two chambers separated by a valve that can be opened or closed without perturbing anything else and without generating any heat, that is there can be no friction.
   - One chamber is filled with a gas at some arbitrary pressure, P, and temperature,T.

- The other chamber is evacuated.

Once the system has equilibrated, we open the valve. As we have discussed at length, the molecules are moving about due to kinetic energy in random directions. Eventually (quite quickly, actually) we expect the molecules to distribute themselves throughout the vessel, with roughly equal numbers on the two sides.

What can we say about what has changed and what hasn't?

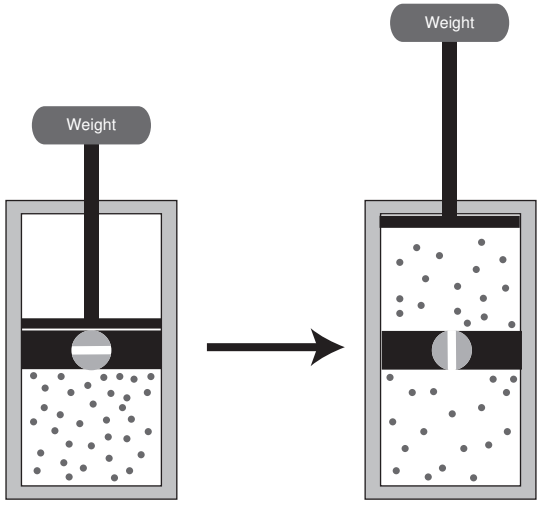
- Nothing has happened to change the average velocities of the molecules. Therefore, the temperature should be the same, and the average kinetic energy is the same.
- The volume has increased. So, by the ideal gas law, $PV = nRT$, the pressure should decrease.
- Because of the insulation, no heat has flowed in or out of the chambers.
- Nothing, except the molecules, has been moved, so no work has been done.

But, something else has happened. Even though no work was done and the energy has stayed the same, we know that we would have to do work, *i.e.,* spend energy, to restore the system to its original state. You probably already know the fancy word for this, "entropy", but let's put off discussing this concept. What we can say, though, is that somehow or other we have wasted the potential to do work, even though the total thermal energy has stayed the same.

Processes in which there is no exchange of heat are called "adiabatic".

2. Gas expansion with work, but without heat flow.

   Now, let's think about another type of apparatus. Again we will keep the apparatus fully insulated. But, now a movable piston has been added to the upper chamber. The upper chamber is still evacuated, and our engineers have made a perfect, frictionless seal to the outside, through which a rod connects the piston to a weight.



When we open the valve, the gas molecules begin to collide with the piston, and the pressure from the gas molecules pushes the piston and the weight upward. Mechanical work is now being done. What is changing as the piston moves?

- Does the temperature change? Consider what happens when the piston is pushed upwards. The gas molecules collide with the piston and transfer some of their energy to it. Unlike when they collide with a fixed wall, their velocity is not quite as great when they bounce off in the other direction. As a consequence the average kinetic energy of the molecules decreases, meaning that the temperature decreases as well.

- If the volume increases by the same amount as in the previous example and the temperature decreases, then the pressure must decrease *more* than in the previous example.

The key point about this example is that some of the kinetic energy of the gas molecules has been converted to mechanical work. From the conservation of momentum, the change in energy should be equal to the amount of work done.

Let's put some labels on the quantities involved:

- E = energy of the molecules. For an ideal gas, this is entirely translational kinetic energy. We will define the change in energy associated with a process as:

$$\Delta E = E_{\text{final}} - E_{\text{start}}$$

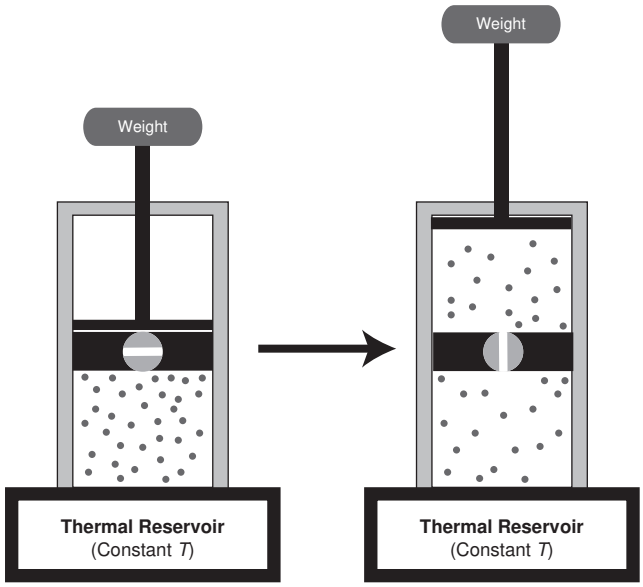
It's important to keep track of the signs! We will define the work involved, w, so that it is positive when work is done on the gas (or, more generally the "system"), and negative when the system does work on the outside world, as in this case.
For this example, we know that the amount of work must equal the change in energy:

$$\Delta E = w$$

Is this always true?

3. Gas expansion at constant temperature with work.

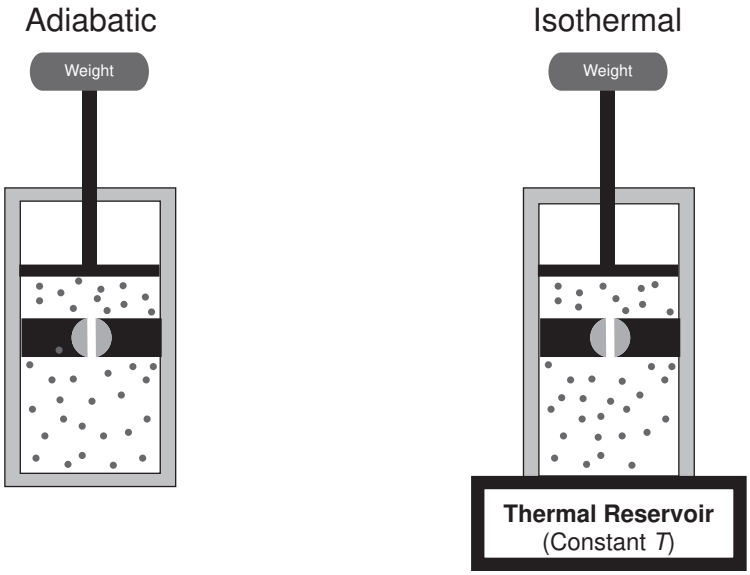
Next, we have the engineers build an even fancier device by adding a thermal reservoir at the bottom of the cylinder:



Basically, the reservoir is a large mass equilibrated at the same temperature as the gas. If the temperature of the gas drops, heat will flow to equalize the two temperatures, without significantly lowering the temperature of the reservoir.

What happens, now, when we allow the piston to move and do work?

- As the piston is pushed upwards, the gas molecules loose some energy, and the temperature starts to drop. But, as soon as that happens, heat flows from the reservoir.
- At the end of the process, the temperature is the same as when it started, so the energy must be the same as well. In this respect, the result is the same as in the adiabatic expansion without work.
- Again, work has been done. Is more work done in the adiabatic or isothermal process? Consider a point part way through each process, where the volumes have increased by the same amount, as illustrated below:



Because the adiabatic process does not allow heat to flow to the gas, the temperature at this intermediate point must be lower for the adiabatic process than the isothermal one. As a consequence, the pressure must be lower for the adiabatic process. Since pressure represents force divided by the area over which it is exerted, the total force on the piston must also be less for the adiabatic process than for the isothermal one. We can thus conclude that the isothermal process can produce more work than the adiabatic process.
The energy for the additional work from the isothermal process is drawn from the heat of the reservoir.

In this example, $\Delta E = 0$, but work has been done, and there has been a flow of heat.

We represent the heat flow by the symbol q and define it so that it is positive when heat flows from the surroundings into the system of interest. In this case, q is positive and w is negative, and:

$$q = -w$$

IV. The first law of thermodynamics.

The common statement of the first law is that the energy of the universe is constant. But, the more formal statement is that for any process, the change in energy, E, is the sum of the work done on the system and the heat absorbed from the surroundings:

$$\Delta E = q + w$$

We can see how this applies to the three examples from above:

- Adiabatic expansion without work.

  $q = 0$
  $w = 0$
  $\Delta E = 0$

- Adiabatic expansion with work

  $q = 0$
  $w < 0$
  $\Delta E < 0$
  $\Delta E = w$

- Isothermal expansion with work

  $q > 0$
  $w < 0$
  $\Delta E = q + w = 0$
  $q = -w$

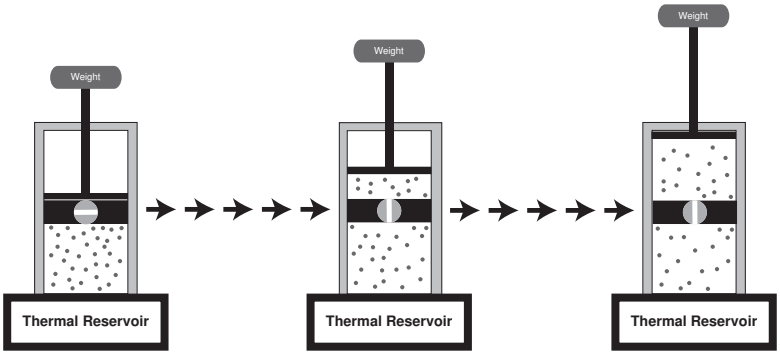
Is there a proof of the first law? No! The laws of thermodynamics are postulates, and our confidence in them comes from the fact that no one has ever found an exception.

In 1775 the French Royal Academy of Science effectively made the first law of thermodynamics (before it was called that) a real law by declaring that it would no longer consider patent applications for perpetual motion machines.

V. Reversible expansion and compression

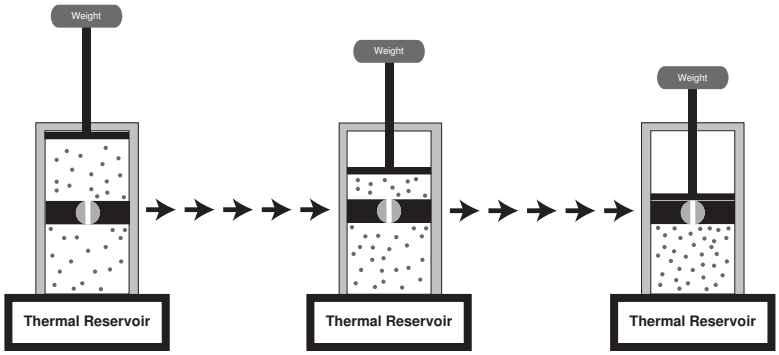
We have already seen that we can get extra work from the expansion of the gas by allowing it to draw heat from its surroundings as it expands. Even if the temperature is the same at the beginning and end of the expansion processes, different paths can lead to different amounts of work being produced and different amounts of heat absorbed.

The maximum amount of work that can be generated from the expansion of a gas is obtained by controlling the movement of the piston so that it is slow enough that the temperature never actually drops. This cannot be done in practice, but it can be approached as a limit, in the sense of limits in calculus, by allowing the piston to move only in infinitesimal steps, as suggested in the drawing below:



Because the piston moves in infinitesimally small steps, the temperature never falls below that of the reservoir. If a larger step ever does take place, the temperature will drop, causing the pressure to drop. As a consequence, less work will be produced by the expansion. This is why we can argue that this is the path that will lead to the maximum production of work (the most negative value of w).

The reverse of this process, diagrammed below, is the one that requires the *least* amount of work to compress the gas to its original volume.



Again, the volume is changed in infinitesimally small steps. As the gas is compressed, the piston imparts extra kinetic energy on the gas molecules, but (in the ideal case) the excess energy is instantly transferred to the thermal reservoir and the temperature remains constant. Because the volume of the gas decreases, the pressure increases, and progressively more work must be done for each decrease in the volume. But, if larger steps were taken, the temperature would increase temporarily, causing a larger increase in pressure and requiring more work for the next step.

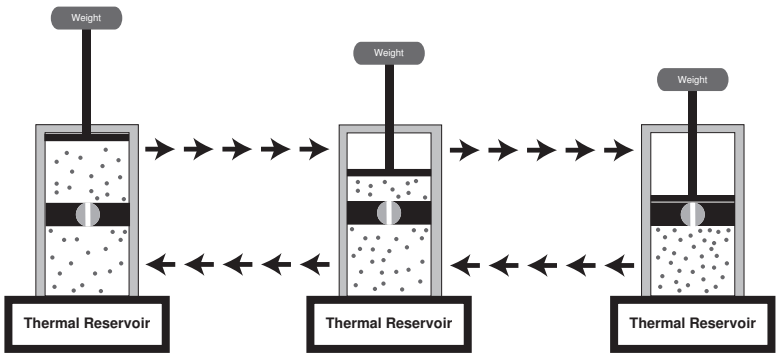
The two processes described above are said to be reversible, and that term can be taken in two senses. In the first sense, the two processes are reversible because they are exact opposites. The maximum work that can be done during expansion is also the minimum amount of work that is required for compression. Thus:

$$w_{\text{exp,rev}} = -w_{\text{comp,rev}}$$

Since $\Delta E = 0$ for both processes, and $q = -w$ for each, we can also conclude that

$$q_{\text{exp,rev}} = -q_{\text{comp,rev}}$$

We can also imagine a cycle formed by the two processes, as illustrated below:



Because $w_{\text{exp,rev}} = -w_{\text{comp,rev}}$ and $q_{\text{exp,rev}} = -q_{\text{comp,rev}}$, w and q for the complete cycle must be zero. Thus, the system is restored to its original volume, pressure, temperature and energy with no net input or output of either work or heat.

Of course, no real process can meet this ideal. As noted above, any real expansion process will produce less work than the ideal, and any real compression will require more work than the ideal. As a consequence, there is a net amount of work applied to the system and w for the cycle will be greater than zero. If the original temperature is restored $\Delta E = 0$ and $q = -w$, and q will be negative, meaning that heat is transferred to the surroundings. This is the general nature of less-than-ideal processes, they require more work (or produce less) than the ideal process and they release more heat.

The ideal expansion and compression processes described above are also reversible in a second sense: Either process can be reversed at a given point by an infinitesimal force in the opposite direction. This is the meaning usually implied by the term in thermodynamics.

VI. The maximum work from expanding a gas at constant temperature

We can calculate the work from the reversible isothermal expansion with a bit of calculus, starting with the integral for any work function:

$$w = \int_{x_1}^{x_2} f dx$$

where x_1 and x_2 are the initial and final positions of the piston, and f is the force. During the reversible gas expansion, the force does not remain constant, but rather drops as the pressure decreases. At any instant, the force is proportional to the pressure:

$$f = -P \cdot A$$

where A is the area of the piston. The negative sign reflects the fact that the work done on the system by expansion will be negative. For each small increment of x, there is a corresponding small increment in the volume of the gas:

$$dV = Adx$$

or, we can write:

$$dx = dV/A$$

We can replace dx in the integral with dV/A, and f with $-P \cdot A$, so that:

$$w = -\int_{V_1}^{V_2} (P \cdot A) dV/A$$

$$= -\int_{V_1}^{V_2} P dV$$

(recall that the product PV has the units of energy, or work!) We can also express pressure in terms of volume, $P = nRT/V$:

$$w = -\int_{V_1}^{V_2} \frac{nRT}{V} dV$$

n, R and T are all constant, so they can be moved outside of the integral.

$$w = -nRT \int_{V_1}^{V_2} \frac{1}{V} dV$$

$$= -nRT \ln(V) \Big|_{V_1}^{V_2}$$

$$= -nRT \ln\left(\frac{V_2}{V_1}\right)$$

Since we have the same number of molecules at the beginning and end of the process, the concentration is inversely related to the volume. So, we can write this expression in terms of concentration:

$$w = -nRT \ln\left(\frac{C_1}{C_2}\right)$$

The form of this expression will likely be familiar to you, or will be soon. This is the origin of all of the expressions like $-RT\ln(C_1/C_2)$ that occur so frequently in chemistry!

Since the energy of the gas molecules is the same at the beginning and end of the process, $\Delta E = 0$, and we can write:

$$\Delta E = 0 = q + w$$
$$q = -w = nRT \ln\left(\frac{V_2}{V_1}\right) = nRT \ln\left(\frac{C_1}{C_2}\right)$$

VII. State functions versus path functions

The quantities E, on the one hand, and q and w, on the other are fundamentally different. The energy of a system depends only on the state of that system, irrespective of how it got there. For an ideal gas, the energy depends only on the temperature and the number of molecules. A change in energy, ΔE, depends only on the starting and ending states.

But, will we always get the same amount of work for a given change in state? Our examples clearly show that we won't. For both the adiabatic expansion without work and the isothermal expansion, ΔE is zero, but q and w are different for the two processes. Furthermore, the amount of work produced by an isothermal expansion (where the temperature is the same at the beginning and end) depends on just how the expansions is carried out.

We say that the energy is a "state function", while heat and work are path-dependent functions. For a gas, the pressure, P and volume, V, are also state functions.

Although work and heat are not state functions, the changes in these quantities associated with the ideal, reversible process separating two states do represent changes in state function. Thus, there are quantities, w_{rev} and q_{rev}, that are associated with any two states that can, in principle, be interconverted by a reversible process.

The quantity w_{rev} is described as being the "free energy change", ΔF, for the conversion of one state to another.[1] ΔF represents the maximum amount of work that can be obtained from a favorable process, or the minimum amount of work required to drive an unfavorable process. The energy change is "free" in the sense that it is the maximum amount of energy that is available to do work. In a more practical sense, free energy is anything but free, since it is the kind of energy that we pay for when we buy, for

[1] More specifically, ΔF is the change in Helmholz free energy, as distinguished from the Gibbs free energy that will be introduced later.

instance, electrical power or gasoline. On the other hand, molecular kinetic energy is "free" in the sense that it is always there, but it can't actually be used to do any work.

The quantity q_{rev} for any two states is also very important, as we will soon see when we discuss the elusive concept of entropy.

5.2 Entropy and the Second Law

The term entropy is commonly associated with the idea of randomness, and this general meaning is used very widely, even entering every-day language. In the context of thermodynamics, where the term originated, a much more specific definition is required, however. There are, in fact, two different and precise ways in which entropy is defined, reflecting the classical and statistical approaches to thermodynamics. Here, we will look at both definitions and show that they are equivalent for at least one process, the expansion of a gas.

I. The classical definition of entropy.

The maximum work that we calculated for the expansion of a gas is simply a function of the temperature and the starting and finishing volumes. This allows us to define this quantity as a state function. It also describes what is *lost* when a gas expands, the ability to do work, and it seems to be related to the increased disorder of the gas, *i.e.,* its entropy.

In fact, the thermodynamic definition of entropy is based on the heat absorbed, q, via the path leading to maximum work, that is the reversible process:

$$\Delta S = \frac{q_{\text{rev}}}{T}$$

where q_{rev} is the heat absorbed by the system in the reversible process, which is $-w_{\text{rev}}$.

We can apply this definition to the case of a gas expansion, when the temperature is the same at the beginning and end of this process. Among the examples we have considered, this could be either the adiabatic expansion without work or the isothermal expansion with work. Because entropy is a state function, the change in entropy is independent of the path between states. *But*, the heat quantity used to calculate the entropy change is the one associated with the reversible process. For the reversible isothermal expansion of a gas, we showed that the heat absorbed is:

$$q_{\text{rev}} = -w_{\text{rev}} = nRT \ln\left(\frac{V_2}{V_1}\right)$$

The entropy change is then calculated as:

$$\Delta S = \frac{q_{\text{rev}}}{T} = nR \ln\left(\frac{V_2}{V_1}\right)$$

The entropy change is easiest to calculate for a process in which the temperature stays constant, but there is still a maximum amount of work that can be obtained for the

transition between two states that have different temperatures, and this maximum amount of work is also obtained through a reversible, (*i.e.,* infinitely slow) path. In this case, the entropy change is calculated as an integral:

$$\Delta S = \int_{T_1}^{T_2} \frac{q_{\text{rev}}}{T} dT$$

where T_1 is the initial temperature and T_2 is the final temperature, and q_{rev} may be a continuously changing function of temperature.

Now, entropy is supposed to be a quantity that increases spontaneously and is related to disorder. In the case of the expanding gas, the quantity we calculated above does seem to be related to disorder, since the molecules are less ordered in a larger volume.

The entropy of a system, as defined here, is a state function. The entropy changes by the same amount irrespective of how much work, for instance, was obtained from the expansion.

But, it is important to point out that this definition only applies to the isolated system. We haven't said anything about how the entropy of the surroundings change. We will come back to this when we talk about the second law, which concerns the entropy of the system *and* its surroundings.

II. The statistical definition of entropy.

The classical definition of entropy is probably not the one that you are most familiar with, which probably refers to the idea of randomness or disorder. In fact, the classical definition is is not an easy concept to work with and is only useful in rather restricted cases, like heat engines.

The important point from the earlier discussions of gas expansion and contraction is that the energy, E, is a function that helps define a state, but the change in energy for a change from one state to another is not sufficient to tell us whether or not the change is favorable or how much work can be obtained. In addition to the internal energy, there is this other quantity that defines the states and determines how much work can be obtained.

In the case of the expanding gas, it appears that the driving force is the tendency of the individual molecules to occupy as large a volume as is made available to them. As we have repeatedly discussed, this is a purely probabilistic phenomenon. If the molecules begin on one side of the container and are allowed to move freely to the other, there is a 50% probability that any one molecule will wind up on the other side. There will be a net flow to the other side until the concentrations on the two sides are equal.

The statistical definition of entropy of a state can be expressed as:

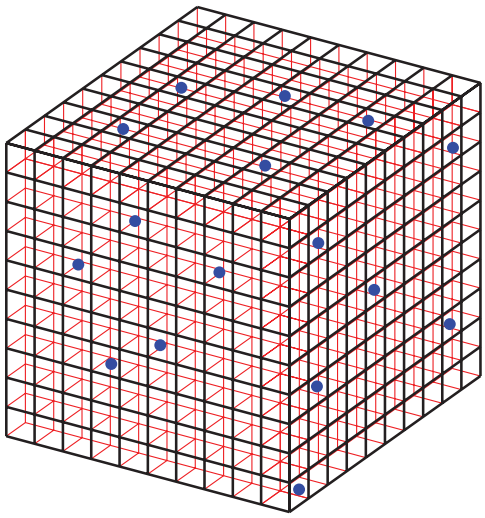
$$S = k \ln \Omega$$

where k is the Boltzmann constant, and Ω is the number of equally probable microstates that make up the state. In other words, Ω is the number of different ways of

arranging the components of the system. This is the fundamental postulate of statistical mechanics. It was deduced (not proven!) by Ludwig Boltzmann, and it is engraved on his tombstone in Vienna. Like the laws of thermodynamics, we believe it because it works!

So, all that we have to do to calculate the entropy is count up the ways of arranging the components. How do we do *that*? This is not at all trivial, and all we can usually do is apply this definition to idealized simple systems. But, the problem is made a little bit easier if we only try to calculate the *change* in entropy for a process. For instance, we can go back to our example of a gas expanding. For even a modest number of molecules, there are a vast number of microstates, each defined by the position and momentum of each molecule. We obviously can't count them all. But, if the temperature stays constant, then the number of possible values for the momentum of each molecule stays the same. So, that all we have to consider is the number of different positions for each molecule.

Suppose that we divide up the initial volume into a grid of small cubes, say 1 nm on a side, as diagrammed below:



If the number of molecules is N and the number of cubes is N_c, then we can calculate the number of ways of arranging the N molecules into the N_c cubes. To simplify things, we can assume that N_c is much larger than N, so that once we place a molecule in a given cube, it doesn't significantly reduce the number of places available to the next molecule. So, taking one molecule at a time, the number of ways of arranging them is:

$$N_c \cdot N_c \cdot N_c \cdot \cdots = N_c^N$$

But, if all of the molecules are indistinguishable, then we have to divide by number of ways of choosing the molecules sequentially, $N!$. So:

$$\Omega = \frac{N_c^N}{N!}$$

So, the initial entropy would be:

$$S = k \ln \Omega = k \ln \left(\frac{N_c^N}{N!} \right)$$

But, this is a little suspect, since the result depends on the number of cubes that I divided the volume up to. Would the entropy suddenly become larger if I decided to use smaller cubes? That doesn't sound right. But, it turns out to be OK if we consider a *change* in volume.

To describe a change in volume, we keep the size of the cubes the same, and the volume change is specified by the change in the number of cubes. If we call the initial number of cubes $N_{c,1}$ and the final number $N_{c,2}$, the change in entropy is given by:

$$\begin{aligned} \Delta S = S_2 - S_1 &= k \ln \left(\frac{N_{c,2}^N}{N!} \right) - k \ln \left(\frac{N_{c,1}^N}{N!} \right) \\ &= k(N \ln(N_{c,2}) - \ln(N!)) - k(N \ln(N_{c,1}) - \ln(N!)) \\ &= kN \ln \left(\frac{N_{c,2}}{N_{c,1}} \right) \end{aligned}$$

Thus, the change in the entropy change depends on the *ratio* of the number of cubes at the beginning and end of the process. If we make the volume of the individual cubes larger or smaller, this ratio is not affected.

