1

Molecules

It is impossible to pack a complete biochemistry course into a single introductory chapter. Some of the
basic properties of the structure of simple biological macromolecules will therefore be covered. The aim of
this chapter is to give the reader a basic grounding in the rich variety of molecules that life presents and some
respect for the extreme complexity of the chemistry of biological molecules in a wide range of cellular
processes.

Cells are predominantly composed of water that is structured and organised by inorganic ions and carbon
containing (organic) molecules. The extracellular matrix in organisms can, in addition, contain solid crystalline
composites such as calcium carbonate and silicates that form bones and exoskeletons. However most of the
processes vital for life occur in aqueous solutions, although they are typically highly congested, with a huge
variety of competing molecular nanoparticles present. How robust well-regulated living processes occur in
such congested environments is still a matter of ongoing research.

There are four main classes of organic macromolecules inside cells: these are the lipids, proteins,
carbohydrates and nucleic acids. Also, mixtures are possible such as glycolipids (carbohydrates fused to
lipids) and glycoproteins (carbohydrates fused to proteins).

The subject of the molecular structure of materials will be first approached and how this arises from
the underlying quantum mechanics. Then, the concept of chirality will be introduced for molecules, cells
and organisms. Finally, a rapid tour will be made of the main classes of biological molecules that occur in
living cells.

1.1 Chemical Bonds and Molecular Interactions

Ernest Rutherford described the atom as a minute nucleus surrounded by a huge expansive cloud of electrons;
‘a mosquito in a cathedral’. Niels Bohr extended this picture, since only certain energy states are permitted
by the quantum principle of Max Planck (1900) and de Broglie’s wave/particle duality. Electrons are confined
to discrete shells as they orbit the nucleus (Figure 1.1). Subsequently, Erwin Schrödinger showed how to
calculate the energies and the spatial distributions of these electronic orbitals.
Exact calculations of electron wave functions that use the Schrödinger equation will be left to more specialised quantum mechanics courses, as too will be more accurate quantum electrodynamics calculations, which are currently the most accurate theories to simulate the behaviour of atoms and molecules. Both approaches with complex biomolecules tend to be horrendously difficult and require extensive computational power. However, a few more details of quantum mechanics are useful for the development of an intuitive picture of molecular structure and dynamics, and will thus be introduced here