If we want to express the entropy change on a molar basis, n, the number of moles, is equal to N divided by Avogadro's number, N_A, and $N = nN_A$:

$$\Delta S = nN_A k \ln \left(\frac{N_{c,2}}{N_{c,1}} \right)$$

Since k is the gas constant, R divided by Avogadro's number, we can write:

$$\Delta S = nR \ln \left(\frac{N_{c,2}}{N_{c,1}} \right)$$

Finally, $N_{c,2}$ and $N_{c,1}$ are proportional to the volumes before and after and we can write the expression as:

$$\Delta S = nR \ln \left(\frac{V_2}{V_1} \right)$$

Notice that this is exactly the same as the expression we derived from the classical definition of the entropy change for this process.

III. Microstates with different probabilities

What if the state is defined by microstates that do not all have equal probabilities? For instance, what if we were to divide up the total volume into cubes of different sizes,

so that they had different probabilities of being occupied. Would the calculated entropy be different? More significantly, what if we were considering a more complicated environment, like the interior of a cell, where certain molecules have specific affinities for different compartments Could we still use this approach to calculate entropy, or at least entropy changes?

The more general expression for the statistical entropy is:

$$S = -k \sum_{i=1}^{N} p_i \ln p_i$$

where N is the number of microstates, and p_i is the probability of microstate i. We won't try to prove this, but we can consider a couple of extreme cases. At one extreme, if we have N states with equal probabilities, the probability of each state must be $p_i = 1/N$. The entropy is then calculated as:

$$S = -k \sum_{i=1}^{N} p_i \ln p_i = -k \sum_{i=1}^{N} \frac{1}{N} \ln \frac{1}{N}$$

$$= -kN \frac{1}{N} \ln \frac{1}{N}$$

$$= k \ln N$$

consistent with the original expression for Ω microstates with equal probabilities. At the other extreme, if there is only a single state, its probability must be 1, and the entropy is:

$$S = -k \sum_{i=1}^{N} p_i \ln p_i$$

$$= -k1 \ln(1)$$

$$= 0$$

So, this state has an entropy of zero. In general, for a given number of microstates, if some of the microstates are more probable than others, the entropy will be lower than if all of the microstates have equal probabilities. So, for instance, if some configurations of a molecule are more probable because certain atoms tend to interact with each other, the molecule will have a lower entropy than if all of the possible conformations had equal probabilities.

IV. Entropy and information

There is another way of thinking about entropy that comes from a quite different discipline, information theory, which is concerned about how to efficiently and accurately transmit and manipulate information. This was a subject that was developed during

the middle decades of the 20th Century, largely at the late great Bell Laboratories, which did basic research for the major U.S. telephone company of the time (AT&T).

In this view, entropy can be described as the amount of information required to specify the exact state of a system. So, for instance, a crystal of a pure substance can be described with relatively little information, because each molecule is in an equivalent position of the crystal. If we know the structure of the molecule and the parameters describing the crystal lattice, then we can completely reconstruct the structure of the crystal. On the other hand, to completely describe a gas of the same molecules requires that we individually specify the position, orientation and momentum of each molecule.

This idea can be quantified, to give a function that calculates, for instance, the number of bits required to transmit a particular message or other information such as images or audio signals. Some messages contain a great deal of repetition and so can be encoded with relatively few bits, meaning that they have low entropy. A message composed of randomly-chosen letters has a high entropy and requires more bits to communicate. The equations for calculating information content have exactly the same form as the statistical definition of thermodynamic entropy, and information content is often described as information entropy.

The science of information theory has become very important in the digital age, as it defines the minimum amount of resources (including time) required to transmit a message or other information. The algorithms that are used to compress computer files, images and mobile-phone transmissions take advantage of the fact that the information is *not* random, *i.e.,* has a relatively low entropy, to represent it with less data.

There is another interesting aspect to the relationship between information and entropy. Reducing thermodynamic entropy requires work, or energy. Similarly, manipulating information and reducing its entropy (such as in compressing it) requires work. There are people working on the theory of quantifying exactly what is the minimum amount of energy required for specific computational tasks.

We don't always think about it, but the information technologies that we take advantage of use huge amounts of energy. Google is a major energy consumer. In 2009, the London Times published an article saying that a Google search produced about 1 g of CO_2, and this number is widely cited. This claim was quickly denied by Google, which a couple of years later released its own estimates of its energy consumption, indicating that a single search uses 0.3 watth (the energy required to power a 60-watt light bulb for about 20 s, or 3.600 J). If produced by burning coal, this would produce about 0.2 g of CO_2, far less than the original estimate. To its credit, Google has also invested heavily in renewable energy sources.

V. The second law

Now that we have defined this new state function, entropy (twice in fact), we can state the second law of thermodynamics. There are several equivalent ways to state this law, but the most common one is probably the following:

> For a spontaneous process, the total entropy of the system and its surroundings increases.

This leaves us with two questions, though:

- What do we mean by spontaneous?
- How do we define, measure or calculate the entropy change for the surroundings?

By spontaneous, we mean that the process will occur without any mechanical work being applied to the system. In other words, $w \leq 0$. If $w < 0$, then we can actually use the process to *do* work. A gas expansion is an example of a spontaneous process, and one that can do work. Compressing a gas, on the other hand, requires work.

For a specified process, the entropy change for the surroundings is defined as:

$$\Delta S_{\text{surr}} = -\frac{q}{T}$$

The important point here is that q is the heat absorbed by the system for the actual process, *not* the maximum-work reversible process. The entropy of the surroundings is *not* a function of the state of the system.

The entropy of the surroundings increases whenever heat flows from the system outwards. This only occurs when the system is warmer than the surroundings, and the flow of heat represents an increase in the disorder of the surroundings. If the system has a lower temperature than the surroundings, heat will flow inward, and the entropy of the surroundings will decrease. These are the only ways that the system can affect the entropy of the surroundings.

If there is a net flow of heat from the system to the surroundings, the entropy of the system can *decrease* in a spontaneous process.

Consider two cases for the expansion of an ideal gas:

1. Adiabatic expansion with no work. $\Delta E = 0$, $q = w = 0$.

   $$\Delta S_{\text{univ}} = \Delta S_{\text{sys}} + \Delta S_{\text{surr}} = nR \ln \left(\frac{V_2}{V_1} \right)$$

   So long as $V_2 > V_1$, $\Delta S_{\text{univ}} > 0$, and the process is spontaneous, which we knew already! But, now we have a way of quantifying the tendency of the gas to expand.

2. Reversible isothermal expansion with maximum work. $\Delta E = 0$, $q = q_{\text{rev}}$

   $$\begin{aligned} \Delta S_{\text{univ}} &= \Delta S_{\text{sys}} + \Delta S_{\text{surr}} \\ &= \frac{q_{\text{rev}}}{T} + \frac{-q_{\text{rev}}}{T} \\ &= 0 \end{aligned}$$

   So, the reversible process is right on the edge of being spontaneous, which we can take as the meaning of a reversible process.

The key thing to keep in mind about the second law is that it refers to the entropy of universe, not just the system. This means that the entropy of the system can decrease in a spontaneous process, as long as there is a flow of heat to the surroundings. Heat flow to the surroundings can only occur if the system is warmer than its surroundings, and the overall effect is to equilibrate the temperature of the universe, which represents an increase in entropy of the universe.

Although we can use the second law in this form, this approach becomes quite awkward when it is applied to things like chemical reactions, which is the major context in which thermodynamics is used in biology. In the next section, we consider the thermodynamics of chemical reactions and introduce a more practical way of applying the second law.

5.3 Thermodynamics of Chemical Reactions

I. E and ΔE reconsidered

Most generally, the change in internal energy for a process, ΔE, is defined by the first law of thermodynamics, stated as $\Delta E = q + w$, where q is the heat absorbed by the system and w is the work done on the system. Heat, in turn, is defined as the flow of energy that can increase temperature (but doesn't necessarily for a given process), and work is the integral of force with respect to distance. The internal energy of the system, E, is a state function, and ΔE is independent of the path taken from the beginning state to the ending state. The heat, q, and work, w, however, are path dependent. But, whatever the path ΔE must equal $q + w$.

Although these definitions stand on their own, it is helpful to have a molecular interpretation. When considering an ideal gas, the only form of energy present is the translational kinetic energy, and E can be calculated from the relationship $E_{\mathrm{k}} = 3kT/2$. Any change in internal energy, ΔE, can only be due only to a change in temperature.

The two distinguishable components of the internal energy are the kinetic energy and potential energy. In the simplest case, an ideal gas, there is only the kinetic energy associated with translational motion, and $E = 3nRT/2$, for n moles at temperature T. Since the particles making up the gas have no internal structure, there is no potential energy or internal thermal motions. If the temperature is the same at the beginning and end of the process, $\Delta E = 0$.

When we deal with real molecules, even in the gas phase, things get more complicated. There are now internal motions that contribute to the kinetic energy, and the internal structures of the molecules give rise to potential energy. The potential energy represents energy that can be converted to heat or mechanical energy if the structures of the molecules change, such as by forming or breaking chemical bonds. Potential energy can also be absorbed or released by changes that do not involve covalent bonds, such as the formation or breaking of hydrogen bonds or other "weak" interactions.

In liquids and solids, there are additional forms of potential energy, due to interactions among molecules.

For systems involving real molecules, the first law ($\Delta E = q + w$) still holds for any process, but there may be redistributions of kinetic and potential energy within the system that do not change the total internal energy, E. These redistributions will not be reflected in ΔE.

II. Enthalpy (H)

For practical problems involving chemical reactions, we usually apply the second law of thermodynamics using an approach developed by Josiah Willard Gibbs (1839–1903), who was arguably the first really great American physical or theoretical scientist. Gibbs, working mostly on his own, put together the ideas of thermodynamics into a consistent system. He spent his entire career at Yale College (as it was then called), and published all of his work in the Transactions of the Connecticut Academy of Sciences. His work was very slow to be fully appreciated, especially in his home country.

One of Gibbs' major accomplishments was formulating the application of thermodynamics to chemical reactions. In doing so, he introduced two new state functions, which we now call enthalpy (H) and the Gibbs free energy (G).

The formal definition of enthalpy is:

$$H = E + PV$$

where P and V are pressure and volume respectively. Like E, P and V are state functions, so H must be also. At first glance, however, determining H or ΔH doesn't look any easier than determining E or ΔE!

But, for chemical reactions, especially in relatively dilute solutions, we can often impose additional restrictions. First, and most important, we can specify that the the process occurs under constant pressure. Using the subscripts 1 and 2 to indicate the beginning and ending states, respectively, the change in enthalpy can then be written:

$$\begin{aligned}\Delta H &= H_2 - H_1 = E_2 - E_1 + PV_2 - PV_1 \\ &= \Delta E + P\Delta V\end{aligned}$$

If ΔV is not zero, the volume change represents work, as we discussed in the case of gas expansion or compression. For most chemical reactions, this is the only form of work associated with the reaction.[2] If the volume of the system increases, then work is done *by* the system, and w is negative. If this form of work is designated w_p(the subscript "p" indicating constant pressure), then we can write:

$$w_\mathrm{p} = -P\Delta V$$

Assuming that this is the only form of work associated with the process at constant pressure, and designating the associated heat absorbed q_p, we have:

$$\begin{aligned}\Delta H &= \Delta E + P\Delta V = q_\mathrm{p} + w_\mathrm{p} - w_\mathrm{p} \\ &= q_\mathrm{p}\end{aligned}$$

[2]Notable exceptions are found in the molecular motors in living organisms, as well as some synthetic systems, but here we are focusing on relatively simple chemical reactions.

This gives the usual working definition for ΔH: The heat absorbed by the system at constant pressure. Although q is not, in general a state function, because q_{p} refers to a specific path between two states, it is a state function. In this respect, the use of q_{p} to define ΔH is analogous to the use of q_{rev} to define ΔS. This makes it relatively easy to measure ΔH for a reaction, using a calorimeter.

With the above restrictions, we can also express ΔH in terms of ΔE:

$$\Delta H = \Delta E - w_{\mathrm{p}}$$

III. ΔG, the change in Gibbs free energy.

The free energy function named for Gibbs, G, is defined as:

$$G = H - TS_{\mathrm{sys}}$$

and the change in G *at constant temperature* is:

$$\Delta G = \Delta H - T\Delta S_{\mathrm{sys}}$$

As noted earlier, the term $T\Delta S_{\mathrm{sys}}$ represents the heat absorbed during the reversible path. As argued above, ΔH is the heat absorbed by the process at constant pressure, and with no work other than w_{p}. Recall that the entropy change for the surroundings for a specific process is defined as:

$$\Delta S_{\mathrm{surr}} = -\frac{q}{T}$$

For the constant pressure process (again assuming constant temperature):

$$\Delta S_{\mathrm{surr}} = -\frac{q_{\mathrm{p}}}{T} = -\frac{\Delta H}{T}$$

and, for any constant temperature process:

$$\Delta S_{\mathrm{sys}} = \frac{q_{\mathrm{rev}}}{T}$$

If we then restrict ourselves to processes where both pressure and temperature are constant, the total entropy change for the universe is:

$$\begin{aligned}\Delta S_{\mathrm{univ}} = \Delta S_{\mathrm{surr}} + \Delta S_{\mathrm{sys}} &= -\frac{\Delta H}{T} - \Delta S_{\mathrm{sys}} \\ &= -\frac{\Delta G}{T}\end{aligned}$$

This result demonstrates that the sign of ΔG indicates whether or not the process is spontaneous: If ΔG is negative, ΔS_{univ} is positive and the process is favorable. If ΔG is positive, ΔS_{univ} is negative and the process is unfavorable. Because ΔG is a state function, as are ΔH, T and ΔS, we can determine whether or not a process is spontaneous without consideration of the surroundings, given the constraint of constant pressure. (Here we are assuming that temperature is also constant, but this assumption is not necessary in order to use the Gibbs free energy to determine if a process is spontaneous; The math just gets more complicated if temperature changes.)

IV. Free energy and available energy for work.

Earlier (page 141), another free energy function, the Helmholz free energy, was introduced and identified with the work associated by a process carried out along a reversible pathway:

$$\Delta F = w_{\text{rev}}$$

Since the reversible process yields the maximum amount of work obtainable from a transition between two states, $-\Delta F$ is the maximum work available from a favorable process. For an unfavorable process, ΔF is the minimum amount of work required to drive the process.

Though it's not very obvious from the relationship above, the Helmoltz free energy, F, is closely related to the Gibbs free energy, G. From the first law, we can write:

$$\Delta E = q_{\text{rev}} + w_{\text{rev}}$$
$$w_{\text{rev}} = \Delta E - q_{\text{rev}}$$

Therefore:

$$\Delta F = \Delta E - q_{\text{rev}}$$

Recall that ΔS, for a process at constant temperature, is q_{rev}/T. Substituting, we have:

$$\Delta F = \Delta E - T\Delta S$$

The formal definition of the Helmholtz free energy is:

$$F = E - TS$$

For comparison, the definition of the Gibbs free energy is:

$$G = H - TS$$

and the change in Gibbs free energy at constant temperature is:

$$\Delta G = \Delta H - T\Delta S$$

Thus, the difference between F and G is the difference between E and H. From before, the definition of H is:

$$H = E + PV$$

where P and V are the pressure and volume, respectively. For a process at constant pressure:

$$\Delta H = \Delta E + P\Delta V$$

Therefore, ΔG and ΔF are related according to:

$$\Delta G = \Delta F + P\Delta V$$

The term $P\Delta V$ represents the work due to any change in volume, and can be either positive or negative. Since $-\Delta F$ is the maximum amount of work available from a process, $-\Delta G$ is that amount of work minus any work due to a volume change.

If the the volume change is negligible, then $\Delta H = \Delta E$, and $\Delta G = \Delta F$. With this provision, ΔG can be interpreted as the minimum work required to drive an unfavorable process (for $\Delta G > 0$) or the negative of the maximum work that can be obtained from a favorable process ($\Delta G < 0$). This condition is often satisfied for chemical reactions in dilute solution and is frequently assumed to be valid for such reactions.

V. Free energy changes for chemical reactions

Most of the transformations of energy in biology involve chemical reactions. When we considered gas expansion problems, one of our underlying assumptions was that the internal energy of the system was only a function of temperature. But, when chemical reactions are involved, there is usually a change in the internal energy of the molecules. Forming and breaking chemical bonds almost always involves a change in potential energy. Bonds form because the atoms involved find a lower energy state in which they "share" electrons. The changes in internal potential energy must be accompanied by either changes in the thermal (kinetic) energy of the system, heat flow to or from the surroundings or mechanical work, or some combination of the above.

In biological systems, the energy changes associated with reactions are frequently coupled to other processes. In that way, the energy change for a favorable process can be used to drive an unfavorable one. The most commonly used tool for keeping track of, and understanding, these energy conversions is the Gibbs free energy.

For the simple case of the interconversion of one compound to another, we can write a generic chemical reaction as:

$$\mathrm{A} \rightleftharpoons \mathrm{B}$$

In principle, any chemical reaction is reversible, in that there is a finite probability that it can occur in either direction. The probability that any molecule of A will be converted to B is the same as for any other molecule of A. As a consequence, the total rate of conversion of A to B will be proportional to the number of A molecules, or equivalently, their concentration. Similarly, the rate of conversion of B to A will be proportional to the concentration of molecules of B. The net rate in the change in the concentration of A will be the difference between these two rates, which we can write as a differential equation:

$$\frac{d[\mathrm{A}]}{dt} = -k_\mathrm{f}[\mathrm{A}] + k_\mathrm{r}[\mathrm{B}]$$

where k_f and k_r are rate constants that relate the probability of each reaction to the concentration of the molecules. The net rate of conversion from B to A is the negative

of the rate of conversion from A to B, which we can write as:

$$\frac{d[\mathrm{B}]}{dt} = -\frac{d[\mathrm{A}]}{dt} = k_\mathrm{f}[\mathrm{A}] - k_\mathrm{r}[\mathrm{B}]$$

We won't worry about solving these equation, except to note that if we start out with all A, some of it will be converted to B. As the concentration of A decreases, the rate of conversion to B will decrease, and the rate of conversion of B to A will increase. At some point, the flow in each direction will be equal, and the rate of change in concentration will be zero. At this point:

$$\frac{d[\mathrm{A}]}{dt} = 0$$

$$k_\mathrm{f}[\mathrm{A}] = k_\mathrm{r}[\mathrm{B}]$$

$$\frac{[\mathrm{B}]}{[\mathrm{A}]} = \frac{k_\mathrm{f}}{k_\mathrm{r}}$$

At this point, we say that the reaction is at equilibrium, and we can define an equilibrium constant:

$$K_\mathrm{eq} = \frac{k_\mathrm{f}}{k_\mathrm{r}} = \frac{[\mathrm{B}]_\mathrm{eq}}{[\mathrm{A}]_\mathrm{eq}}$$

where the subscripts indicate that the concentrations are those that are measured at equilibrium. This is not the most rigorous derivation of an equilibrium constant, but it will do.

One of the implications of the equilibrium state is that there is no way in which to obtain work from the reaction. If the system is at equilibrium, then an infinitesimal amount of work could shift it in either direction, by an infinitesimal amount. Looked at another way, if we were to have an immense volume of the reactants at the equilibrium concentrations, one mole of A could be converted to one mole of B (or vise versa) without doing any work, provided that the concentrations didn't change significantly. Therefore, the free energy change for the reaction must be zero.

Suppose, though, that we were to change the concentrations from their equilibrium values, say by adding one or the other of the reactants. Then, we would expect the concentrations to shift towards those that would satisfy the equilibrium constant, and during this process we could, at least in principle, extract some work. At concentrations that do not satisfy the equilibrium conditions, ΔG is negative if the concentrations favor the forward reaction, and ΔG is positive if the concentrations favor the reverse reactions.

For any set of reactant and product concentrations, ΔG is defined as the free energy for converting one mole of reactants to one mole of products, A to B in the example used here, in a volume large enough that the concentrations do not change significantly.

We can derive an expression to calculate ΔG for an arbitrary set of concentrations by considering the sum of three processes:

1. $\mathrm{A}([\mathrm{A}]_{\mathrm{eq}}) \longrightarrow \mathrm{B}([\mathrm{B}]_{\mathrm{eq}})$
2. $\mathrm{A}([\mathrm{A}]) \longrightarrow \mathrm{A}([\mathrm{A}]_{\mathrm{eq}})$
3. $\mathrm{B}([\mathrm{B}]_{\mathrm{eq}}) \longrightarrow \mathrm{B}([\mathrm{B}])$

where the concentrations of A and B in the three processes are indicated in the parentheses: $[\mathrm{A}]_{\mathrm{eq}}$ and $[\mathrm{B}]_{\mathrm{eq}}$ represent the equilibrium concentrations and [A] and [B] are the concentrations at which ΔG is to be calculated. The sum of the three reactions is:

4. $\mathrm{A}([\mathrm{A}]) \longrightarrow \mathrm{B}([\mathrm{B}])$

A general principle of chemical thermodynamics states that the changes in state functions (including ΔE, ΔH, ΔG and ΔS) associated with individual reactions can be added to yield the change in state functions for the corresponding sum of the reactions. Thus, we can write:

$$\Delta G_4 = \Delta G_1 + \Delta G_2 + \Delta G_3$$

Since process 1 represents the equilibrium condition, $\Delta G_1 = 0$. Process 2 represents a change in the concentration of A, from the concentration of interest, [A], to the equilibrium concentration, $[\mathrm{A}]_{\mathrm{eq}}$. If we assume the results for ideal gasses can be extended to the reaction in dilute solution, the reversible work for the change in concentration, w_{rev}, is calculated as:

$$w_{\mathrm{rev}} = RT \ln \frac{[\mathrm{A}]_{\mathrm{eq}}}{[\mathrm{A}]}$$

If we further assume that the volume change at constant pressure is negligible, then $w_{\mathrm{rev}} = \Delta G$, and we can write the free energy change for process 2 as:

$$\Delta G_2 = RT \ln \frac{[\mathrm{A}]_{\mathrm{eq}}}{[\mathrm{A}]}$$

From analogous arguments, the free energy change for process 3 is:

$$\Delta G_3 = RT \ln \frac{[\mathrm{B}]}{[\mathrm{B}]_{\mathrm{eq}}}$$

The free energy change for process 4, the reaction at concentrations [A] and [B] is then

given by:

$$\begin{aligned}\Delta G_4 &= \Delta G_1 + \Delta G_2 + \Delta G_3 \\ &= 0 + RT \ln \frac{[\mathrm{A}]_{\mathrm{eq}}}{[\mathrm{A}]} + RT \ln \frac{[\mathrm{B}]}{[\mathrm{B}]_{\mathrm{eq}}} \\ &= RT \ln \frac{[\mathrm{B}][\mathrm{A}]_{\mathrm{eq}}}{[\mathrm{A}][\mathrm{B}]_{\mathrm{eq}}} \\ &= RT \ln \frac{[\mathrm{B}]}{[\mathrm{A}]} - RT \ln \frac{[\mathrm{B}]_{\mathrm{eq}}}{[\mathrm{A}]_{\mathrm{eq}}} \\ &= RT \ln \frac{[\mathrm{B}]}{[\mathrm{A}]} - RT \ln K_{\mathrm{eq}}\end{aligned}$$

In order to define and calculate the free energy change for a reaction at different concentrations, it is useful to define a standard state for measuring and reporting free energy changes, and the most widely used standard is to specify that all of the reactants and products are at 1 M concentration. (Or, for the gas phase, that the pressures of each component is 1 atm.) The free energy change under these conditions is defined as the standard free energy change, ΔG°. From the equation above, the standard free energy change is calculated as:

$$\begin{aligned}\Delta G^\circ &= RT \ln \frac{1\,\mathrm{M}}{1\,\mathrm{M}} - RT \ln K_{\mathrm{eq}} \\ &= -RT \ln K_{\mathrm{eq}}\end{aligned}$$

This then provides a link between the standard free energy change and the equilibrium constant. Both represent essentially the same thing: the extent to which one side of the reaction is favored. If the forward reaction (A $\longrightarrow$ B) is favored, then $K_{\mathrm{eq}} > 1$ and $\Delta G^\circ < 0$. A negative value of ΔG° also implies that work can be obtained from the reaction if the reactants are initially at their standard-state conditions.

Measuring the equilibrium constant is generally the most straight forward way of measuring ΔG°. Once this value is known, the free energy change for other concentrations can be easily calculated. If ΔH is also know, for instance by measurement with a calorimeter, the entropy change, ΔS, can then be calculated from the relationship $\Delta G = \Delta H - T\Delta S_{\mathrm{sys}}$.

The commonly used form of the equation for calculating the free energy change at concentrations other than the standard state ones is:

$$\Delta G = \Delta G^\circ + RT \ln \frac{[\mathrm{B}]}{[\mathrm{A}]}$$

Especially for reactions that involve more than one reactant or product, it is convenient to introduce a quantity called the *reaction quotient*, Q. This quantity is defined as the

product of the reaction product concentrations divided by the product of reactant concentrations. For the reaction:

$$\mathrm{aA} + \mathrm{bB} \rightleftharpoons \mathrm{cC} + \mathrm{dD}$$

The reaction quotient is:

$$Q = \frac{[\mathrm{C}]^c[\mathrm{D}]^d}{[\mathrm{A}]^a[\mathrm{B}]^b}$$

When the reaction is at equilibrium, Q is equal to the equilibrium constant, K_{eq}, and the free energy associated with the reaction is zero. At other concentrations, the free energy change is calculated as:

$$\Delta G = \Delta G^\circ + RT \ln Q$$

VI. Concentrations and standard states

Things become a little more complicated when the number of reactant and product molecules are not equal. For instance, consider the reaction:

$$\mathrm{A} + \mathrm{B} \rightleftharpoons \mathrm{C}$$

We can write an equilibrium constant:

$$K_{\mathrm{eq}} = \frac{[\mathrm{C}]_{\mathrm{eq}}}{[\mathrm{A}]_{\mathrm{eq}}[\mathrm{B}]_{\mathrm{eq}}}$$

And the expression for the free energy change is:

$$\Delta G = \Delta G^\circ + RT \ln \frac{[\mathrm{C}]}{[\mathrm{A}][\mathrm{B}]}$$

But, there are some funny things here:

- The equilibrium constant has dimensions of inverse concentration. If different units are used for concentration, the value of the equilibrium constant will change.
- The reaction quotient (Q) also has units of inverse concentration. How can we take the logarithm of a quantity with units?
- The value of ΔG will also depend on the concentration units chosen.

We can solve the problem of the logarithms by stipulating that the reaction quotient should be written in terms of the ratios of the actual concentrations and the standard state concentrations. If the standard state concentrations are 1 M, the the reaction quotient is written as:

$$Q = \frac{[\mathrm{C}]/1\,\mathrm{M}}{([\mathrm{A}]/1\,\mathrm{M})([\mathrm{B}]/1\,\mathrm{M})}$$

Now, Q is dimensionless, and it is OK to take a logarithm. But, it is still true that K_{eq}, ΔG and ΔG° will all depend on the units of concentration chosen.

The underlying reason for this is that when the number of molecules changes in a reaction there is an intrinsic change in entropy. In the example above, there will be a reduction of entropy when two molecules are converted to one, and this will disfavor the reaction. If the standard state is, for instance, 1 mM instead of 1 M, then the loss of entropy will be larger, and this will be reflected in the numerical value of the equilibrium constant and the free energy change.

VII. Calculating the entropy change for a bimolecular reaction.

Following the approach used when considering the expansion of a gas (pages 144–145), we can use the statistical definition of entropy to estimate the entropy loss due to converting two molecules into one.

As in the example of gas expansion, we begin by dividing up the volume of interest into small cubes with volume V_c. The total volume is $N_\text{c}V_\text{c}$, where N_c is the number of small cubes.

Considering just one molecule each of A and B, the number of ways of placing these two molecules in the lattice of cubes, $\Omega_{\text{A,B}}$ is N_c^2. This again assumes that the number of cubes is much larger than the total number of molecules. Because, the two molecules are assumed to be different, N_c^2 is not divided by two, since swapping the positions of A and B results in a distinct arrangement. If the two reacting molecules were identical, then $\Omega_{\text{A,B}}$ would equal $N_\text{c}^2/2$

The number of arrangements of a single molecule of C in the volume is $\Omega_\text{C} = N_\text{c}$.

The entropy change for a single pair of A and B molecules being converted to C is then given by:

$$\Delta S = k \ln \left(\frac{\Omega_\text{C}}{\Omega_{\text{A,B}}} \right) = k \ln \left(\frac{N_\text{c}}{N_\text{c}^2} \right)$$

$$= -k \ln N_\text{c}$$

This result indicates that the entropy decreases for the reaction, as we expect. But, the result is also somewhat problematic, because the number of cubes, N_c, does not cancel out, as it did for the example of a gas expansion. Therefore, the size of the cubes does matter in this case.

Though there is no absolutely certain way to define the size of the cubes, we can at least make a reasonable estimate. The cubes should be just big enough to hold one of the reactant or product molecules. The product is probably larger than either of the reactants, making it difficult to specify a single size, but a reasonable estimate would be a cube 1 nm on each side, or a volume of 1 nm^3. The number of cubes is then calculated by dividing the reaction volume, V, by the volume of a single cube, V_c. We

will specify that the total volume is given in liters. Taking into account the necessary conversion factors, N_c can be expressed as:

$$N_\mathrm{c} = \frac{V(\mathrm{L})}{1\,\mathrm{nm}^3} \times \frac{10^3\,\mathrm{m}^3}{L} \times \left(\frac{10^9\,\mathrm{nm}}{\mathrm{m}}\right)^3 \approx V(\mathrm{L}) \times 10^{24}$$

The entropy change is then:

$$\Delta S = -k \ln(N_\mathrm{c}) = -k \ln(V(\mathrm{L}) \times 10^{24})$$

Note that the entropy change is closely related to the reaction volume, V, and becomes more negative with larger volume. This can be understood by recognizing that all of the molecules gain entropy when the volume is increased, but the two reactant molecules, together, lose more than the single product molecule.

For one mole each of A and B converted to C, the result above is multiplied by Avogadro's number:

$$\begin{aligned}\Delta S &= -N_\mathrm{A} k \ln(V(\mathrm{L}) \times 10^{24}) \\ &= -R \ln(V(\mathrm{L}) \times 10^{24})\end{aligned}$$

If the volume is specified as 1 L, a numerical result can be calculated:

$$\begin{aligned}\Delta S &= -R \ln(10^{24}) = -8.31\,\mathrm{J/K} \times 53 \\ &= -460\,\mathrm{J/K}\end{aligned}$$

The contribution to the free energy change at 298 K is:

$$-T\Delta S = -460\,\mathrm{J/K} \times 440\,\mathrm{J/K} = 140\,\mathrm{kJ}$$

In order for the reaction to be favorable, this large contribution would have to be compensated for by other factors, such a favorable enthalpy change due to covalent changes in the molecule or other entropic factors.

This simple calculation shows that the change of entropy associated with a reaction in which the number of molecules changes can be very significant. It is particularly significant in the context of biological systems, where very large molecules, such as proteins, nucleic acids and polysaccharides, are assembled from small building blocks (amino acids, nucleotides and sugars). However, accurately calculating these entropy changes is challenging and a source of some controversy. The calculation above considers only the loss of translational degrees of freedom, and a large reduction in rotational freedom is also to be expected in a bimolecular reaction. Other factors, including the changes in internal motions (bond rotations and vibrations) and interactions with solvent molecules may also contribute to the total entropy change. Among various estimates, the one above represents the lower range of calculated entropy changes for the association of two molecules.

VIII. Activity versus concentration

The expression relating the free energy and concentrations is strictly valid only for ideal gas molecules. If there is any tendency for the molecules to interact with one another or with solvent molecules, this can affect the free energy change. The formal way of dealing with this problem is to replace the concentrations with "activities" for each of the reactants and products:

$$\Delta G = \Delta G^\circ + RT \ln \frac{a_A a_B}{a_C a_D}$$

The activities are dimensionless quantities, but they are defined relative to a specific standard state (*e.g.,* 1 M). At relatively low concentrations, the activities are often close to being proportional to concentration and are often expressed in terms of activity coefficients, γ:

$$a_A = \gamma_A [\mathrm{A}]$$

The activity coefficient can depend on concentration. If $\gamma = 1/(1\,\mathrm{M})$, then the species behaves ideally, and the activity is equal to the concentration divided by the standard state concentration.

For most biochemical reactions, at relatively low concentrations, the deviations from ideal behavior are not very large (or at least we hope that they aren't), and we rarely worry about them. As a practical matter, we usually assume that we can use concentrations in the equilibrium expression.

5.4 "Chemical Energy" and Metabolism

It is common to hear people talk about "chemical energy" or refer to certain compounds as having "high energy" or "high energy bonds". We can even eat "energy bars" and drink "energy drinks". This language can be rather vague, or even misleading. Here we consider some specific examples of biochemical processes and the meaning of chemical energy in this context.

I. Glucose oxidation

The most useful measure of chemical energy is the free energy change. For instance, the oxidation of glucose by molecular oxygen is written as:

$$\mathrm{C_6H_{12}O_6} + 6\,\mathrm{O_2} \rightleftharpoons 6\,\mathrm{CO_2} + 6\,\mathrm{H_2O}$$

Like any reaction, this reaction is reversible and has an equilibrium constant. The extent to which it is favorable depends on the concentrations of all of the reactants and products. The standard free energy change for this reaction is about $-2.700\,\mathrm{kJ/mol}$, meaning that it is an extremely favorable reaction when all of the reactants are at their standard states (1 M or 1 atm). But, if the O_2 concentration is very low, the reaction is much less favorable. Prior to about 2.5 billion years ago, the concentration

of oxygen in earth's atmosphere was about 10^5-fold less than it is now. Under those conditions, the oxidation of glucose was not favorable at all. Glucose was used as a source of "energy" through glycolysis (a partial enzymatic breakdown of glucose), and this is still an important metabolic reaction. But glycolysis is a much less favorable overall reaction than complete oxidation (at the present atmospheric concentrations of O_2 and CO_2), meaning that glucose or other carbohydrates can provide much less "energy" under anaerobic conditions.

The nutritional calories that are listed for foods are *not* free energy changes. They are measured by burning the foods (or their digestible ingredients) and directly measuring the heat in a calorimeter at constant volume, with an excess of oxygen. So, these calories actually represent ΔE. For glucose, the dietary value is about 4 kcal/g, which corresponds to about 175 kJ/mol, much less than the standard free energy change. This is because there is a large entropy change in the reaction as well, as 7 molecules are converted to 12.

The nutritionists then estimate how much work must be done by the body in order to offset the metabolism of, for instance, 1 g of sugar. Most of the free energy is lost as heat. Relating food calories to free energy changes is not straight forward (to me at least!)

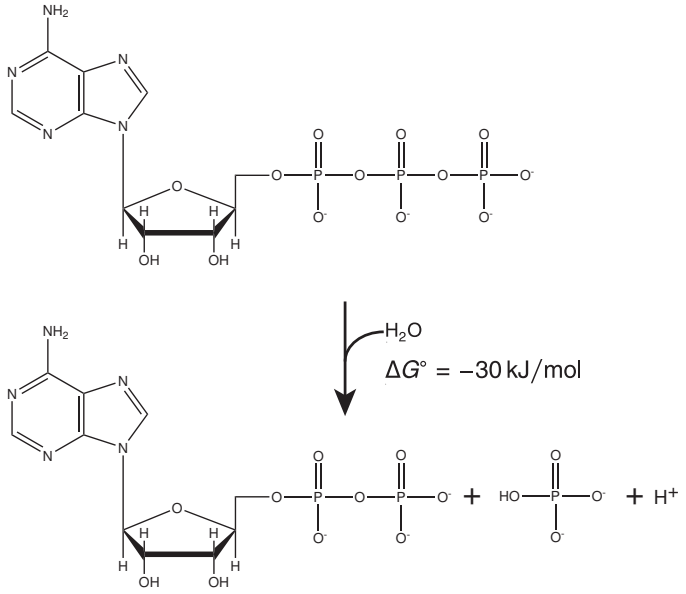
The important point is that in our atmosphere, the oxidation of glucose and related molecules is very favorable, and organisms can couple this very favorable reaction to processes that otherwise would not be favorable.

The synthesis of glucose and related molecules is *unfavorable* and depends on coupling to other favorable reactions, which are driven by absorbing the energy of light from the sun.

For our cars and some electric power plants, we (mostly) use as a fuel hydrocarbons or coal, which also are oxidized with a large negative free energy change. All of the energy in hydrocarbons and coal, was originally captured by photosynthesis. Plants and animals that ate plants died and the carbon was converted into the forms we now extract and burn. We are quickly consuming the vast amounts that were accumulated over billions of years, and we are moving the carbon into the atmosphere.

II. ATP hydrolysis

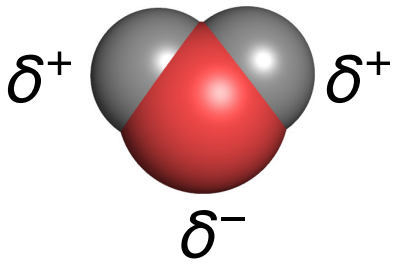
Another compound that is often described as having "high energy" is ATP, with the hydrolysis reaction shown below:



The bond linking the second and third phosphate groups is often described as a "high energy bond". In fact, there is nothing very special about this bond. What is important is that the reaction has a large negative free energy change under physiological conditions, about $-30\,\text{kJ/mol}$.

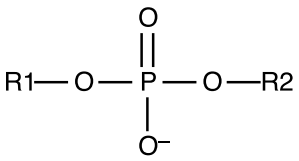
We won't worry for now how this compound gets made in the first place. Basically, ATP serves as a kind of energy currency, it is formed during the oxidation of glucose (and other favorable reactions) and it is hydrolyzed to provide energy for other processes.

In considering why this reaction is so favorable, the first thing to emphasize about this reaction is that it occurs in the presence of water, and water is one of the reactants. We will talk more about water later, but for now, the important thing is that water is a polar molecule, meaning that the electrons of the molecule are unevenly distributed leading to partial charges on the hydrogen and oxygen atoms:



The partial charges of the water molecule can interact with charges on other molecules making the charged forms much more stable than they would be otherwise.

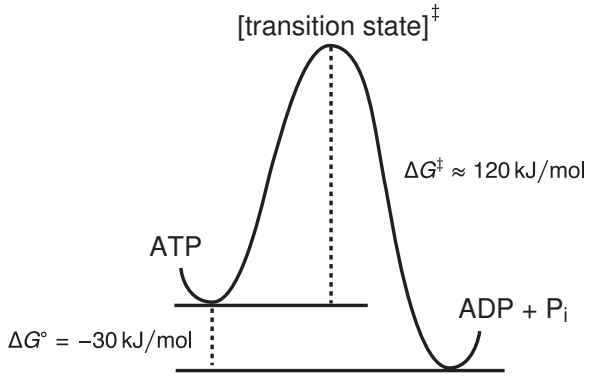
ATP, as suggested by its name, has three phosphate groups:



In water, at neutral pH, one or more of the oxygen atoms will be in an ionized, negatively charged, state. In ATP, there are four negative charges on the three phosphates (one each on the first and second and two on the third phosphate). This results in a quite high density of negative charge. When the third phosphate is removed by hydrolysis, the charges are not so close together and are shielded by the water molecules, leading to a reduction in potential energy. There are other factors, but this is the major one. Further hydrolysis, of the remaining phosphate-phosphate ester bond, is also favorable, but not by quite so much.

Even if a reaction is thermodynamically favored, it may not occur very rapidly. This is obviously true for oxidation of glucose, and it is also true for ATP hydrolysis, for which the half time is about 20 days at neutral pH and 60 °C.

It's useful to represent the relationship between thermodynamics and kinetics as a reaction coordinate, or energy profile:



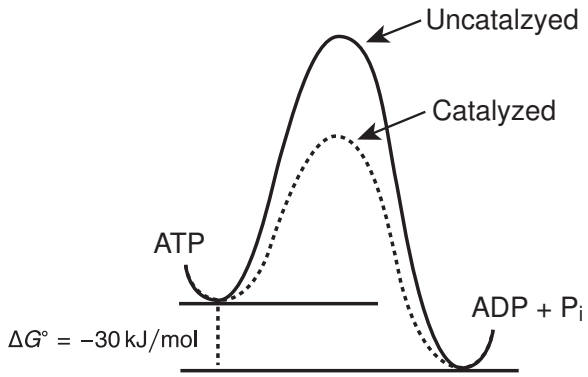
The central idea here is that the molecules have to acquire a high energy state in order for the reaction to proceed. This high energy state is called a "transition state" or "activation complex" and is a bit of a mythical beast. It is imagined to be in equilibrium with the products and reactants, but present at only very low concentrations. As soon as it forms, it breaks down to either products or reactants, with equal probability. The higher the transition-state energy, the slower the reaction.

The theory for relating reaction rates to energies was developed by Henry Eyring. See the painting in the lobby of building named for him!

In the absence of a catalyst, the reaction rate under physiological conditions would be insignificant. This is important: If ATP spontaneously hydrolyzed, it wouldn't be an effective way of storing energy. It would be like trying to use a fuel that spontaneously combusts.

The key to using chemical energy in biology is controlling the reaction by catalysts, enzymes. Enzyme can increase the rate of a reaction by several orders of magnitude.

A somewhat glib way of saying what enzymes, or other catalysts, do, is to say that they lower the transition state energy[3]:

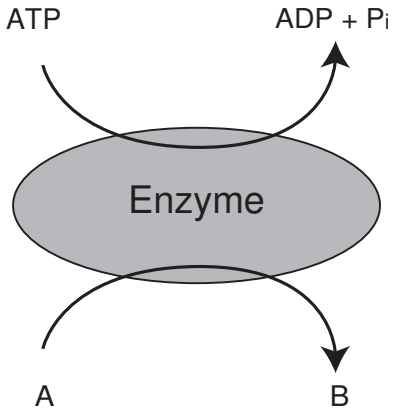


How do enzymes increase reaction rates? This is a major sub-discipline of biochemistry, and there are many different effects, not all of which are fully understood or agreed on. But, the basic idea is that the enzyme, which is a protein (or sometimes RNA) molecule, binds to the reactants and creates a local chemical environment that makes the reaction much more likely to proceed. One important part of this is that the local concentrations are much higher than they are when the molecules are free in solution. The enzyme can also help stabilize charges that form in the transition state.

An enzyme cannot alter the thermodynamics of a reaction! If it catalyzes the forward reaction, it must also catalyze the reverse reaction by the same factor. Any scheme that suggests that both reactions are not catalyzed equally violates the first law!

III. Enzymatic coupling

An enzyme that simply catalyzed ATP hydrolysis would not be very useful. Essentially all enzymes that catalyze ATP hydrolysis do so in a way that couples the favorable reaction with another reaction that would otherwise be unfavorable. The structure and mechanism of the enzyme is such that the hydrolysis reaction can only occur if the other reaction also occurs.



[3]The statement is glib because it doesn't really provide any additional information about the mechanism of the catalyst. But, considering the transition state and what determines its free energy is a valuable way of framing mechanistic questions

The enzyme links the two reactions together so that one can't occur without the other, so that the overall reaction is:

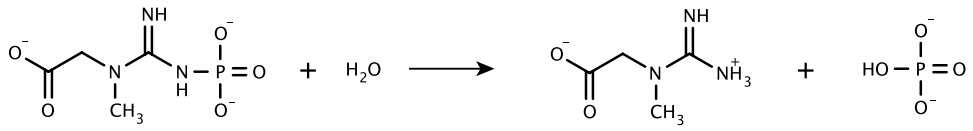
$$\mathrm{A} + \mathrm{ATP} \rightleftharpoons \mathrm{B} + \mathrm{ADP} + \mathrm{P_i}$$

The free energy changes for the individual reactions are added together. Even if ΔG for the second reaction is positive, the overall reaction can be favorable.

The coupled process can be a chemical reaction, but it can also be a physical process such as movements of molecules across a membrane against a concentration gradient or the generation of mechanical force, as in muscle.

The details of how this coupling occurs depends on the structures of the proteins that catalyze the reactions.

The enzyme creatine kinase offers a good example of enzymatic coupling. Like ATP, creatine phosphate (also called phosphocreatine) has a large negative free energy of hydrolysis



$$\Delta G^\circ = -43\,\mathrm{kJ/mol}$$

The enzyme creatine kinase catalyzes the exchange of phosphate between ATP/ADP and creatine phosphate/creatine (CP/C):

$$\mathrm{ADP} + \mathrm{CP} \rightleftharpoons \mathrm{ATP} + \mathrm{C}$$

The standard free energy change for the overall reaction is: $-43\,\mathrm{kJ/mol} + 30\,\mathrm{kJ/ml} = -13\,\mathrm{kJ/mol}$

Creatine, creatine phosphate and the enzyme creatine kinase are found in a variety of animal tissues, including muscle and brain cells, where creatine phosphate serves as a short-term source of reserve chemical energy.

In resting muscle cells, the typical concentrations are:

4 mM ATP
0.013 mM ADP
25 mM creatine phosphate
13 mM creatine

We can calculate the free energy change for the formation of ATP from CP under these

conditions:

$$\begin{aligned}\Delta G &= \Delta G^\circ + RT \ln \frac{[\text{ATP}][\text{C}]}{[\text{ADP}][\text{C}]} \\ &= -13\,\text{kJ/mol} + RT \ln \frac{4\,\text{mM} \cdot 13\,\text{mM}}{0.013\,\text{mM} \cdot 25\,\text{mM}} \\ &= -13\,\text{kJ/mol} + 8.314\,\text{J/K} \cdot 310\,\text{K} \ln(160) \\ &= -13\,\text{kJ/mol} + 13\,\text{kJ/mol} \\ &\approx 0\end{aligned}$$

The four compounds are at equilibrium, because the enzyme quickly equilibrates them. But, when there is a large demand for ATP, the concentration of ATP goes down, the concentration of ADP goes up, and the forward reaction, as written above, becomes favorable, to restore ATP concentration.

But, the reserve of creatine phosphate is relatively limited and lasts about 4 seconds in a sprinter. After that, either ATP is restored by oxidative phosphorylation, or glucose is metabolized by glycolysis. The latter provides much less ATP than oxidative metabolism and leads to the accumulation of lactic acid.

Chapter 6

Formation of Biomolecular Structures

Now, we are going to shift direction a bit and start discussing the question of how biological structures form on the molecular and cellular level. The thermodynamic principles that we have been discussing create a framework for this subject. At a first glance, it seems that biology somehow violates or circumvents the second law of thermodynamics, since highly ordered structures seem to form spontaneously. Our goal is to understand how that can happen within the constraints of the thermodynamic laws.

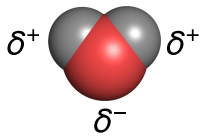
6.1 Water, Ionization and the Hydrophobic Effect

A key factor in biological assemblies is water and its special properties. We tend to take water for granted, since it is the liquid that we know the best, but it is actually a very special liquid.

The unique properties of water become obvious when we try to mix it with other liquids, especially oils. Everyone knows about this experiment: oil and water don't mix. This is, in fact, the major driving force for the formation of biological structures at the molecular level. But, *why* don't they mix? What is so fundamentally different about the two kinds of liquids?

I. Hydrogen bonding

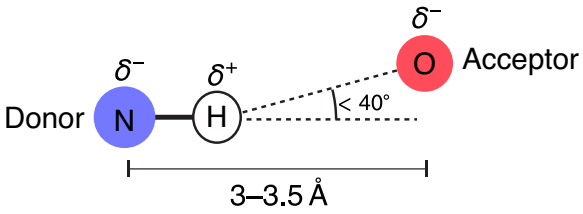
The key property of water is something that we briefly discussed earlier, the unequal sharing of electrons in the chemical bonds between oxygen and hydrogen:



This results in a partial positive charge on the hydrogen atoms and a partial negative charge on the oxygen atoms, which lead to a strong tendency of the molecules to interact. This is an example of a more general phenomenon, the hydrogen bond.

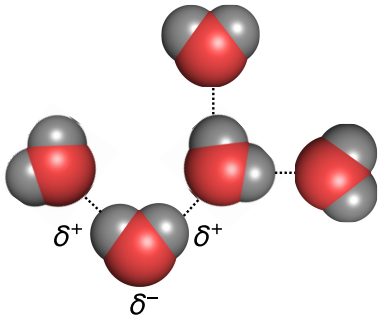
Hydrogen bonds form when a hydrogen atom is covalently bound to an electronegative atom (most often oxygen or nitrogen) and a second electronegative atom is in the vicinity. To a first approximation this is an interaction between charges, but there is also a small degree of covalent bonding, that is electron sharing, involved.

To form a hydrogen bond, the two molecules (or groups within a single molecule) must be arranged so that the two electronegative atoms are about 3–3.5 Å apart, as shown below:



The electronegative atom that is covalently bound to the hydrogen atom is referred to as the hydrogen-bond donor and the other electronegative atom is called the acceptor. In order to form a stable interaction, the three atoms must be approximately collinear with the angle indicated in the drawing no more than about 40°.

The oxygen atom of a water molecule can accept two hydrogen bonds and can act as a donor for two, thus enabling a water molecule to form up to 4 hydrogen bonds.

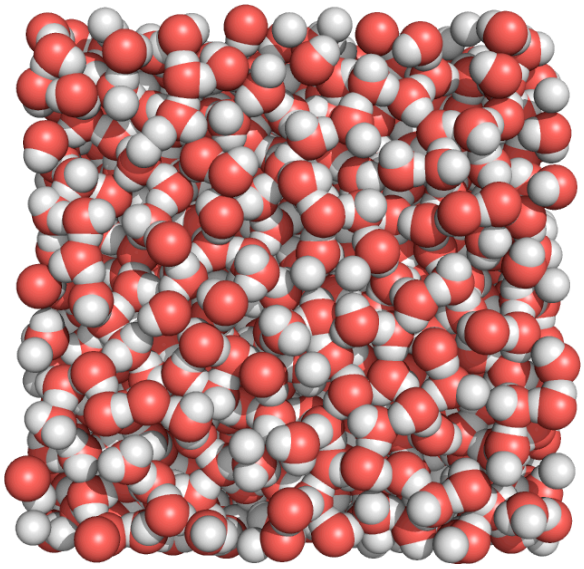


When water freezes it forms a lattice in which each molecule forms all four of the possible hydrogen bonds. The geometry is very similar to the lattice in a diamond, except that the bonds are much weaker.

In liquid water, each molecule forms, on average, three hydrogen bonds at any instant. Thus, only 1/4 of the hydrogen bonds break when ice melts. This is a major reason that the boiling temperature of liquid water is relatively high for a molecule of its size. (In general, the boiling points of liquids increase with the size of the molecules because they can form more extensive van der Waals interactions.)

The hydrogen bonds in liquid water rapidly break and reform constantly. At any instant, there is an extensive network of hydrogen bonds that can be traced from one side of a beaker to another (in principle), but this network is constantly being rearranged.

The picture below is a "frame" from a simulation of 1,000 water molecules, provided by Prof. Valeria Molinero of the University of Utah Chemistry Department:



II. Ionization

Another consequence of the uneven sharing of electrons in water is that the covalent hydrogen-oxygen bonds break rather easily, generating H^+ and OH^- ions. This is a reversible and very rapid process:

$H_2O \rightleftharpoons H^+ + OH^-$

A major reason that this (and similar) reactions occur to a significant degree in water is that the ionic species can interact favorably with the other water molecules. In fact, the H^+ ions are not really free. Instead, they interact with groups of water molecules through hydrogen bonds. The hydrogen ion in solution is often represented as H_3O^+ to indicate that the ion is in close association with water molecules, but this, too, is a simplification. In non-polar solvents, there is essentially no tendency for molecules to ionize, because the solvent does not interact favorably with charged species.

Like any other reversible chemical reaction, there is an equilibrium constant for the dissociation of water:

$$K = \frac{[H^+][OH^-]}{[H_2O]} = 1.8 \times 10^{-16}\,\mathrm{M}$$

Because only a tiny fraction of the water ionizes, the concentration of neutral water is essentially constant, and the usual representation of the equilibrium constant ignores the water:

$$K_{\mathrm{wat}} = [H^+][OH^-] = 10^{-14}\,\mathrm{M}^2$$

It is common to write the concentrations of the H^+ and OH^- ions in a logarithmic form:

$$\mathrm{pH} = -\log[H^+]$$

$$\mathrm{pOH} = -\log[OH^-]$$

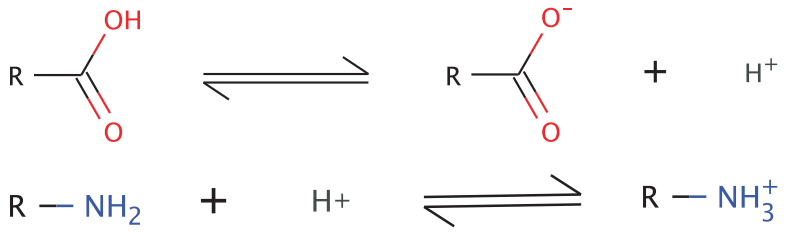
The equilibrium expression can then be written as:

$$\mathrm{pH} + \mathrm{pOH} = 14$$

This equation implies that if, for instance, we add H^+ ions to water (by adding an acid), the concentration of OH^- ions will go down. The reason for this is that some of the added H^+ ions combine with OH^- ions to form water. In general, pH is used much more commonly than pOH, but they both convey the same information, the balance between H^+ and OH^- ions.

The reason that chemists and biochemists give so much attention to pH is that other molecules can release or bind hydrogen ions as well, and the pH determines the balance of charged species.

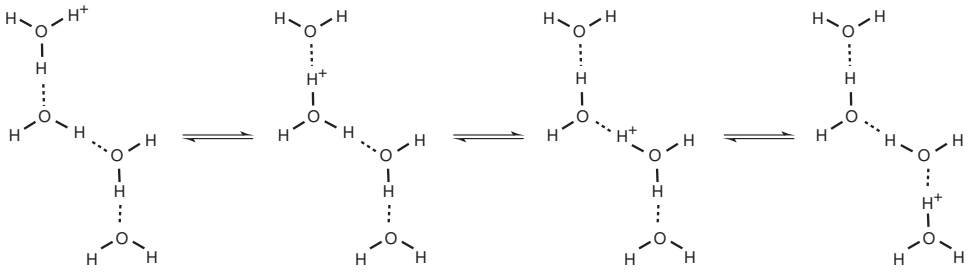
Two important examples of ionizing functional groups in organic molecules are the carboxyl groups and amino groups:



Groups that ionize do so for same basic reason as water does, an uneven distribution of electrons in covalent bonds. Different functional groups have different tendencies to release or take up H^+ ions. Ions are shuttled among different molecules in solution, including the water molecules. These reactions are typically very fast, on the order of microseconds. The exact balance between different charged forms of a molecule depends on the total concentration of free H^+ ions. Because a change in ionization results in a change in the electrical charge of a molecule, its chemical, structural and functional properties can be very sensitive to pH.

III. Dynamics of hydrogen ion diffusion.

Water, and molecules in it, form a highly dynamic solution, with the charges of molecules rapidly changing. Rates of exchange of H^+ ions from one molecule to another occur with times on the order of 1 μs, or less, allowing electric charge to be displaced in water very rapidly through relay processes. A mechanism for the rapid diffusion of hydrogen ions in water was was proposed by Theodor Grotthuss in 1806 and is illustrated below:



If you examine this diagram closely, you will see that the hydrogen ion that starts on the top-most water molecule doesn't really change position, but the electric charge moves through the rearrangement of hydrogen bonds. The basic idea of this mechanism is still thought to be correct, but details are still being studied and debated. This type of mechanism may also be important for the transport of hydrogen ions through some membrane channels.

IV. The hydrophobic effect

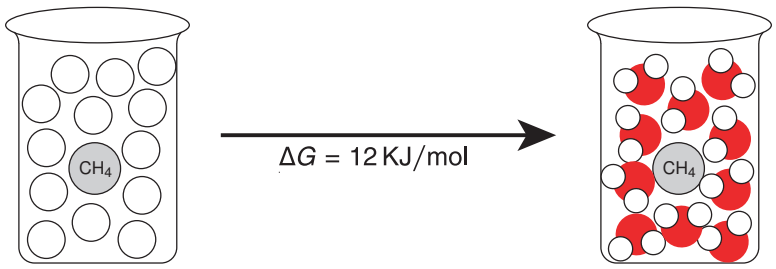
As mentioned earlier, one of the most important features of water is the fact that it doesn't mix well with non-polar molecules. This is a major driving force for the assembly of biological structures, because it leads to structures in which non-polar parts of molecules are sequestered away from water.

Why are non-polar molecules not very soluble in water? This is often referred to as the "hydrophobic effect", but, as we will see, non-polar molecules aren't really afraid of water.

Remember, water molecules *love* to form hydrogen bonds. What happens if a non-polar molecule does try to enter water? One thing it *doesn't* do is form hydrogen bonds with the water!

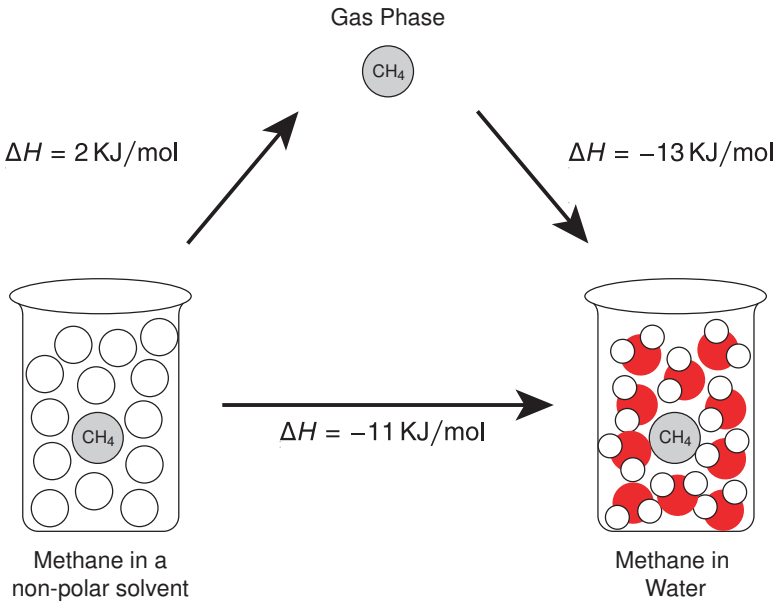
Does the non-polar molecule cause hydrogen bonds to break? It seems plausible. How can we find out?

Most of what we know about this phenomenon (or think we know) comes from thermodynamic measurements. The hydrophobic effect can be quantified by measuring the thermodynamics of transferring a non-polar molecule from a non-polar solvent to water. For instance, a molecule of methane from octanol to water:



In practice, the free-energy change is measured by determining the solubility of the molecule in each of the two solvents. As expected, the free energy of the process is positive, which simply means that oil and water don't mix.

We can also measure the enthalpy change for this process:



What is surprising is that ΔH is actually negative, which means that heat is released. This implies that, on average, there are more or stronger bonds in the aqueous solution with the non-polar molecule dissolved than when the non-polar molecule is in a non-polar solvent.

It's also possible to measure the enthalpy change for transfer of the non-polar molecule from each of the solutions to the gas phase. There is actually only a small positive ΔH for moving from the non-polar liquid to the gas phase, but a large negative ΔH for moving from the gas phase to water. Water and non-polar molecules actually interact quite strongly.

If ΔG is positive and ΔH is negative, ΔS must be negative for the transfer of the non-polar molecule to water. Somehow or other, the molecules become more ordered when a non-polar molecule is dissolved in water. Favorable processes for which ΔH is less than zero are often said to be *entropically driven*.

These observations lead to a model in which the water molecules rearrange themselves in some way around the non-polar molecule so that they lose entropy but actually form more or stronger hydrogen bonds. This seems rather counter intuitive, since it would seem easy for the water to just give up a few hydrogen bonds to accommodate the non-polar molecule. None the less, it seems that giving up some entropy is least bad way for water molecules to live with a non-polar molecule in their midst.

Further support for this model comes from another thermodynamic parameter, the heat capacity change at constant pressure, ΔC_{p}. This parameter is the derivative of ΔH with respect to temperature:

$$\Delta C_{\mathrm{p}} = \frac{d\Delta H}{dT}$$

ΔC_{p} can be measured by measuring ΔH as a function of temperature. For the transfer processes we are discussing, this gives a linear, or very nearly linear, plot and the slope is ΔC_{p}.

We can think about heat capacity as the amount of heat that is required to raise the temperature of a substance by 1°C. For either side of the reaction, we can write:

$$dH = C_p dT$$

Different substances have different heat capacities, because they have different ways of absorbing heat, including modes of motion and potentially breaking bonds.

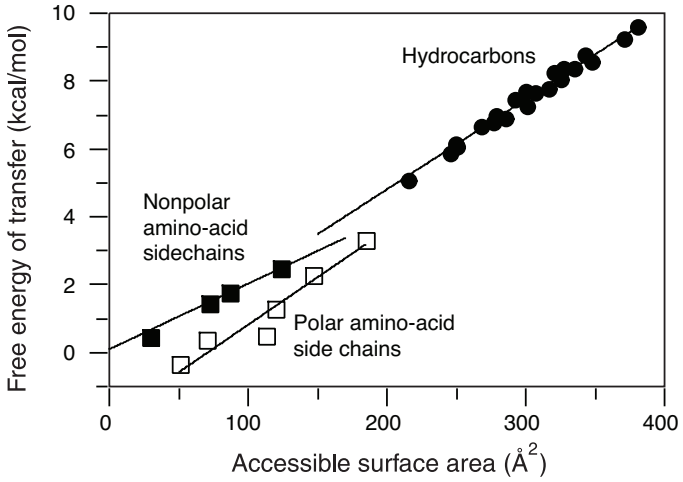
The heat capacity change for the transfer of a non-polar molecule to water is quite large and positive. This means that it takes more heat to raise the temperature of water when a non-polar molecule is present. This is consistent with the idea that the water forms more or stronger hydrogen bonds when the non-polar molecule is present, and these bonds break or weaken when the temperature is increased.

The positive heat capacity change also means that ΔH becomes less negative as the temperature increases. At higher temperatures, ΔH becomes positive, suggesting that introducing a non-polar molecule in water *does* lead to a net loss of hydrogen bonds.

This model is sometimes referred to as an "iceberg" model, but no one *really* understands it on a detailed structural basis.

"Hydrophobic effect" is a bad name, but it's somewhat better than "hydrophobic bonds" or "hydrophobic interactions", which are also used. The non-polar molecule doesn't fear water; it actually likes it pretty well! It's the water that has problems with its guest.

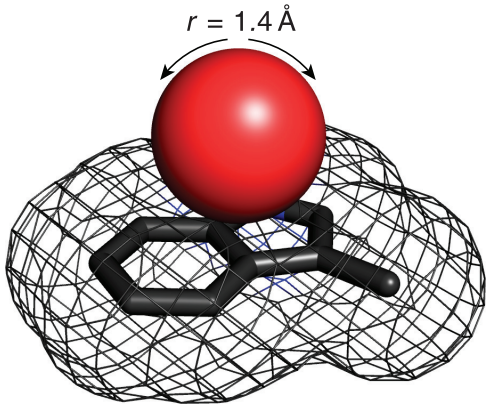
Another important observation about the hydrophobic effect is that the magnitude of the transfer free energy is proportional to the size of the non-polar molecule. This was pointed out in a classic review article by Fred Richards in 1977, which included a graph like the one below[1]:



In this graph, the filled circles represent hydrocarbons, the filled squares represent non-polar amino-acid side chains and the open squares represent polar amino-acid side chains.

[1]Figure adapted from Richards, F. M. (1977). Areas, Volumes, Packing and Protein Structure. *Annu. Rev. Biophys. Bioeng.*, 6, 151–176. http://dx.doi.org/10.1146/annurev.bb.06.060177.001055

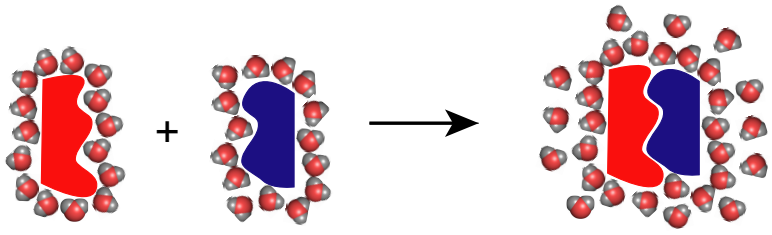
The quantity plotted on the horizontal axis is called the *accessible surface area* (ASA) and has units of $Å^2$. The process for calculating the accessible surface area for a known structure can be visualized as rolling a sphere representing a solvent molecule over the surface of the molecule and recording path followed by the point of the sphere closest to the molecule, as diagrammed in the figure below:



In this illustration, the molecule of interest is the side chain of a tryptophan residue and probe sphere has a radius of 1.4 Å, representing a water molecule. The surface area is represented by the wire cage surrounding the stick representation of the tryptophan side chain. For these calculations, the hydrogen atoms of both the molecule of interest and the water molecule are often ignored because they contribute relatively little to volume and surface area. Other measures of molecular size, such as volume, do not correlate with free energy of transfer nearly as well as surface area does. This is consistent with the idea that it is the interaction of the non-polar molecule with water and the interference with hydrogen bonding that define the hydrophobic effect.

Another important point that is illustrated in the graph on the previous page is that the amino-acid side chains that include polar groups lie on a distinct line from the one for those that contain only carbon and hydrogen atoms. The free energy of transfer for the side chains with polar groups are systematically less positive than the hydrocarbons. This reflects the more favorable interactions of non=polar groups with water.

The relationship between the hydrophobic effect and accessible surface area helps to explain how the hydrophobic effect can favor the assembly of biological structures, as illustrated in the cartoon below:



The cartoon represents two molecules with non-polar surfaces. As discussed above, these surfaces favor the ordering of water molecules in their vicinities, a process that

leads to the reduction of solvent entropy. When the two molecules interact via their complementary surfaces, the water that is closely associated with those surfaces is displaced, leading to an increase in entropy. Because many water molecules may be released by the association of two larger molecules, the increase in the solvent entropy can be significantly greater than the decrease in the associating molecules.

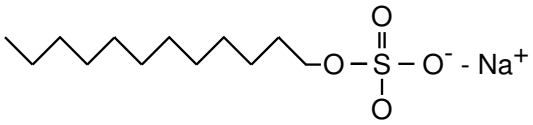
Other molecules or ions that are closely associated with the surfaces of relatively large molecules can also contribute to a favorable association reactions. An important example is the association of nucleic acids with proteins. Because they have a high density of negative charge at their surfaces, nucleic acids in solution are frequently associated with divalent cations, such as Mg^{2+}. The proteins that bind nucleic acids often contain positively charged side chains that interact with the negative charges of the nucleic acids and displace the cations. This results in an increase in entropy of the cations and contributes to a favorable free energy change for interaction between the protein and nucleic acid.

6.2 Lipid Bilayers and Membranes

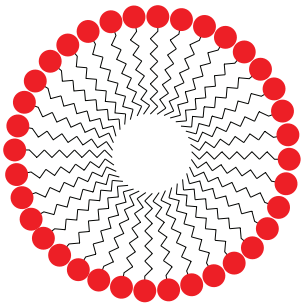
Some of the most important structures in living organisms are the membranes that separate the contents of cells from the extracellular environment and separate different intracellular compartments. The structures and properties of these membranes are largely determined by the hydrophobic effect discussed in the previous section.

I. Amphiphilic molecules, micelles and bilayers

When non-polar molecules exceed their solubility in water they simply form a separate phase, like in salad dressing. But, there are molecules that contain both polar and non-polar parts, and these molecules, called amphiphiles, can form more specific structures. An example of this type of molecule is a detergent. The structure of a typical detergent, sodium dodecyl sulfate (SDS), is shown below:



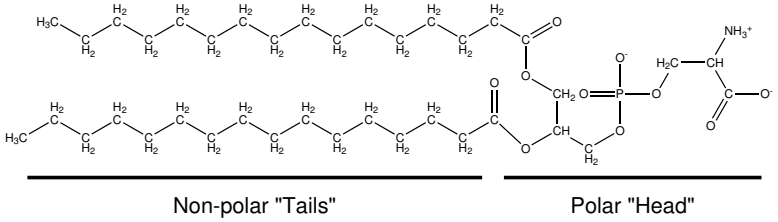
SDS is commonly used in biochemistry experiments, especially electrophoresis, and is also a common ingredient of household cleaning products, such as shampoos. The two parts of the molecule are a hydrocarbon chain (commonly referred to as the tail) with 12 carbon atoms and a sulfate group, a highly charged ion (the polar head). When molecules like this exceed their solubility, they assemble into a structures that sequester the non-polar part away from water while keeping the charged polar group exposed. These structures are called micelles and are roughly spherical, as illustrated in cross-section below:



These structures have a modest degree of specificity. For a given molecule, there will be a characteristic preferred size of micelle that optimizes the packing of the hydrophobic tails, while keeping the polar head-groups solvated with water. Typical micelle diameters lie in the range of about 3 to 50 nm.

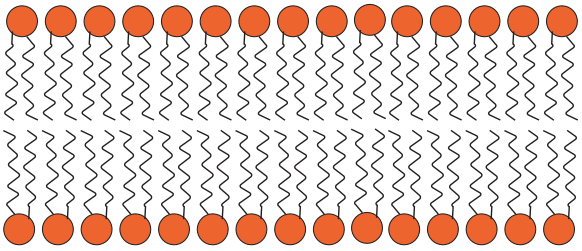
Molecules like this help us clean things up by solubilizing greasy molecules in the middle of the micelles.

The major amphiphilic molecules in biology have a slightly more complicated structure. These molecules typically have two hydrocarbon tails that are linked together by a glycerol molecule, a three carbon sugar, which is also linked to a polar phosphate group:



These molecules are called phospholipids, and there is wide variety in their structures. Different phospholipids have different hydrocarbon tails and different chemical groups attached to the phosphate.

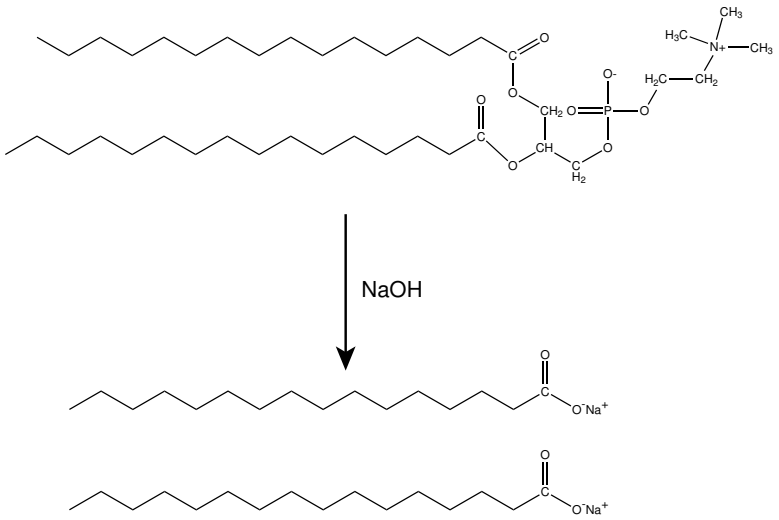
Like detergents, phospholipids form structures in water, with the non-polar groups sequestered and the polar groups interacting with water. But, rather than forming spheres, these molecules form extended (nearly) flat structures with two layers, as diagrammed below:



These structures form the membranes of cells, creating compartments with different chemical compositions.

The very different structures formed by detergents and phospholipids is largely due to the difference in the shapes of the molecules. Detergents tend to have a wedge shape that leads to sharp curvature, while phospholipids, because they contain two lipid tails, are more rectangular and form flat structures.

Treating phospholipids with strong bases (lye) hydrolyzes the ester bonds between fatty acids and glycerol, resulting in salts of the fatty acids, which are soaps.

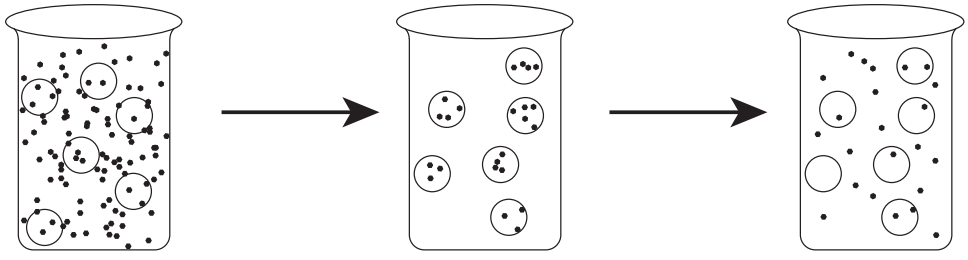


Like the detergents, soaps have a bit of a wedge-like structure, leading them to form micelles, rather than bilayers, so that they act in the same way as synthetic detergents. This reaction is probably one of the earliest examples of practical chemistry, dating back to the Roman Empire.

II. Permeability of bilayers

The major role of lipid bilayers is to enclose cells and form compartments within them with distinct chemical compositions. So, an important property is their permeability to different molecules.

Permeability can be measured by forming vesicles in the presence of specific molecules, separating the vesicles from free molecules, and then measuring the rates at which the molecules diffuse out of the vesicles, as illustrated below:



The rate at which the concentrations equilibrate is determined by Fick's first law:

$$J = -D\frac{dC}{dx}$$

where J is the flux (with units of $\text{mol} \cdot \text{s}^{-1}\text{m}^{-2}$); D is the diffusion coefficient (with units of m^2s^{-1}) and dC/dx is the concentration gradient across the membrane.

In this case, the area is the total area defined by the surface of the vesicle. The concentration gradient is the difference in the concentration divided by the thickness of the bilayer. The diffusion coefficient is a property of the molecule and the bilayer, and will generally be very different (smaller) than the diffusion coefficient of the same molecule in water. A typical vesicle might have a diameter of about 50 nm and typical bilayers have thicknesses of 3–4 nm. Thickness is a bit ambiguous, since it depends on how much of the polar head group is included, and different lipids have different lengths. In fact, the diffusion coefficient is not so easily defined because of the heterogeneous structure that the molecule has to cross.

Instead of using diffusion coefficients, the common practice for describing diffusion across bilayers is to introduce a *permeability coefficient* defined so that:

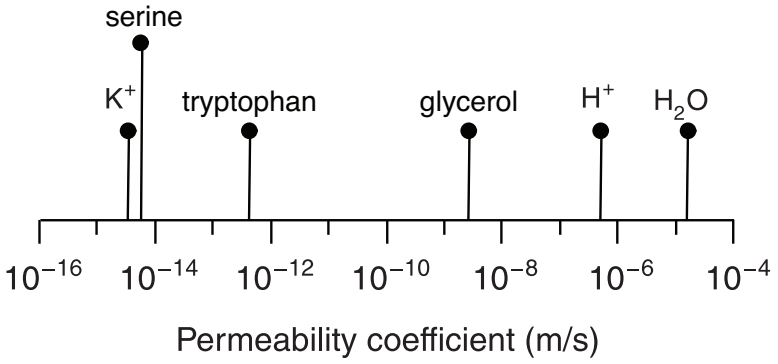
$$J = -D\frac{dC}{dx} = -P\Delta C$$

where ΔC is the difference in concentration across the membrane. In effect, the permeability coefficient combines the diffusion coefficient and the thickness of the membrane, so that:

$$P = \frac{D}{\Delta x}$$

where Δx is the thickness of the membrane, which is about 4 nm. The units for the permeability coefficient are m/s.

Some experimental values for the permeability coefficient are plotted below on a logarithmic scale[2]:



[2]Values in the figure are from: Chakrabarti, A. C. & Deamer, D. W. (1992). Permeability of lipid bilayers to amino acids and phosphate. *Biochim. Biophys. Acta - Biomembranes*, 111, 171–177. https://doi.org/10.1016/0005-2736(92)90308-9
and
Paula, S., Volkov, A. G., Van Hoek, A. N., Haines, T. H. & Deamer, D. W. (1996). Permeation of protons, potassium ions and small polar molecules through phospholipid bilayers as a function of membrane thickness. *Biophys. J.*, 70, 339–348. https://doi.org/10.1016/S0006-3495(96)79575-9

Some important points to note from these values are:

- The permeabilities cover 9 orders of magnitude.
- Ions have extremely low permeabilities, but H^+ is a notable exception.
- Polar molecules also have low permeabilities.
- Water has a quite high permeability.

We can compare the permeability coefficients to the diffusion coefficients for small molecules in water. The two parameters are related to one another according to:

$$D = P\Delta x$$

For the amino acid serine, assuming $\Delta x = 4\,\text{nm}$:

$$\begin{aligned} D &= 5 \times 10^{-15}\,\text{m} \cdot \text{s}^{-1} \times 4 \times 10^{-9}\,\text{m} \\ &= 2 \times 10^{-23}\,\text{m}^2\text{s}^{-1} \end{aligned}$$

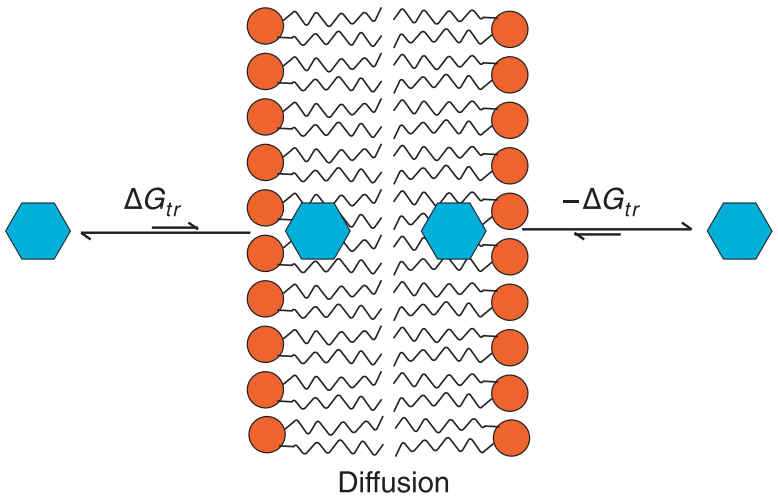
This is *much* less than the diffusion coefficient of a small molecule in water, which is about $10^{-10}\,\text{m}^2\text{s}^{-1}$.

For water:

$$\begin{aligned} D &= 1.6 \times 10^{-5}\,\text{m} \cdot \text{s}^{-1} \times 4 \times 10^{-9}\,\text{m} \\ &= 6.4 \times 10^{-14}\,\text{m}^2\text{s}^{-1} \end{aligned}$$

This is still quite small.

For many molecules, the observed permeability coefficients can be accounted for by a model in which there is an equilibrium between the molecule in water and in the lipid bilayer, coupled to diffusion within the bilayer and then rapid escape from the bilayer back to the water phase:



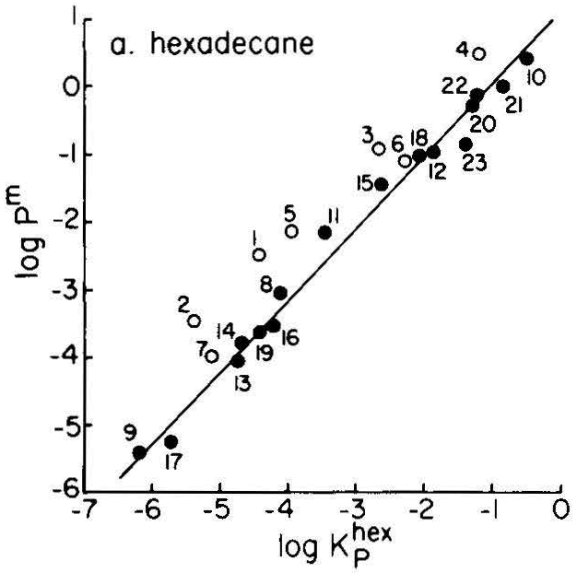
$\Delta G_{\rm tr}$ is the free energy of transfer from water to a non-polar environment and is very unfavorable for a polar molecules . This model, referred to as the *solubility-diffusion model*, can be used to predict the permeabilities of molecules from the values of ΔG_{tr} and the diffusion coefficients in oils. This model assumes that the overall rate of crossing the bilayer is determined by the fraction of molecules that are present in the bilayer, relative to the water phase on each side, and the diffusion coefficient in the non-polar part of the bilayer. The fraction of molecules in the bilayer is equal to the equilibrium constant for transfer from water to the non-polar environment, $K_{\rm tr}$, provided that this equilibrium constant is much less than 1. The effective diffusion coefficient is then given by:

$$D_{\rm eff} = K_{\rm tr} \times D$$

and the permeability coefficient can be expressed as:

$$P = P = \frac{D_{\rm eff}}{\Delta x} K_{\rm tr} D/\Delta x$$

This model predicts that the permeability coefficients for different molecules should be correlated with their relative solubilities in non-polar liquids. This prediction has been borne out for many, but not all, molecules that have been examined, as illustrated in the figure below:[3]



This correlation provides a strong argument that the solubility-diffusion model is a good description for the permeabilities of small molecules, both polar and non-polar, and most small ions. The largest discrepancies appear to be for water and H^+ ions, which have anomolously high permeability coefficients.

[3]Figure from Walter, A. & Gutknecht, J. (1986). Permeability of small nonelectrolytes through lipid bilayer membranes. *J. Membrane Biol.*, 90, 207–217. http://dx.doi.org/10.1007/BF01870127

An alternative model would be that holes transiently form in the bilayer and allow molecules to pass through. However, this model would suggest that the permeability coefficients would be relatively independent of the polarity of the molecules, which is clearly not the case. On the other hand, the permeability data for some species, especially H^+ ions and water, do not fit the solubility-diffusion model, indicating that other factors may play a role. It has been suggested that chains of hydrogen-bonded water molecules may cross the bilayer and allow the net movement of both water molecules and H^+ ions, the latter by a Grotthuss mechanism (page 170). In addition, transient defects may form and contribute to permeability, but this appears to be a relatively small factor for most molecules.

One important application of kind of data is in the design of pharmaceuticals, since drug molecules generally have to cross multiple membranes to reach their targets. It may also be important to keep the drug from crossing other membranes. This can make or break a potential drug.

We can apply the permeability coefficients to estimate the rate of molecules entering or leaving cells by passive diffusion across the bilayer (in the absence of transport by specific membrane proteins). Suppose that we have a cell with a diameter of $20\,\mu$m, and it contains no glucose, but is surrounded by a solution that is 0.1 M in glucose. How rapidly will glucose, with a permeability coefficient of about 5×10^{-10} m/s, enter this cell? From the permeability coefficient and the concentration difference, we can calculate the flux, J, in units of $\text{mol} \cdot \text{s}^{-1}\text{m}^{-2}$:

$$\begin{aligned} J &= P\Delta C \\ &= 5 \times 10^{-10}\,\text{m/s} \times 0.1\,\text{M} \\ &= 5 \times 10^{-11}\,\text{mol} \cdot \text{L}^{-1}\text{m} \cdot \text{s}^{-1} \\ &= 5 \times 10^{-8}\,\text{mol} \cdot \text{s}^{-1}\text{m}^{-2} \end{aligned}$$

Next, we calculate the surface area of the cell:

$$\begin{aligned} A &= 4\pi r^2 \\ &= 4\pi(10^{-5}\,\text{m})^2 \\ &\approx 10^{-9}\,\text{m}^2 \end{aligned}$$

The total flow into the cell is the flux multiplied by the surface area:

$$\begin{aligned} \text{flow} &= 5 \times 10^{-8}\,\text{mol} \cdot \text{s}^{-1}\text{m}^{-2} \times 10^{-9}\,\text{m}^2 \\ &= 5 \times 10^{-17}\,\text{mol} \cdot \text{s}^{-1} \end{aligned}$$

That's not a lot of moles per second, but it is about 30 million molecules per second.

How rapidly would the intracellular concentration change? We need to calculate the volume:

$$\begin{aligned} V &= \frac{4}{3}\pi r^3 \\ &= \frac{4}{3}\pi(10^{-5}\,\mathrm{m})^3 \\ &\approx 4\times 10^{-15}\,\mathrm{m}^3 \times \frac{10^3\,\mathrm{L}}{1\,\mathrm{m}^3} \\ &\approx 4\times 10^{-12}\,\mathrm{L} \end{aligned}$$

So, the rate of change in concentration is:

$$\begin{aligned} \frac{dC}{dT} &= \frac{5\times 10^{-17}\,\mathrm{mol}\cdot\mathrm{s}^{-1}}{4\times 10^{-12}\,\mathrm{L}} \\ &\approx 10^{-5}\,\mathrm{mol}\cdot\mathrm{L}^{-1}\mathrm{s}^{-1} \end{aligned}$$

It will take a long time for the intracellular concentration to equilibrate with the extracellular environment.

III. Primitive membranes

The impermeability of bilayers raises an interesting question with regard to the origins of life, one of the great intellectual challenges. The basic problem is that modern life forms appear to be so perfect and complicated that it is hard to imagine how any part of it could have evolved by itself. Modern organisms have to:

- Collect nutrients
- Convert nutrients into useable forms of energy
- Build complicated macromolecules, including enzymes and genetic material
- Create compartments bounded by membranes
- Reproduce themselves

For a long time, a big issue was whether proteins or nucleic acids came first. Proteins are needed for enzymes, but DNA and RNA are needed to encode proteins. In the 1980s, however, it was discovered that some RNA molecules have catalytic activities. It's now widely believed that the earliest biological macromolecules were RNA molecules that had very limited ability to catalyze their own replication.

Membranes also present a problem. A key event in the evolution of early cells must have been the formation of membranes, so that nutrient molecules could be sequestered and not shared with competing cells or molecules. Modern lipid bilayers are extremely impermeant to polar molecules and ions. Translocation of molecules across membranes depends on protein molecules embedded in the bilayer, which allows for the transport to be controlled. But, how could primordial membranes have worked?

It turns out that some fatty acids can form bilayers that are much more permeant to polar and even charged species. A current idea is that vesicles of this type formed and trapped RNA, or related molecules, that could polymerize. Precursors to polymers could diffuse across the membranes, but when they were incorporated in polymers they were trapped. As molecules inside the vesicles got larger and more numerous, vesicles were forced to expand and eventually divide.

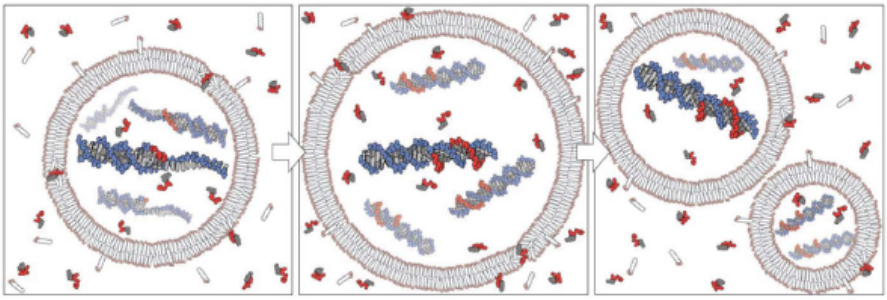


Figure from:

Mansy, S. S., Schrum, J. P., Krishnaurthy, M., Tobé, S., Treco, D. A. & Szostak, J. W. (2008). Template-directed synthesis of a genetic polymer in a model protocell. *Nature*, 454, 122–125.
http://dx.doi.org/10.1038/nature07018

Another, related, reference:

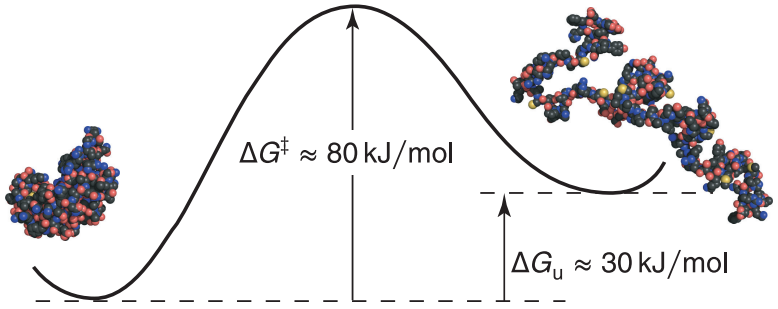
Monnard, P.-A., Luptak, A. & Deamer, D. W. (2007). Models of primitive cellular life: polymerases, and templates in liposomes. *Phil. Trans. Royal. Soc. Lond. B*, 362, 1741–1750.
http://dx.doi.org/10.1098/rstb.2007.2066

6.3 Protein Folding and Unfolding

One of the most important examples of biological structure formation is the folding of polypeptide chains into stable three dimensional structures. This process occurs largely after proteins are synthesized by ribosomes, although the first segments of a protein that are synthesized may begin to fold while the rest of the chain is still being synthesized. For most proteins, the folded conformation is required for function, and this conformation is specified by the amino acid sequence. As a consequence the folding process represents a point in which one-dimensional information encoded in the genome is expressed as a three-dimensional structure. With many proteins, it is possible to unfold the folded, or native, structure by altering the solution conditions and then reform the structure by again changing the conditions. This provides the means of studying the thermodynamics and mechanisms of protein folding. Because of its central importance in biology, this process has been studied extensively, both experimentally and by theoretical and computational methods.

I. Native and unfolded protein states

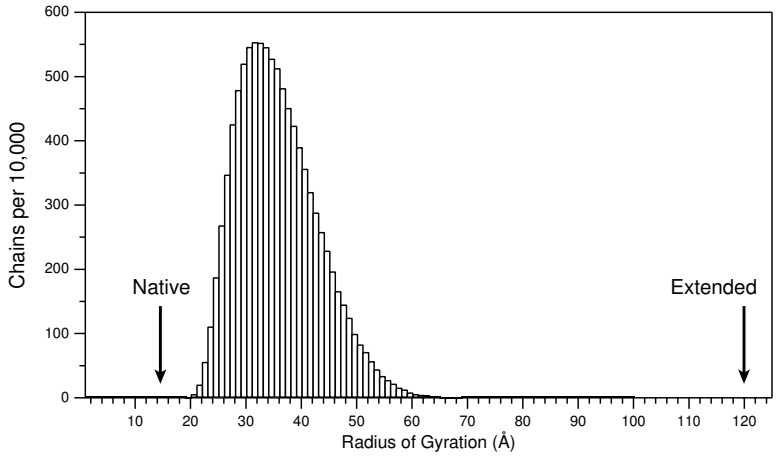
Although there is tremendous variety in the size and architecture of protein molecules, the great majority of the experimental work that has been done to date has focused on relatively small proteins, containing about 50–200 amino-acid residues, and our discussion will be largely limited to this class of proteins. Proteins of this size typically fold into a single compact structure, termed a domain, whereas larger proteins are often formed of multiple domains. One important characteristic feature of single-domain proteins is that the unfolding and folding processes is highly cooperative, meaning that partially folded structures are significantly less stable than either the native state or completely disordered molecules. In addition, the kinetics of folding and unfolding can often be described by models involving a single major transition state or energy barrier, as diagrammed below:



The structures in the diagram represent ribonuclease A (RNAse A), which contains 124 amino-acid residues, in its folded conformation (on the left) and a disordered conformation. As indicated in the diagram, the native state is more stable than the unfolded by a few tens of kJ/mol. The activation free energy indicated in the diagram, 80 kJ/mol corresponds to a rate constant for folding of about $0.1\,\mathrm{s}^{-1}$, or a half time of about 6 s. This value falls within the quite wide range of observed folding times, from minutes to microseconds.

Whereas the folded conformation is a relatively unique structure, the unfolded state is a broad distribution of rapidly interconverting conformations, contrary to the impression that may be conveyed by the figure above. The conformation shown in the diagram is one of many generated in a computational simulation designed to explore the properties of unfolded proteins[4] This simulation generated approximately 200,000 conformations, and the figure below shows their distribution with respect to overall dimension:

[4]Goldenberg, D. P. (2003). Computational simulation of the statistical properties of unfolded proteins. *J. Mol. Biol.*, 326, 1615–1633. http://dx.doi.org/10.1016/S0022-2836(03)00033-0



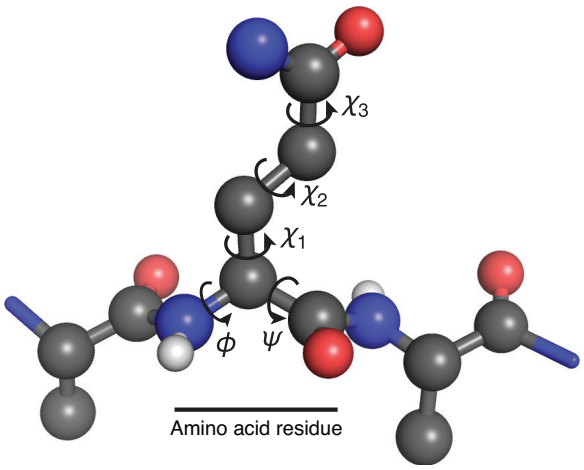
The radius of gyration, as plotted on the horizontal axis, of a polypeptide chain is the root-mean-square distance between the center of mass and each non-hydrogen atom in the molecule[5]. It is thus a useful measure of the overall size of a molecule in a particular conformation. For reference, the figure also indicates the radii of gyration of native RNAse A and a fully extended conformation. Although the distribution includes conformations that are nearly as compact as the folded state, there are none that approach the radius of gyration of a fully extended chain. One consequence of the very broad distribution of unfolded states is that it has much more entropy than the native conformation, as discussed further below.

II. Entropy of the unfolded state

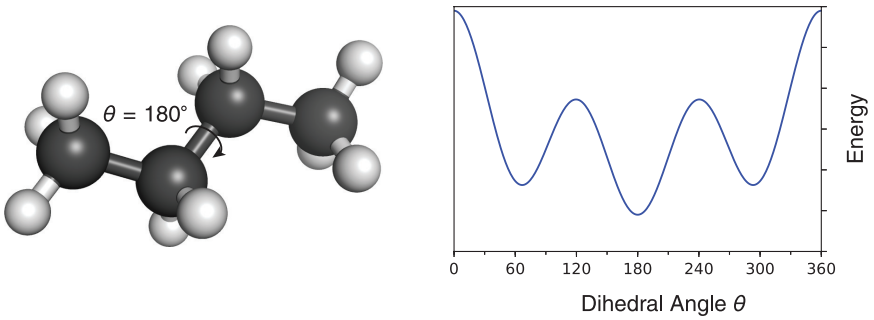
The large difference in entropy between the native and unfolded states of a protein is expected to disfavor the native state. It is thus useful to try to estimate the magnitude of this difference. If we can somehow count the number of microstates making up the two states, we can use the Boltzmann equation to calculate the entropy difference. To do this rigorously would require consideration of all of the possible alternate conformations in the two states, which is a formidable challenge. We can, however, make some simplifying assumptions that allow for a very rough approximation. The key assumptions are that the folded protein has a single, unique conformation and that the individual amino-acid residues are able to take on multiple, independent conformations in the unfolded state.

The conformations accessible to an individual reside can be defined by the dihedral angles that describe the rotations about single covalent bonds, as illustrated below for a glutamine residue:

[5]This definition is somewhat simplified from the more general definition of the radius of gyration, but it is a good approximation for an object in which all of the elements have approximately the same mass.



For each residue (other than prolines), there two rotatable bonds in the polypeptide backbone, labeled ϕ and ψ, and additional rotatable bonds in the side chain, which are labeled χ_1, χ_2 and so on, depending on the particular residue type. These bonds can undergo rotation, to varying degrees, in both the folded and unfolded states. However, these rotations are associated with changes in the potential energy of the molecule, due to steric and other interactions among the atoms. These effect create a pattern of valleys and peaks in the energy as a bond is rotated, as illustrated in the diagram below for the simple case of a propane molecule:



As the central bond is rotated, the lowest energy is found when the methyl groups on the two sides of the bond are pointed in opposite directions, and the highest energy is observed when the methyl groups are on the same side of the central bond. There are two other minima, where the bond is rotated by 120° from the lowest energy position. Bonds in different contexts have different rotational energy profiles, but the patterns are similar.

Within a folded protein, most of the rotatable bonds are restricted a single minimum, but the dihedral angles fluctuate about those minima. In the unfolded state, the dihedral angles can sample all of the minima and fluctuate within these minima. Thus, we can think of the reduction in conformational entropy as a reduction in the number of accessible minima, and we can define the microstates in term of distinct conformations with the dihedrals in specified minima. This approximation is referred to as the *rotational isomeric state* model.

For the native protein we assume that there is a single microstate, recognizing that this microstate includes all of the fluctuations about the dihedral minima, as well as vibrational motions. For the unfolded state, we assume that the number of microstates of each residue is approximately 10-times the number in the native state. Therefore the ratio of the number of microstates for a single residue in the two states is:

$$\frac{\Omega_{\mathrm{U}}}{\Omega_{\mathrm{N}}} = 10$$

For two residues, if the accessible conformations are independent, the ratio is 10^2; for three residues, the ratio is 10^3, and so on. For an n-residue protein, therefore, the ratio is:

$$\frac{\Omega_{\mathrm{U}}}{\Omega_{\mathrm{N}}} = 10^n$$

The entropy change is then:

$$\Delta S_{\mathrm{conf}} = k \ln \frac{\Omega_{\mathrm{U}}}{\Omega_{\mathrm{N}}} = k \ln 10^n$$

This quantity is designated ΔS_{conf} to emphasize that it reflects only the change in polypeptide conformation for unfolding and that there are other contributions to the overall entropy change for the process. For a protein containing 100 amino-acid residues, the conformational entropy change is:

$$\Delta S_{\mathrm{conf}} = k \ln 10^{100} = 3.3 \times 10^{-21}\,\mathrm{J/K}$$

On a molar basis:

$$\Delta S_{\mathrm{conf}} = R \ln 10^{100} = 1900\,\mathrm{J/(K \cdot mol)}$$

and the free energy contribution to unfolding at 300 K is:

$$-T\Delta S_{\mathrm{conf}} = -570\,\mathrm{kJ/mol}$$

Note that this is a large factor favoring unfolding. It is approximately ten-fold greater than the net free energy change that favors folding for a typical protein of this size.

One important, and unrealistic, assumption that went into this calculation is that the conformations of individual amino-acid residues are independent of one another in an unfolded protein. Manipulating a physical model of a peptide containing only a few residues will demonstrate that many combinations of dihedral angles will lead to steric clashes among the atoms. Suppose that we assume that only one in 10^{10} (one in 10 billion) of the conformations that we assumed in our calculations is actually possible. For the 100-residue protein, the ratio of microstates in the unfolded and native states is then reduced to:

$$\frac{\Omega_{\mathrm{U}}}{\Omega_{\mathrm{N}}} = 10^{100} \div 10^{10} = 10^{90}$$

The corresponding molar entropy and free energy changes are:

$$\Delta S_{\text{conf}} = R \ln 10^{90} = 1700\,\text{J/mo}$$
$$-T\Delta S_{\text{conf}} = 520\,\text{kJ/mol}$$

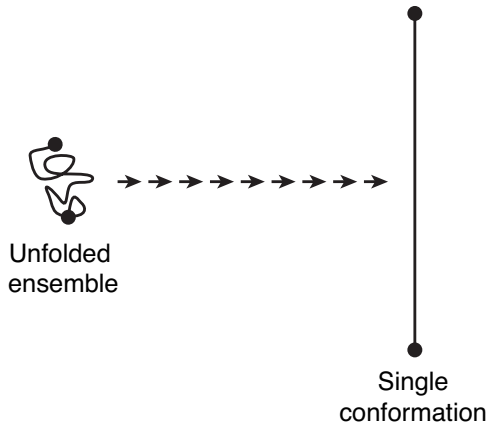
Thus, even reducing the initial estimate of the number of possible conformations by a factor of 10 billion does not alter the general conclusion from this calculation.

It is also possible to estimate the value of ΔS_{conf} from experimental measurements. Recall, from Chapter 6, the general definition of the entropy change (for the system) for a process at constant temperature:

$$\Delta S_{\text{sys}} = \frac{q_{\text{rev}}}{T}$$

where q_{rev} is the heat absorbed by the system during the reversible change from one state to another. If the process is carried out at constant temperature, and there are no changes in the potential energies of the molecules making up the system, $\Delta E = 0$ and $q_{\text{rev}} = -w_{\text{rev}}$, where w_{rev} is the work done on the system during the process. If we could somehow measure the work required to convert the very broad ensemble of unfolded conformations (very slowly) to a single conformation, we could determine the associated entropy change.

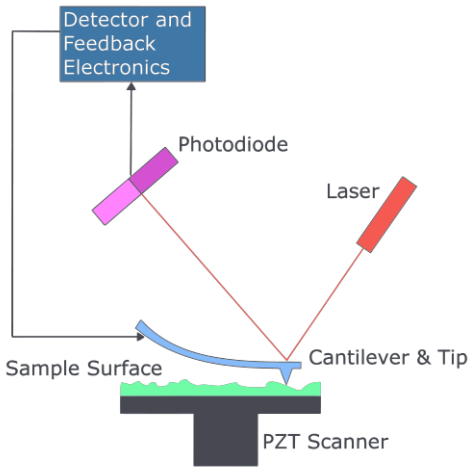
It might seem that the process to examine in this way would be the conversion of the unfolded state to the native state. But, this is problematic because folding is associated with the formation and rearrangement of numerous non-covalent interactions that alter the potential energy of the molecule, so that ΔE is not equal to zero. On the other hand, we can consider the conversion of the unfolded ensemble to a fully extended conformation, as diagrammed below:



Because the intramolecular interactions in the folded state are largely broken in the unfolded state, there should not be significant changes in the interactions of the polypeptide chain with either itself or the solvent when the molecules is stretched to its maximum extension. Thus, in terms of conformational entropy, the fully extended chain is equivalent to the fully folded chain! A caveat to this assumption is that the side

chains of many residues will be more restricted in the native state than in the fully extended conformation. Thus, the decrease in conformational entropy for stretching the unfolded ensemble to a fully extended conformation is expected to be somewhat less than that for folding.

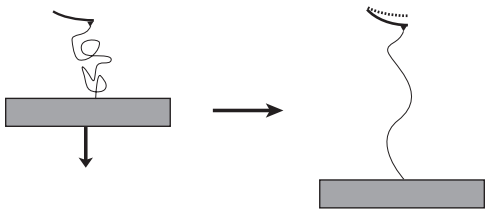
To actually measure the work for stretching a single protein requires very sensitive instrumentation, and two types of instrument have been employed in this kind of experiment, optical tweezers and atomic-force microscopes (AFM). Here we will focus on the AFM, a version of which is diagrammed below[6]:



This instrument was invented primarily for recording very high resolution images of surfaces. The essential elements are a very fine probe, with a tip diameter of a few nm or less, and a sample stage that can be moved with nanometer precision. The probe is attached to a flexible cantilever and brought into contact with the sample surface mounted on the stage. As the surface is moved in two directions, the probe tip follows the surface, and the cantilever bends slightly to accommodate this motion. The cantilever contains a reflective surface, and the light from a stationary laser is reflected from it. As the sample is scanned, the direction of the reflected light changes in response to the changes in vertical position of the tip. The fluctuations are recorded and converted into a record of the height of the surface as a function of position, thus generating an image. The tip can also be used to manipulate or move individual molecules, and even atoms. The key to this precision is the ability to move the stage in tiny, reproducible steps. This is made possible by the use of piezoelectric crystals that undergo small size changes in response to electrical voltage changes.

An AFM can also be used to measure forces generated by the motion of the stage. For this purpose, the cantilever, which acts as a spring, is calibrated so that the displacement of the tip can be converted into a force value. An arrangement for stretching a polypeptide chain is shown in the diagram below:

[6]Figure from https://en.wikipedia.org/wiki/Atomic-force_microscopy



Using some combination of genetic engineering and protein chemistry, one end of the chain is attached to the probe tip and the other to the movable stage. The stage is then moved downward, which reduces the entropy of the chain and creates a downward force on the tip. In response, the cantilever bends and, in doing so, exerts an opposite force, which increases as the stretching proceeds. By carrying out this process very slowly, to approach the ideal of reversibility, and continuously recording the force of the cantilever, the total work can be determined as:

$$w = \int F dx$$

Actually carrying out this kind of experiment, and properly analyzing the data is very challenging. None the less, the measurement has been performed for a variety of proteins and the quantitative results are quite consistent with the estimates of ΔS_{conf} based on counting rotational isomers[7].

To summarize, these experiments and calculations based on the rotational isomeric state model indicate that a reasonable estimate for the change in conformational entropy for a polypeptide of n residues is on the order of:

$$\Delta S_{\text{conf}} = k \ln 10^n$$

For a 100-residue protein, ΔS_{conf} is calculated from this relationship to be $1900\,\text{J}/(\text{K}\cdot\text{mol}$. This factor favors unfolding, by $570\,\text{kJ/mol}$ at $300\,\text{K}$. In the next section, we consider the factors that overcome this entropy penalty to make the folded structures of proteins stable under physiological conditions.

III. Protein-stabilizing factors

The thermodynamics of unfolding have been studied experimentally for a large number of single-domain proteins, most with chain lengths ranging from about 50 to 200 amino acid residues. Although the values of the parameters vary substantially, the values below, for hen egg-white lysozyme at 25°C[8], are typical for a protein in this size class:

[7]Thompson, J. B., Hansma, H. G., Hansma, P. K. & Plaxco, K. W. (2002). The backbone conformational entropy of protein folding: Experimental measures from atomic force microscopy. *J. Mol. Biol.*, 322, 645–652. http://dx.doi.org/10.1016/S0022-2836(02)00801-X

[8]Baldwin, R. L. (1986). Temperature dependence of the hydrophobic interaction in protein folding. *Proc. Natl. Acad. Sci., USA*, 83, 8069–8072. http://dx.doi.org/10.1073/pnas.83.21.8069

ΔG_{u}	$60.7\,\mathrm{kJ/mol}$
ΔH_{u}	$236\,\mathrm{kJ/mol}$
ΔS_{u}	$586\,\mathrm{J/(K \cdot mol)}$
-TΔS_{u}	$175\,\mathrm{kJ/mol}$

Hen lysozyme contains 129 amino-acid residues, leading to a predicted value of ΔS_{conf} of about $2500\,\mathrm{J/(K \cdot mol)}$, based on the assumptions introduced in the previous sections. Note that this value of ΔS_{conf} is approximately four-fold greater than the observed entropy change for unfolding. Further, the magnitude of $-T\Delta S_{\mathrm{conf}}$, $-740\,\mathrm{kJ/mol}$, is about ten-fold greater than that of ΔG_{u}, but of the opposite sign, representing a large factor favoring unfolding. We thus need to account for two apparent discrepancies associated with the large calculated value for ΔS_{conf}:

- The conformational entropy change for unfolding is much larger than the observed entropy change for the overall unfolding process.
- The free energy change associated with the conformational entropy change is far larger than the free energy change for unfolding and greatly favors the unfolded protein under conditions where the folded state is stable.

There must be at least one other factor that contributes a large negative change in entropy upon unfolding and a positive contribution to the free energy change for unfolding. The most likely explanation is the transfer of non-polar parts of the protein, which are buried in the native state, to the water solvent in the unfolded state. Recall from the discussion of the hydrophobic effect (pages 171–175) that the transfer of non-polar molecules from a non-polar liquid to water is associated with both a positive free energy change and a negative entropy change. Although it is not completely understood, it is generally believed that the decrease in entropy is due to an increase in order of water molecules directly surrounding the non-polar atoms, sometimes referred to as an iceberg effect.

As discussed earlier, the magnitude of the positive free energy change for transfer of a non-polar molecule to water is closely correlated with the accessible surface area (ASA) of the molecule (pages 173–174). From transfer measurements of numerous molecules and careful analysis of the data, simple relationships have been derived between surface areas of non-polar parts the molecules (A_{np}) and the thermodynamic parameters for transfer. The values vary quite strongly with temperature, and the expressions below are for 298 K.

$$\begin{aligned}
\Delta H_{\mathrm{tr}} &= A_{\mathrm{np}} \times 7\,\mathrm{J/mol} \\
\Delta S_{\mathrm{tr}} &= -A_{\mathrm{np}} \times 0.3\,\mathrm{J/(mol \cdot K)} \\
-T\Delta S_{\mathrm{tr}} &= -A_{\mathrm{np}} \times 90\,\mathrm{J/mol} \\
\Delta G_{\mathrm{tr}} &= A_{\mathrm{np}} \times 97\,\mathrm{J/mol}
\end{aligned}$$

As discussed earlier, the overwhelmingly predominant component of the unfavorable transfer free energy change is entropic. Polar surface area also influences the transfer thermodynamics, but this effect is much smaller than the influence of the non-polar surface area and will be ignored here.

From the structure of a folded protein, it is relatively straight forward to calculate the accessible surface area, and to distinguish between the polar and non-polar components of that surface. For the unfolded ensemble, some kind of model must be used because of the broad range of conformations. Calculations for hen lysozyme lead to the values listed below for the accessible surface areas of the native and unfolded states:

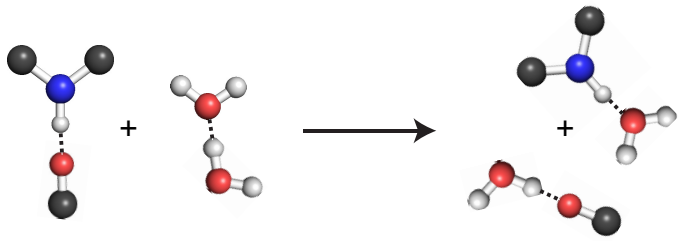
	Native (Å^2)	Unfolded (Å^2)	ΔASA(Å^2)
Total	6,670	15,800	9,130
Non-polar	3,400	9,700	6,300
Polar	3,300	6,100	2,800

Note that both polar and non-polar groups are accessible to solvent in the native state, contrary to the common perception that non-polar groups are almost entirely buried in folded proteins. None the less, non-polar groups are disproportionately buried in the native state and the non-polar accessible surface area increases greatly upon unfolding. Using these values and the expressions relating non-polar surface area to the thermodynamic parameters, we can estimate the contributions of the hydrophobic effect to the overall unfolding thermodynamics. The table below includes these estimates for hen lysozyme in an overall balance sheet.

	ΔH kJ/mol	ΔS J/(mol $\cdot$ K)	ΔG kJ/mol
Conformational entropy		2,500	-740
Hydrophobic effect	44	-1,900	610
Other	192	-14	190
Overall	236	586	61

The row labeled "Other" in the table above represents the additional contributions to the enthalpy, entropy and free-energy changes that must be added to those from conformational entropy and the hydrophobic effect in order to match the observed values for unfolding of hen lysozyme. Note that the hydrophobic effect balances out about 75% of the favorable conformational entropy change for unfolding, leading to a residual (600 J/(mol $\cdot$ K)) that very closely matches the overall entropy change for unfolding. On the other hand, there is an additional contribution of about 200 kJ/mol to ΔH_{u} that is not yet accounted for.

Qualitatively, the positive enthalpy change for unfolding indicates that attractive interactions are broken during unfolding. These interactions likely include hydrogen bonds and van der Waals interactions, both of which are apparent in the folded structures of proteins. Estimating the energetic contributions of individual interactions of this type is quite difficult, however, because interactions in the native state that are disrupted upon unfolding are likely compensated for by new interactions between the protein and the solvent. For instance, the folded proteins contain a large number of hydrogen bonds between the amide nitrogen atoms and carbonyl oxygen atoms of different residues, as illustrated in the right-hand side of the figure below:



At the same time, the water molecules surrounding the protein are extensively hydrogen bonded to each other. When the protein unfolds and the intramolecular hydrogen bonds are broken, the protein nitrogen and oxygen atoms can readily form new hydrogen bonds with water molecules. Breaking a hydrogen bond of this type, without replacing it, requires an energy input of about 50 kJ/mol. However, estimating the net energetic effect of breaking two hydrogen bonds (one in the protein and one between water molecules, as in the illustration above) is very difficult and has been the subject of controversy for nearly sixty years. The net energy difference is likely to be far smaller than the 50 kJ/mol for breaking a hydrogen in isolation, but the number of hydrogen bonds in a folded protein is quite large, about 100 in a protein the size of hen lysozyme. Thus the total contribution to hydrogen bonds could be quite significant.

One important source of information about the contributions of individual interactions to protein stability is experiments in which specific amino-acid residues have been modified by genetic engineering and the effects of these changes on unfolding thermodynamics have been been measured. For instance, a serine residue, in which the side-chain hydroxyl group forms a hydrogen bond with another protein group, can be changed to an alanine residue to eliminate the hydrogen bond. A large number of experiments of this type have been performed over the last few decades, and one of the major observations has been that the effects of this kind of change can vary greatly from protein to protein and among different sites within the same protein. Thus, the contributions of different types of interactions appear to be highly context dependent. For amino-acid replacements that remove hydrogen bonding groups, there is typically a reduction in ΔG_{u} in the range of 5–10 kJ/mol per hydrogen bond. When multiplied by the number of hydrogen bonds in a folded protein, these effects could readily account for a large fraction of the enthalpy change observed for unfolding.

Estimates of the kind discussed here provide a reasonably satisfying accounting for the observed thermodynamics of protein folding, with the predominant contributions

assigned to conformational entropy and the hydrophobic effect. The balancing of the conformational entropy change and the entropy change attributed to the hydrophobic effect seems particularly close. Some caution in interpreting these estimates is called for, however, since they are based on assumptions with significant uncertainties. Some of the greatest uncertainties concern the properties of the unfolded state, which influences both the estimate of ΔS_{conf} and the net contribution of the hydrophobic effect. Unfortunately, the very nature of unfolded states makes them much more difficult than folded proteins to characterize experimentally. This, among other areas of protein conformation and dynamics, continues to be an important area of research.

Chapter 7

Molecular Motors

The cells of nearly all (if not all) organisms contain protein complexes that are able to convert chemical energy into mechanical work. These motors include:

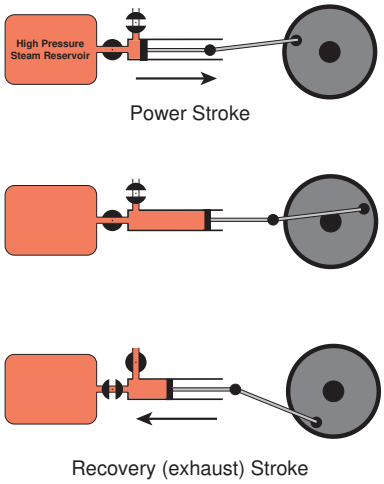
- Myosin, which generates force by interacting with actin in muscle and other cells.
- Kinesin and dynein, which both move along microtubules, but in opposite directions.
- The ATP synthase that produces ATP from ADP in mitochondria and aerobic bacteria, using the electrochemical potential of a H^+ gradient. Force generation is not part of the normal function of this protein, but its catalytic activity is coupled to rotary motion.
- The bacterial flagellar motor, which uses electrochemical gradients to rotate flagella and enable swimming.

In all of these motors, conformational changes in protein complexes are coupled to chemical changes, either ATP synthesis/hydrolysis or motion of ions across membranes.

7.1 Some Basic Principles

I. Steam engines

In thinking about how molecular motors might work, it is useful to consider first a simple engine on a larger scale, such as a steam engine.



Note the following important features of the engine illustrated above:

- The energy source for the engine is a pressure difference, which is created by a temperature difference.
- The free energy of the steam (its ability to do work) is lost as it expands, and its entropy increases.
- Expansion of the gas is coupled to movement of the piston and, in turn, turning the wheel and shaft.
- After the powerstroke, the momentum of the flywheel returns the engine to its starting state so that the cycle can be repeated.
- If the coupling between the steam expanding and the mechanical motions is disrupted, the free energy of the steam is wasted.
- The valves that control the flow of steam are essential, and their function must be coupled to the movement of the piston to ensure that each one opens and closes at the correct time in the cycle.

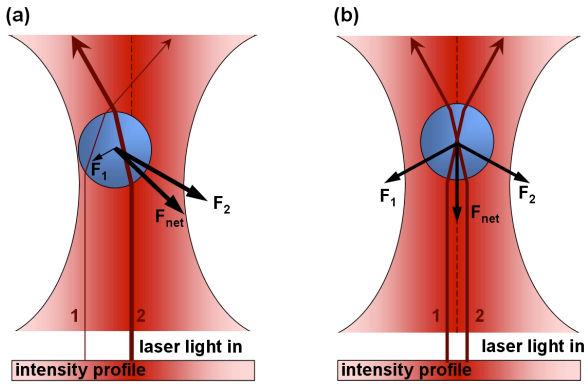
Some of these features can be found in molecular motors. Perhaps most important are the mechanisms that couple a highly favorable process with an unfavorable mechanical process. Molecular motors also function in a cyclical fashion. On the other hand, there are important differences between molecular motors and steam engines:

- There cannot be significant temperature differences at the molecular scale.
- The molecular structures have no significant inertia, so that a flywheel cannot restore the motor to its initial state.

II. Measuring forces at the molecular scale and stretching a DNA molecule

One of the technical advances that has enabled the study of molecular motors is the development of instruments capable of measuring forces at the molecular level. In the previous chapter, we briefly discussed one such instrument, the atomic force microscope (pages 189–190). Another instrument used to measure molecular forces and manipulate individual molecules is called an *optical trap* or *optical tweezers*. This type of instrument can be used to manipulate objects as small as individual atoms, but when used in studies of biological macromolecular strategies, the usual strategy is to attach the molecules to small glass or silica beads, typically about 1 μm in diameter, which are then manipulated. The optical trap is generated by focusing a beam of light into a narrow spot. When viewed from the side, this beam has an hour-glass shape with a narrow waist, as illustrated below[1]:

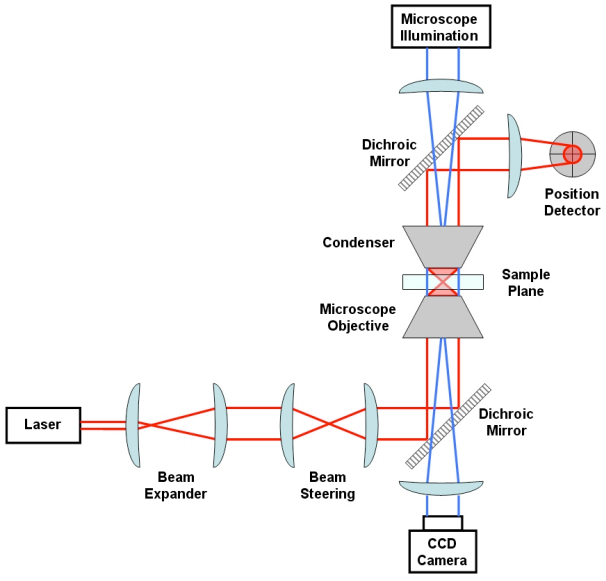
[1]Illustrations of the optical trap device are from https://en.wikipedia.org/wiki/Optical_tweezers



When a transparent bead is placed within the light beam, it acts as a small lens and refracts, or bends, the light. This effect represents a force of the bead acting on the light, and, by Newton's third law of motion, there must be an equal and opposite force acting on the bead. As shown in the diagram above, light rays entering the bead at different points generate forces in different directions, and the magnitudes of these forces are determined by the intensity of light at different postions in the beam. All of these forces, along with the gravitational force acting on the bead, are balanced when the bead is in the center of the beam and slightly above the beam waist, as shown in the right-hand side of the illustration.

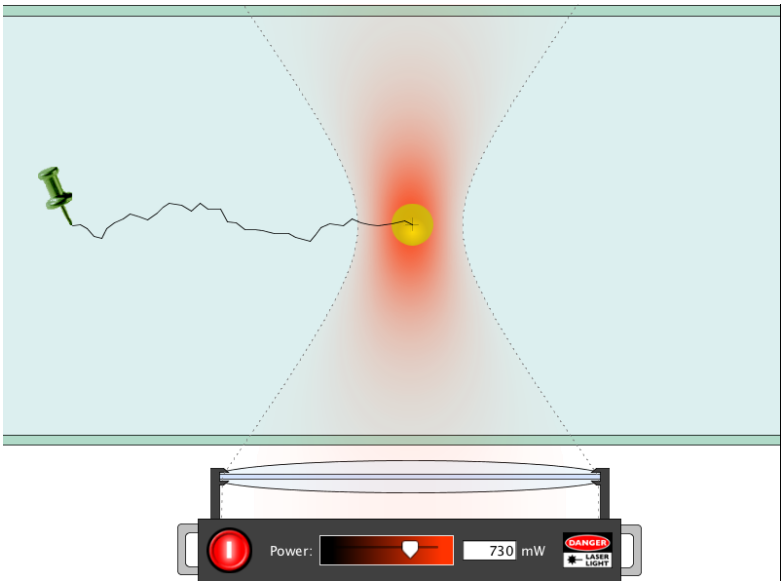
If an outside force acts on the bead to move it away from the center of the beam, the forces from the focused light will try to return the bead to the center. The magnitude of this restoring force increases as the bead is displaced from the beam center, until the outside force becomes great enough to pull the bead out of the beam altogether. The beam thus acts like a spring, and the displacement indicates the magnitude of the outside force, much like a spring scale, as found in a grocery store.

The construction of an optical trap device is quite intricate and requires very precise engineering. The schematic diagram below shows the main elements:



The apparatus is typically built around a conventional optical microscope, and the trapping beam is generated by a laser and focused by the microscope objective. In addition, there are lenses that are used to move, or steer, the beam. After passing though the sample, which contains the glass bead, the beam is focused on a position detecter, which can very precisely record the position of the beam as it is steered. In addition, illumination from the top of the microscope allows the position of the glass bead to be monitored from an image created in a camera below.

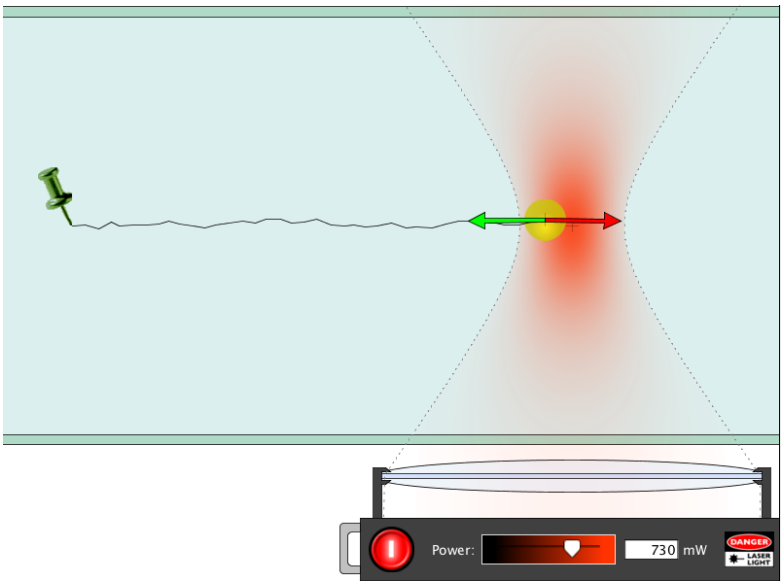
One application of optical traps has been to measure the forces exerted on DNA molecules as their ends are pulled appart. This is conceptually very similar to the protein stretching experiment described in the previous chapter. A highly schematic representation of the experimental apparatus is shown below:



This figure is a screen capture from a very clever educational simulation that allows the user to manipulate the virtual optical trap and follow the effects on the trapped bead and DNA molecule[2]. In this kind of experiment, one end of the DNA is somehow fixed to the microscope slide (as represented by the thumbtack), and the other is attached to the transparent bead. The focused laser beam is then manipulated to capture the bead and then move it about.

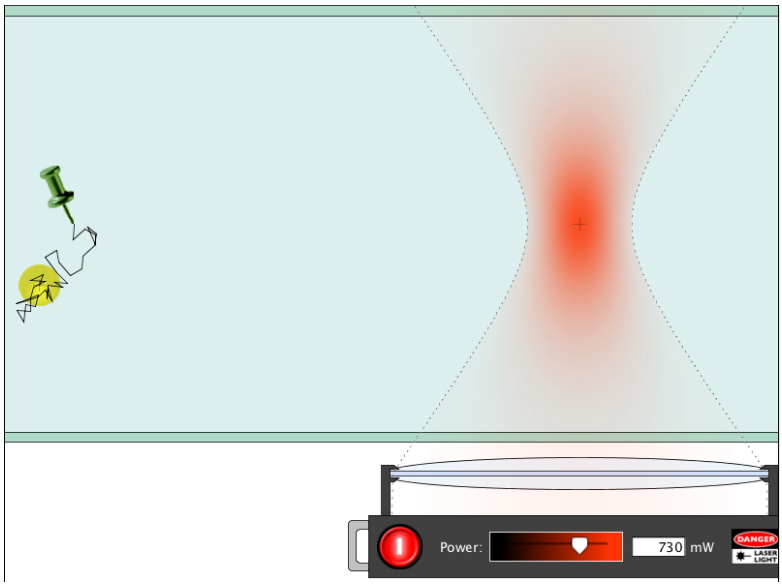
As the bead is moved away from the point at which the other DNA end is attached, the forces acting on the bead increase, as shown by the red and green arrows in the illustration below:

[2]https://phet.colorado.edu/en/simulation/legacy/stretching-dna



The green arrow represents the force exerted by the DNA, and the red arrow represents the net force generated by the optical trap as the bead is pulled away from the center of the beam.

If the bead is pulled too far, the force exerted by the DNA will exceed the maximum force of the optical trap, and the bead will break away. Rather than staying in one position, or moving randomly by brownian motion, the bead will gradually (via a biased random walk) move towards the fixed end of the DNA, as shown below:



As discussed in the context of the protein-stretching experiment, the force generated by the DNA is entropic in nature. As the ends of the molecule are separated, the number of possible conformations (microstates) is reduced and the free energy increases. Although this argument offers a thermodynamic explanation for the work, and therefore force,

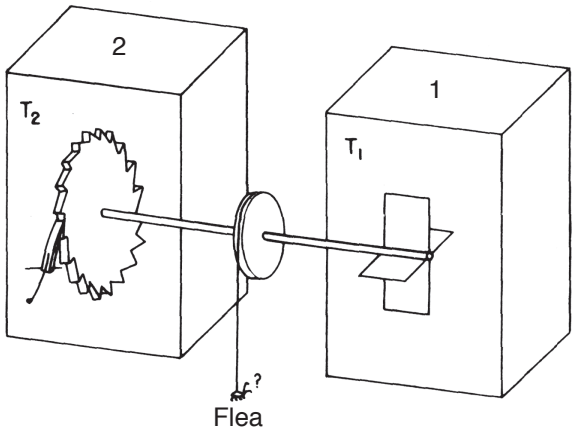
required to stretch the molecule, it is worth giving some thought to exactly what causes the force and net motion of the bead when it is released from the optical trap.

It might seem that some of the force could come from the stretching of covalent bonds in the DNA molecules. However, the forces generated in this experiment are far smaller than those required to stretch a bond by even a very small amount. The actual motions of the DNA and the bead are due to thermal motion and the collisions with surrounding water molecules, that is Brownian motion. Over time, a very large number of the possible conformations of the DNA are sampled due to these motions, and all of them have roughly equal energies. However, many more of the possible conformations are ones where the two ends are relatively close together than far apart. This is very similar to what we discussed in the context of two- and three-dimensional random walks, and exactly the same mathematics can be used to describe the entropic stretching force.

The protein- and DNA-stretching experiments demonstrate that thermal motion can generate macromolecular forces and net motion along a single direction. But, can these motions be used in a cyclic motor?

III. A Brownian ratchet and Maxwell's demon

To help understand the requirements for capturing thermal energy to produce mechanical energy, it is helpful to consider a hypothetical heat engine designed to capture the thermal energy of gas molecules. One version of this engine, sometimes called a "Brownian ratchet", was described by Richard Feynman in his classic book of physics lectures [3], as shown below:



In this device, there are two isolated compartments. In compartment 1, there is a paddle wheel surrounded by gas molecules at temperature T_1. The gas molecules randomly collide with the paddles and can, in principle, cause periodic rotations in either direction. But, the paddle is connected via a shaft to a ratchet and pawl mechanism in compartment 2, which contains a gas at temperature T_2. Because of the shape of the

[3]Figure from Feynman, R. P., Leighton, R. B. & Sands, M. (2013). *The Feynman Lectures on Physics*, volume I, chapter 46. Basic Books, new millennial edition. http://www.feynmanlectures.caltech.edu/I_46.html

teeth on the ratchet wheel, the wheel can rotate in the clockwise direction, as viewed in the drawing, but counter-clockwise rotation is blocked by the pawl, which is held in position by a spring. It would appear that this mechanism can capture the thermal energy of the gas molecules in compartment 1 to cause rotation in the clockwise direction, perhaps even doing a little bit of work, such as lifting the flea in the drawing.

One statement of the second law of thermodynamics is that work can only be obtained from thermal energy when there is a net flow of heat from a warm object to a cooler one. In the Brownian ratchet, however, it is not so obvious why work could only be obtained if T_1 is greater than T_2. Nor is it obvious how heat could flow from compartment 1 or 2. The key to both paradoxes lies in the ratchet mechanism itself. In order for the wheel to move in the clockwise direction, the force generated by the collision of the gas molecules on the paddles must be great enough to lift the pawl. However, if the temperatures of the two sides of the apparatus are equal, then this amount of thermal energy is also available to lift the pawl directly, which will allow the wheel to turn in either direction.

The notion that the ratchet mechanism could allow rotation in either direction with equal probability may still seem rather counterintuitive, since it certainly looks as though it is much harder to lift the pawl to allow counter-clockwise rotation. And, anyone who has used a wrench with a ratchet mechanism knows that a ratchet does, indeed, allow rotation in one direction but not the other. This, however, is a case in which our intuition based on the macroscopic world fails us when we move to the microscopic scale, where motions are determined by random collisions. If we examine the mechanism closely, it is apparent that the clockwise rotation involves a gradual lifting of the pawl along the slope of the teeth. Since the collisions of gas molecules on the paddle wheels cause only very small movements, moving the ratchet wheel to the next stopping point requires a series of microsteps in the same direction. Otherwise, the force of the pawl will move the wheel back to its starting point. Probabilistically, this is equivalent to flipping a coin 10 times, say, and seeing 10 heads. This is unlikely, but not impossible. On the other hand, for the pawl to be lifted by thermal energy all the way up to allow a counter-clockwise step is equivalent to many unlikely events all at once. This would be equivalent to flipping 10 coins all at once and, again, seeing 10 heads. This is unlikely, but no more so than 10 heads from 10 sequential coin flips.

The only way in which net work can be extracted by this apparatus is if the two parts, the paddle wheel and the ratchet, are isolated and the temperature of the paddle wheel chamber (T_1) is greater than that of the ratchet chamber (T_2). Under these conditions, more thermal energy is available to move the wheel than to spontaneously lift the ratchet pawl, thus favoring the forward direction. During this process, the temperature of the paddle chamber will decrease, as the gas molecules lose kinetic energy. In the ratchet chamber, the temperature will increase because each time that the ratchet pawl falls back on the wheel heat is generated. Thus, there is a net flow of heat associated with the work generated, as required by the second law. Eventually the two chambers will reach the same temperature, and no further work can be extracted.

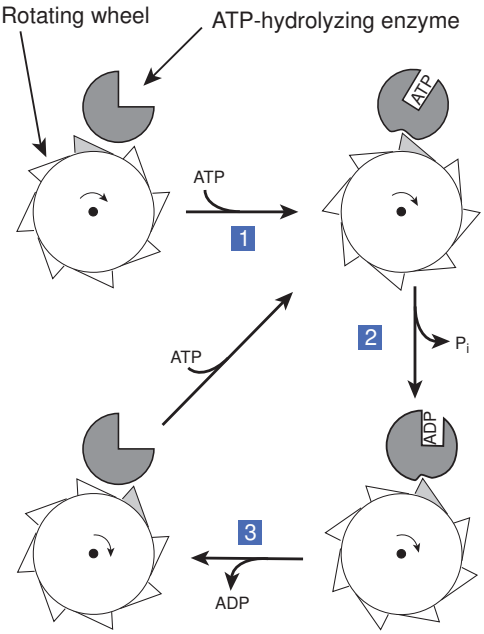
The one way in which work could be obtained with the ratchet and paddle at the

same temperature is if some "intelligent" being could monitor the direction of the small fluctuations and allow the wheel to go in one direction, but not the other. This mythical creature is usually referred to as "Maxwell's demon".

The original version of the demon, proposed by George Clerk Maxwell in 1871, was poised near an opening between two chambers containing a gas. By allowing only the faster gas molecules to move in one direction, and the slower ones in the opposite direction, the demon is able to take a system at thermal equilibrium and raise the temperature of one side and lower that of the other. The key, however, is that the demon must, herself, expend energy to create this temperature difference.

IV. A hypothetical ATPase ratchet

To consider what features might be important for an ATP-driven motor, we can design an imaginary motor in which an ATP-hydrolyzing enzyme plays the role of Maxwell's demon. One possible scheme is drawn below:



Note the following features of this device:

- The motor has two components:
  - A rotating wheel with teeth like those on a ratchet wheel.
  - An ATPase enzyme that undergoes conformational changes as it binds and hydrolyzes ATP.
- The teeth on the rotating wheel have an asymmetric structure, so that they interact with the ATPase differently depending on which direction the wheel moves.
- The ATPase has three different conformations, depending on whether ATP or ADP, or neither is bound.

In one forward cycle of the motor, the following steps take place:

1. The enzyme binds ATP and changes conformation; The wheel rotates clockwise.
2. The ATP is hydrolyzed, phosphate is released and the enzyme changes conformation; The wheel rotates clockwise.
3. ADP is released and the enzyme returns to its starting conformation; The wheel rotates clockwise.

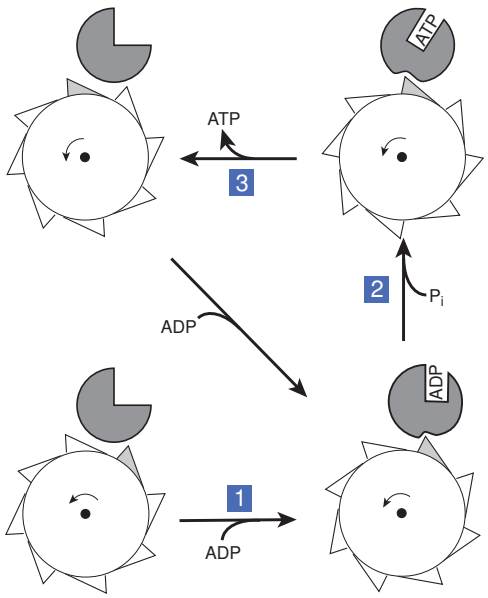
At the end of these steps, the motor is in its original state, except that the wheel has rotated by one full step from its starting position. The cycle can then begin again with the binding of another ATP molecule.

Each of the steps outline above actually includes three kinds of events:

- Nucleotide binding or release, or catalysis.
- A conformational change.
- Rotation of the wheel.

The three events could, in principle, occur in a specific order or in a concerted process. For instance, the first step might begin with the Brownian motion that brings a tooth of the wheel into contact with the ATPase, which then causes a change in the enzyme's conformation, which then favors ATP binding. Or, the process could begin with ATP binding, followed by the conformational change and rotation. Dissecting the details of such processes is an important aspect of the study of real molecular motors.

What is most important about this scheme is that there is an order to the conformational changes, and this order is determined by the order of steps in the chemical reaction. If the concentrations of ADP and P_i are much higher than that of ATP, then the reverse chemical reaction will be favored thermodynamically. Under these conditions, the open enzyme is more likely to bind ADP than ATP, leading to rotation in the counter-clockwise direction, as illustrated below:



Each step in the counter-clockwise cycle is the exact reverse of the corresponding step in the clockwise cycle, including nucleotide binding or release, catalysis, conformational change and rotation.

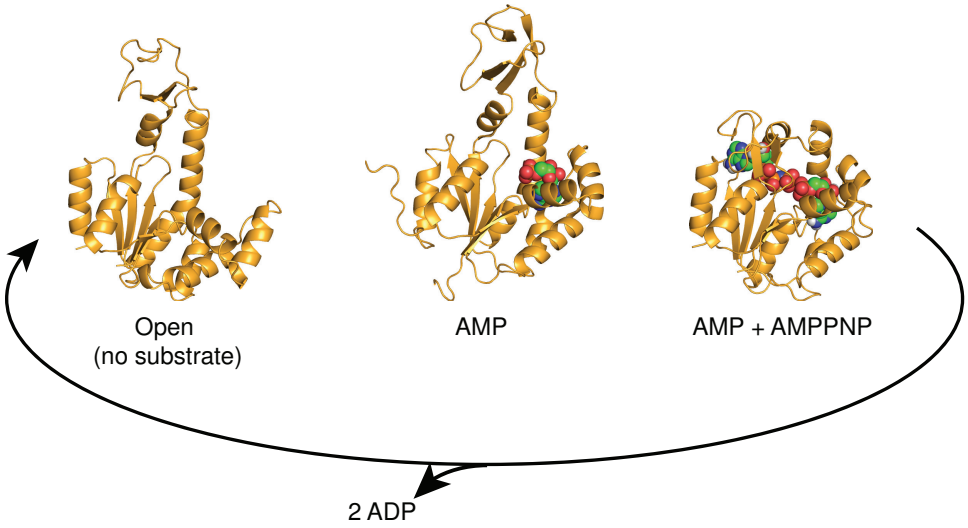
Another critical aspect of the motor is that the catalytic cycle and the rotation are tightly coupled, so that the wheel cannot rotate without the enzyme catalyzing the reaction in one direction of the other. Conversely, the conformational change of the enzyme and rotation of the wheel cannot occur in the absence of the catalytic cycle. It should then be possible for mechanical energy to be used to make ATP from ADP and P_i. This is what, indeed, happens in the ATP synthase that we will discuss later. In this case, rotation of one portion of the enzyme is driven by the movement of H^+ ions down a concentration gradient, and this mechanical motion is coupled to ATP synthesis, by a mechanism similar in principle to that suggested above.

7.2 Adenylate kinase: Coupling a chemical reaction to conformational change

Before discussing some real molecular motors, we will consider an enzyme that illustrates how a cyclical structural change can be coupled to a chemical reaction. The reaction catalyzed by adenylate kinase is the transfer of a terminal phosphate group from one adenosyl nucleotide to another:

$$\mathrm{ATP} + \mathrm{AMP} \rightleftharpoons \mathrm{ADP} + \mathrm{ADP}$$

This reaction has an equilibrium constant close to 1. The major physiological role of adenylate kinase is in muscle tissue, where the reverse of the reaction drawn above regenerates a reserve of ATP, from ADP, when ATP levels are depleted. The structure of the enzyme has a cleft in which the substrates bind, and the protein undergoes large structural changes upon substrate binding, as illustrated in the drawings shown below:



These drawings are based on crystal structures of adenylate kinase with no substrate bound (PDB entry 4AKE), AMP bound (PDB entry 2AK3) and AMP and a non-hydrolyzable ATP analog (AMPPNP) bound (PDB entry 1ANK). The non-hydrolyzable substrate was used to trap the enzyme in the conformation presumed to exist when it is about to carry out the catalytic reaction.

As illustrated above, the cleft closes tightly around the substrates, and it is believed that the major role of this structural change is to exclude water, which otherwise could act to hydrolyze the ATP. Once the chemical reaction has taken place, in either direction, the cleft must open up again to allow release of the products. One could imagine how this motion might, if connected to other structures, be used to carry out mechanical work, and it turns out that the structural motif found in adenylate kinase is present in several ATP-driven molecular motors, including myosin, kinesin and ATP synthase.

Consideration of adenylate kinase also raises a bit of a paradox. Suppose that ATP, ADP and AMP were all present, along with the enzyme, and the reaction had reached equilibrium. Under these conditions, the free energy change for the reaction is zero, and no work should be available. However, the enzyme will continue to catalyze the reaction in both directions, undergoing a continuous opening and closing process. It seems as though it is generating work without consuming energy. This would be equivalent to the Brownian ratchet performing work when the two compartments are at the same temperature, which we have argued violates the second law of thermodynamics. However, as soon as we try to couple the enzyme to some other structure, in the hopes of producing mechanical work, the direction of the reaction will be biased in one direction or the other. Only if the reactant and product concentrations favor the opposite reaction will we be able to produce motion against the force acting on the enzyme.

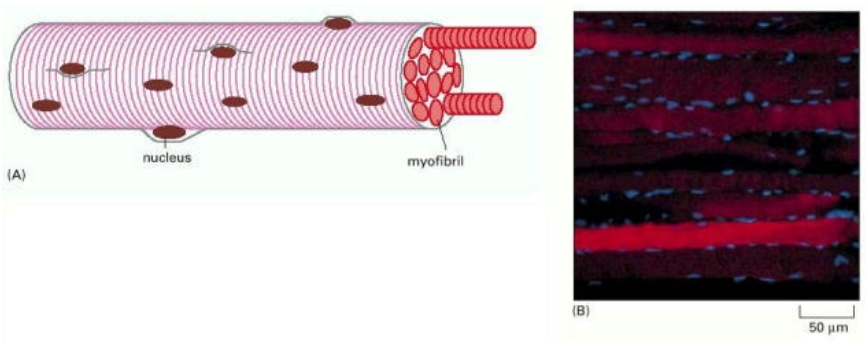
7.3 Myosin and Muscle Contraction

The molecular motors with which we have the most readily visible experience are found in our muscle cells. The primary components of these motors are two proteins, myosin and actin, which interact with one another and link ATP hydrolysis to a relative motion of one with respect to the other. Before considering the molecular details of these motors, it is useful to describe the larger structures in which they are found.

I. The structure of muscle fibers

The force-generating cells of muscles are called *myocytes* and extend the full length of the muscle. During development, these cells are generated by the fusion of multiple precursor cells, *myoblasts*, and the nuclei of the individual precursors are retained in the myocytes, as illustrated in the figures below[4]:

[4]Figure from Alberts B, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26888/#A3065



As shown in the diagram in Panel A, the nuclei are located at the periphery of the cell, and the central part of the cell is filled with roughly cylindrical structures, which extend along the full length of the cell (and the muscle), called *myofibrils*. Panel B shows a fluorescent micrograph of muscle tissue stained with a blue-fluorescent dye that binds to DNA and highlights the nuclei.

The figure below shows several parallel myofibrils as visualized by electron microscopy at relatively low resolution[5]:

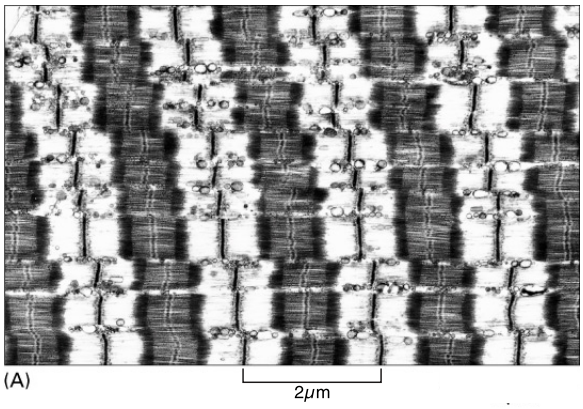


Figure 16–69 part 1 of 2. Molecular Biology of the Cell, 4th Edition.

The sample used for this micrograph was stained with a heavy-metal compound that reacts with protein molecules, so that the darker regions represent the most protein-dense regions. The proteins are organized into repeating bands of high and low density. These repeating structures are called *sarcomeres* and have a length of about $2\,\mu$m in this sample.

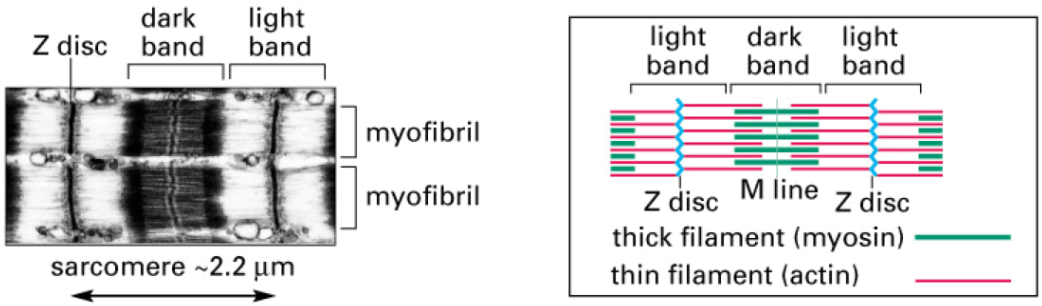
The protein composition of myofibrils was extensively studied in the 1930s and 40s, particularly by Albert Szent-Györgyi and his colleagues at the the University of Szeged, Hungary. These studies identified the two major protein components of muscle cells:

- Actin, with a molecular mass of 42000 Da.
- Myosin, a much larger protein, with a molecular mass of about 500000 Da.

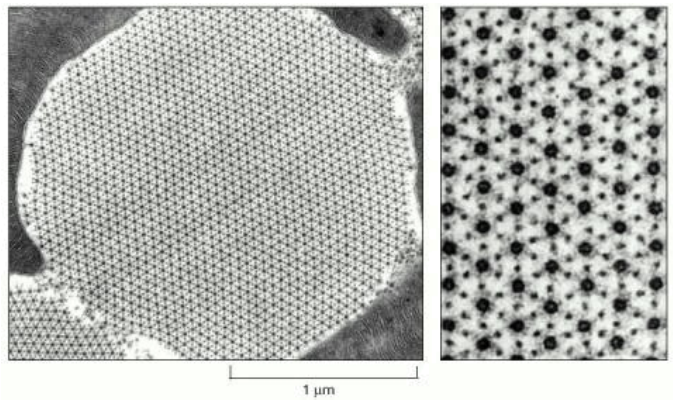
[5]Figure from Alberts B, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26888/#A3065

Actin can be isolated as a monomer, but readily assembles into long, thin fibers. Myosin can also be isolated in a soluble form, but assembles in to filaments that are much thicker than those formed by actin. Myosin was also discovered to have an ATP hydrolyzing activity that is stimulated by the presence of actin filaments.

Further studies, in the 1950s, established that the darkly staining regions of the sarcomeres contained predominantly the thick filaments formed by myosin, and that the lightly staining regions contained actin thin filaments. The figure below shows, in an electron micrograph and a diagram, the organization of the sarcomere and the locations of the thick and thin filaments[6]:



A better sense of how the filaments are arranged in three dimensions can be gained from a cross-sectional view of a myofibril, as shown below for an insect flight muscle:



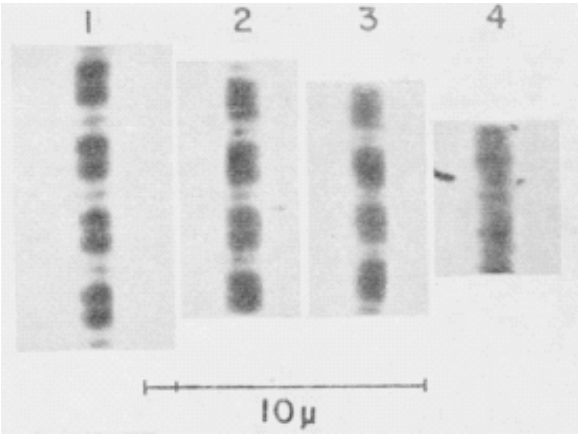
The myofibrils of insect flight muscles are more highly organized than most, with the thick and thin fibers in overlapping hexagonal arrays, but the same interdigitation of the two types of fibers is found in all muscle cells.

The year 1954 brought a major advance in the understanding of muscle contraction, based on studies by two groups, who published their findings in side-by-side articles[7]

[6]Both figures on this page are from Alberts B, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26888/#A3065

[7]Huxley, A. & Niedergerke, R. (1954). Structural changes in muscle during contraction: Interference microscopy of living muscle fibres. *Nature*, 173, 971–973. http://dx.doi.org/10.1038/173971a0 and

Both groups used special forms of light microscopy to enhance the contrast of images of myofibrils, allowing the alternating patterns of high and low protein density to be visualized as the fibrils underwent contraction. A set of images from the paper by Huxley and Hanson is shown below:



As in the electron micrographs shown earlier, the dark bands in these images represent the protein rich regions containing the myosin thick filaments, and the light bands contain primarily actin. The sequence of images, labeled 1 through 4, represent the same four sarcomeres during contraction induced by the addition of ATP. The important observation, reported by both groups, was that the light bands became progressively shorter during contraction, while the dark bands remained largely unchanged. Importantly, there was no thickening of the fibrils during contraction, as might have been expected from the familiar bulging of muscles when they contract.

From these observations, both pairs of authors proposed what came to be known as the *sliding-filament* model of contraction, which is diagrammed below[8]:

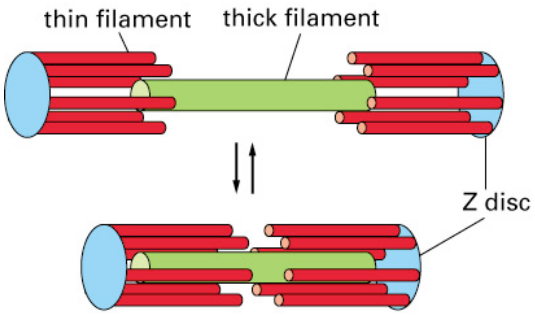


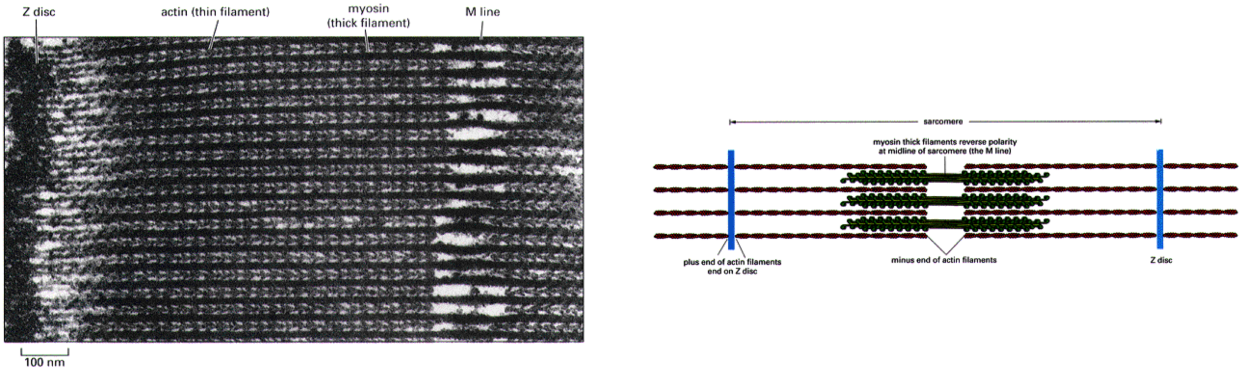
Figure 16–71. Molecular Biology of the Cell, 4th Edition.

Huxley, H. & Hanson, J. (1954). Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. *Nature*, 173, 973–976. http://dx.doi.org/10.1038/173973a0 (The two Huxleys are unrelated, but both made major contributions to physiology in the 20th century.)

[8]Figure from Alberts B, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26888/#A3065

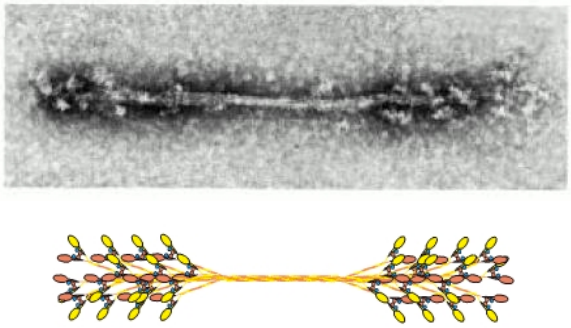
The major feature of the model is that contraction is based on the sliding of the thin filaments past the thick filaments, thus accounting for the shortening of the light regions in the micrographs. In addition, the actin filaments are attached the densely staining structures, called *Z disks*, that separate the sarcomeres, thus linking the sarcomeres together.

Following the 1954 studies, the major question to be resolved was what generates the force that moves the filaments past each other. Again, a major source of insight was electron microscopy. In the figure below[9] , the left-hand panel shows myofibrils stained in a way so that the proteins appear light against a dark background (negative stain).



A schematic interpretation of the electron micrograph is shown on the right. A prominent feature visible in the micrograph is the array of globular protrusions located between the thick and thin filaments. If you look closely, you will see that these protrusions are angled and that they point in opposite directions on the two sides of the region identified as the M line.

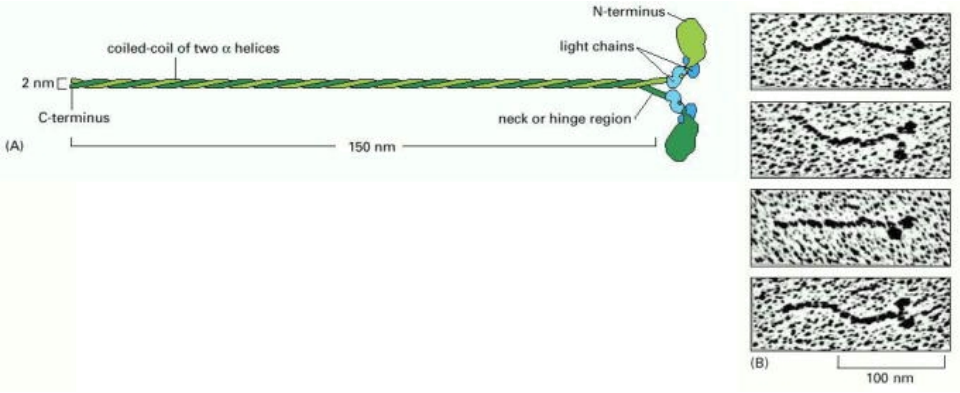
Based on micrographs like the one above, Hugh Huxley proposed that these cross-bridges were the location of ATP hydrolysis and force generation. This interpretation was supported by electron micrographs of isolated thick filaments, as shown below[10], as well as biochemical experiments.



[9]This figure and the one at the top of the following page are from Aberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994). *Molecular Biology of the Cell*. Garland Publishing, 3^{rd} edition.

[10]Figure from Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. https://www.ncbi.nlm.nih.gov/books/NBK22418/

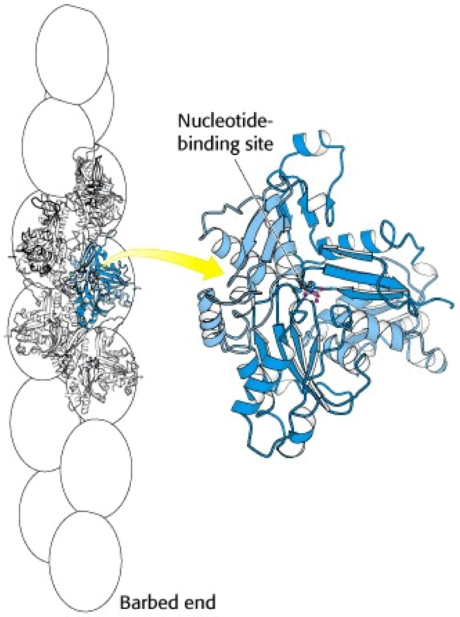
The micrograph of the isolated filament shows that the cross-bridges are restricted to the outer thirds of the filament and that the middle third is relatively thin. The cartoon representation below the micrograph reflects the structures of individual myosin molecules, which are shown in more detail in the next figure.



Each myosin molecule contains two very large and identical polypeptide chains (called the heavy chains and composed of about 2,000 amino acid residues each) and four smaller polypeptides, called light chains. The C-terminal portions of the heavy chains form α-helices, and the helices of the two chains wrap around one another to form a structure known as a coiled coil. The coiled-coil regions of individual myosin molecules assemble further to form the thick filaments. The N-terminal regions of the heavy chains fold into a globular structure. The coiled-coil tails of the molecules are linked to the globular heads by a hinge region that includes two light chains bound to each heavy chain. The ATPase activity is located in the globular heads.

The thin filaments, composed of actin, have a structure very different from that of the thick filaments. Actin folds into single globular domain and assembles into filaments as illustrated below[11]

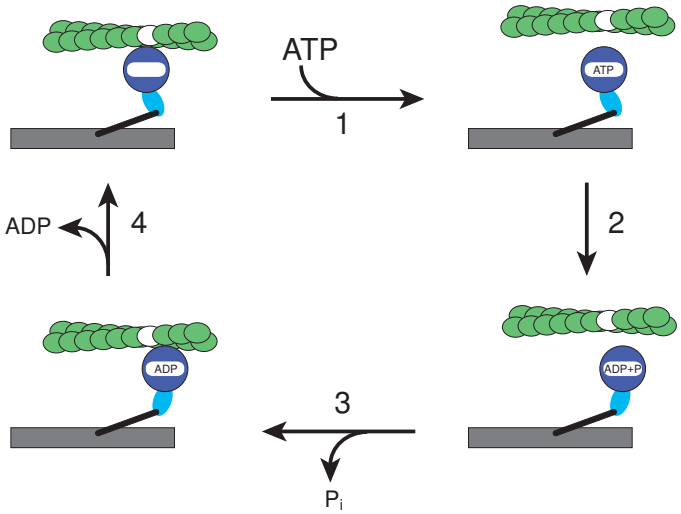
[11]Figure from Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. https://www.ncbi.nlm.nih.gov/books/NBK22418/



Interestingly, the actin monomers contain a nucleotide binding site, but do not posess catalytic activity. Nucleotide binding helps to control the polymerization of actin, but does not appear to play any role in force generation.

II. The ATPase cross-bridge cycle

In addition to structural studies, our understanding of the mechanism of muscle contraction (and other molecular motors) comes from extensive biochemical studies, including kinetic experiments in which the rates of the catalytic reaction have been studied using a range of substrate concentrations and in the presence or absence of actin. Work by Edward Taylor and his colleagues[12], in particular, along with subsequent work, has led the model diagrammed below:



[12]Lymn, R. W. & Taylor, E. W. (1971). Mechanism of adenosine triphosphate hydrolysis by actomyosin. *Biochemistry*, 10, 4617–4624.
http://dx.doi.org/10.1021/bi00801a004

In this figure, one of the actin monomers is left uncolored to highlight the movement of the filament with respect to the myosin during the cycle, which consists of the following steps:

1. Binding of ATP to the enzyme, which is initially bound to the actin filament. Binding of ATP is coupled to the release of the actin by myosin.

2. Hydrolysis of ATP, which is coupled to a conformational change in the myosin and a change in the position of the myosin head with respect to the thin filament and the rest of the thick filament.

3. Release of the inorganic phosphate, which is coupled to rebinding of the myosin head to the actin filament.

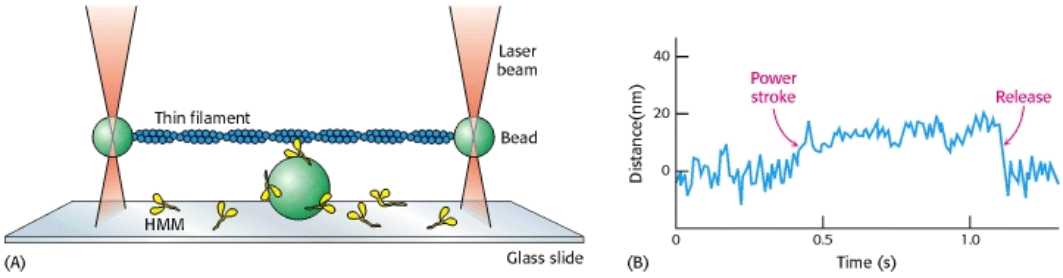
4. Release of ADP, which is coupled to a conformational change in the myosin, which leads to a shift in the relative positions of the two filaments.

The last step described above is called the powerstroke and is the point in the cycle in which motion and force are generated. Note also the critical role of the conformational change associated with the ATP hydrolysis step. Although no force is generated in this step, the conformational change positions the myosin head so that when it rebinds it is most likely to do so at a position further along the thin filament. As in the hypothetical ATPase ratchet discussed on pages 202–204, the catalytic cycle is coupled to both conformational changes of the ATPase and binding and release of the second component of the motor, in this case the actin filament. The actual motions produced by myosin and actin arise from thermal Brownian motions of the molecules. But, the motion is biased by the conformational change which increases the probability that a head will rebind further along the filament.

Contraction of a myofibril requires the action of millions of myosin heads, and one could imagine that there is some mechanism to coordinate them all. However, the myosin heads all act independently. This independence requires a limited degree of elasticity in the filaments, which comes primarily from the connections between the myosin heads and the core of the thick filament. The elasticity allows the fibril to continue contracting as individual myosin heads move through their catalytic cycles independently. Each myosin head only spends about 5% of the time bound to actin.

The action of individual myosin molecules can be studied using the optical trapping method described on pages 196–198. One such experiment is illustrated below[13]:

[13]This figure is from Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. https://www.ncbi.nlm.nih.gov/books/NBK22418/ and is adapted from:
Finer, J. T., Simmons, R. M. & Spudich, J. A. (1994). Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature*, 368, 113–119. http://dx.doi.org/10.1038/368113a0



In this experiment glass beads were attached to both ends of a thin filament and each was held in an optical trap. A fragment of myosin, containing the head domains and enough of the stalks to hold the two heads together, was attached to the surface of another glass bead, so as to hold it in place above the surface of the glass slide. The optical traps were then steered to move the actin filament into the proximity of the bead with myosin heads and the actin and myosin bound to each other in the absence of ADP or ATP. When ATP was added, the myosin underwent its catalytic cycle and exerted force on the actin filaments. During this process, the displacement of the glass beads attached to the actin was monitored, to generate plots such as shown in the right-hand side of the figure, which represents a single cycle.

From analysis of the distance versus time plots, the investigators estimated that the average step size was about 10 nm and the average force was 3–4 pN. (1 piconewton = 10^{-12} N). From these values, the work done by the motor is:

$$\begin{aligned} w &= \int F dx \approx F_{\text{ave}} \times \text{distance} \\ &= 3\,\text{pN} \times 10\,\text{nm} = 30\,\text{pN}\cdot\text{nm} \\ &= 3 \times 10^{-12}\,\text{N} \times 10^{-8}\,\text{m} = 3 \times 10^{-20}\,\text{N}\cdot\text{m} \\ &= 3 \times 10^{-20}\,\text{J} \end{aligned}$$

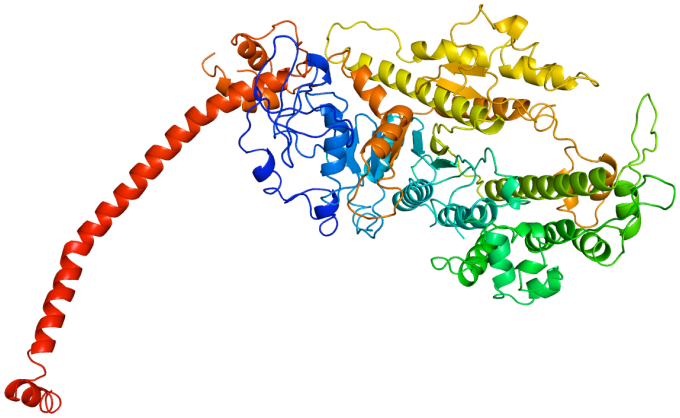
It is instructive to compare this amount of work with the standard free energy change for hydrolysis of a single molecule of ATP:

$$\begin{aligned} \Delta G^\circ &= -30 \times 10^3\,\text{J/mol} \div 6.02 \times 10^{23}\,\text{molecules/mol} \\ &= 5 \times 10^{-20}\,\text{J/molecule} \end{aligned}$$

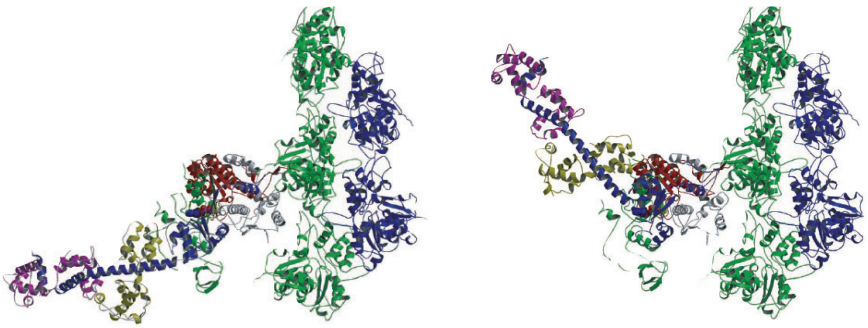
This would appear to be remarkably consistent with the work generated by the myosin, which is expected to be at least somewhat less than the theoretical energy available from ATP hydrolysis. It is important to note, however, that ΔG° represents the free energy change when the reactant and products are all present at 1 M concentrations. In the single-molecule experiment, only ATP was added to the reactions, at concentrations of 1 μM to 2 mM. From this information, it is difficult to estimate the reaction quotient, Q, or the actual free energy change, ΔG, under the experimental conditions. But, forces on the order of a few piconewtons and distances of a few nanometers appear to be common feature of molecular motors, leading to the use of the pN $\cdot$ nm as a common unit of work and energy in the field of molecular motors.

III. Atomic resolution structures of myosin and actin

In the early 1990s, a high-resolution structure of a myosin head and stalk region was determined by x-ray crystallography, providing a much more detailed view of the mechanism of muscle contraction[14]. A ribbon diagram of this structure is shown below:



As can be seen, the structure confirms the overall structure deduced earlier from electron microscopy. At about the same time that this structure was determined, the structure of actin monomers was also determined by x-ray crystallography and was used, in conjunction with EM images, to construct models of the thin filament, as shown in the figure on page 211. These structures were used to construct models of the actin-myosin complexes in different states of the catalytic cycle. The figures below represent models of the complex before and after the powerstroke[15].



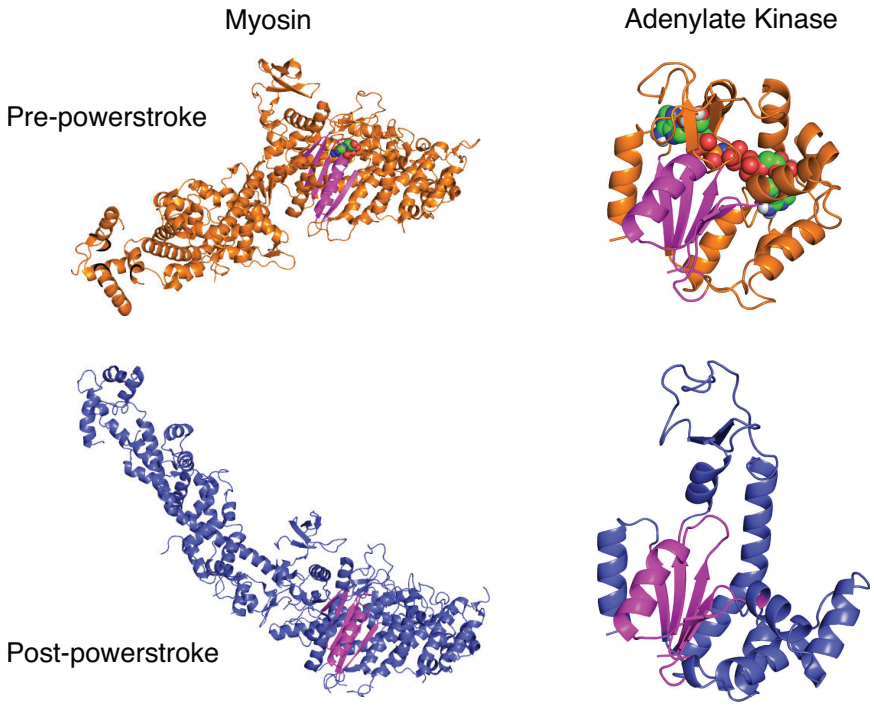
As can be seen in the figure, the position of the myosin head remains essentially fixed relative to the actin filament during the powerstroke, but the orientation of the head with respect to the myosin stalk changes dramatically. If the other end of the stalk

[14]Rayment, I., Rypniewski, W. R., Schmidt-Base, K., Smith, R., Tomchick, D. R., Benning, M. M., Winkelmann, D. A., Wesenberg, G. & Holden, H. M. (1993). Three-dimensional structure of myosin subfragment-1: a molecular motor. *Science*, 261, 50–58. http://dx.doi.org/10.1126/science.8316857

[15]Figure from Geeves, M. A. & Holmes, K. C. (1999). Structural mechanism of muscle contraction. *Ann. Rev. Biochem.*, 68, 687–728. http://dx.doi.org/10.1146/annurev.biochem.68.1.687

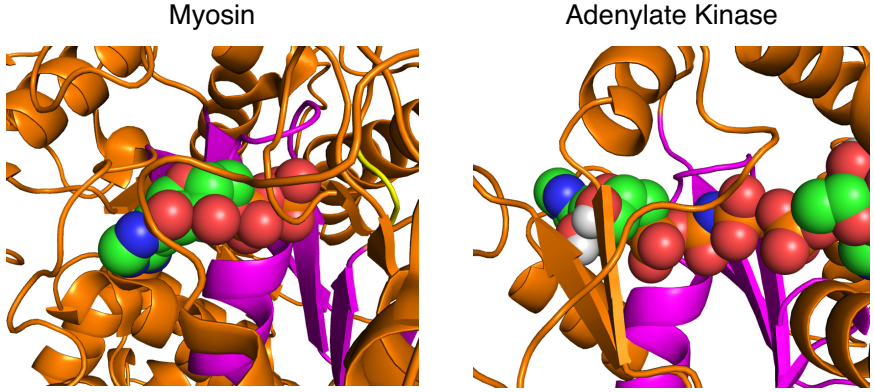
were fixed, as in an intact thick filament, the conformational change would lead to a motion of the thin filament with respect to the thick one.

Earlier, we considered the conformational changes of adenylate kinase as an example of how a catalytic cycle can be coupled to molecular movement. In fact, the relationship between adenylate kinase and myosin is more than just a formal one. The two proteins appear to be evolutionarily related and show significant conservation, particularly in the nucleotide binding site. The ribbon diagrams below highlight this relationship:



In all four drawings, the nucleotide binding region is colored magenta, and the molecules are oriented to emphasize the similarity of the structures. In both proteins, the conformational changes are centered at and near the binding sites and are propagated over long distances, though the nature of the propagated motions are different.

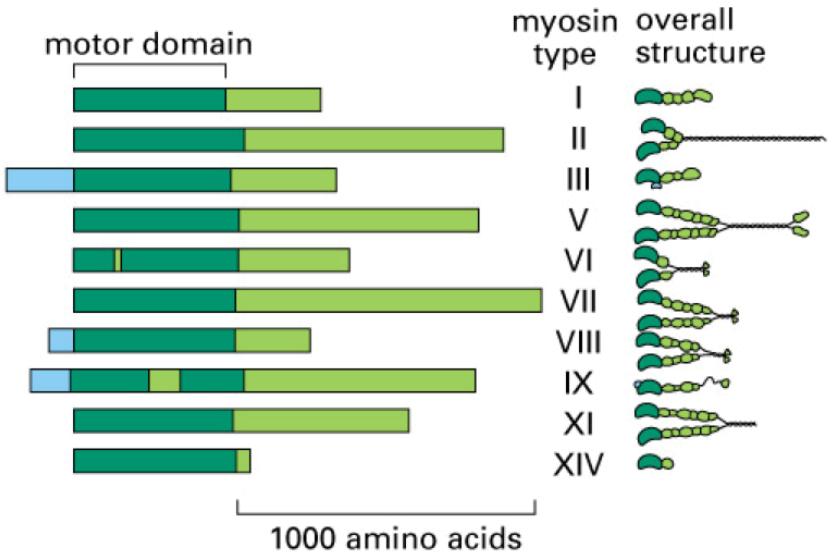
The similarity between the binding sites is further illustrated in the close-up drawings below:



The drawing on the left represents myosin in the pre-powerstroke state, with ADP bound, and the one on the right is of adenylate kinase with AMP and an ATP analog bound. The similarity of these structures, and the ubiquitous occurrence of adenylate kinase (in eukaryotes, bacteria and archaea), indicates that the structural motif at the core of myosin is ancient and has evolved to carry out very different functions.

IV. Non-muscle myosins

Although the form of myosin found in skeletal muscle cells (called type II) is probably the best known and best studied, it is only one of a large family of myosins, which display a wide range of structural organization and carry out a variety functions. The figure below diagrams the sequence relationships among the heavy chains of the major myosin types[16]:

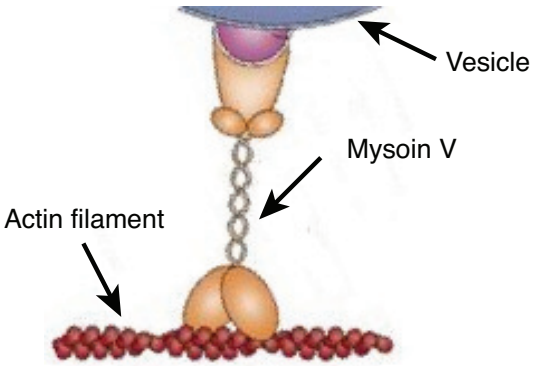


In all of these myosin types, the motor domain (in the N-terminal portion of the heavy chain) is highly conserved, but there is great variety in the C-terminal regions. Only type II myosin possesses the extended coiled-coil structure that leads to assembly into muscle thick filaments. In most, but not all, of the others, the C-terminal region leads to dimerization via a coiled-coil, but the coiled-coil regions are shorter than in type II myosin, and other domains are present in the other types.

One of the better studied of the non-muscle myosins is the type V form. This myosin functions to move intracellular cargoes, such as vesicles along actin filaments, as illustrated in the diagram below[17]

[16]Figure from Alberts B, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. https://www.ncbi.nlm.nih.gov/books/NBK26888/#A3065

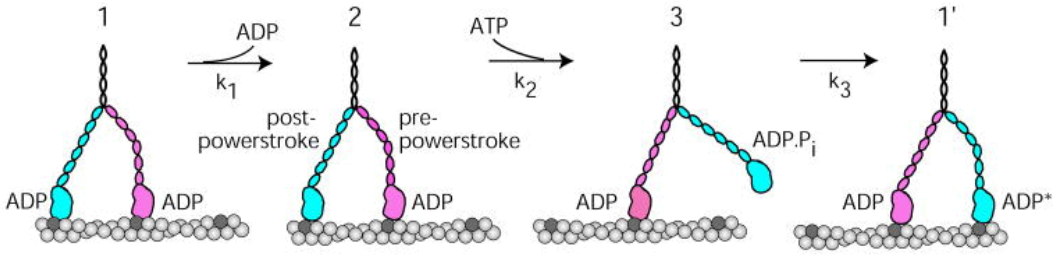
[17]Figure adapted from Soldati, T. & Schliwa, M. (2006). Powering membrane traffic in endocytosis and recycling. *Nature Rev. Mol. Cell. Biol.*, 7, 897–908. http://dx.doi.org/10.1038/nrm2060



As shown in the diagram, myosin V interacts with the actin filament via its two head domains, just as myosin II does in muscle cells, and the two myosin heavy chains are dimerized via a coiled-coil. However, the C-terminal region of the heavy chain has a specialized function to interact with proteins linked to the transport vesicle.

As discussed earlier, the heads of myosin II are bound to the thin filaments only about 5% of the time during muscle contraction. This is described as a low *duty cycle* and is a requirement to minimize the potential interference among the thousands of myosin molecules in a sarcomere. For a myosin molecule engaged in vesicle transport, however, a light duty cycle would cause the vesicle to fall off of the actin filament. In this context, it is essential that one or both of the myosin heads be bound to the actin almost all of the time, resulting in a high duty cycle.

A model for the myosin V mechanism for moving along actin filaments is diagrammes=d below[18]:



Each of the myosin heads undergoes the ATPase catalytic cycle, which is coupled to conformational changes and binding to and release from the actin filament. In this representation, the stalk of the myosin head in the pre-powerstroke state (with ADP bound) is bent, whereas the stalk is straight after the powerstroke. The powerstroke leads to the movement of the other myosin head along the filament in a hand-over-hand fashion. As in the muscle cross-bridge cycle, ATP binding is required for release of the myosin heads after the powerstroke.

In contrast to myosin II, the catalytic cycles of the two myosin V heads must be tightly coordinated in order to ensure that the complex does not fall of the filament and that

[18]Trybus, K. M. (2008). Myosin V from head to tail. *Cell. Mol. Life Sci.*, 65, 1378–1389. https://dx.doi.org/10.1007%2Fs00018-008-7507-6

the motion proceeds directionally. In particular, the powerstroke of the leading head is coordinated with the ATP binding, filament release and forward motion of the trailing head. The detailed molecular mechanism of this coordination remains to be elucidated.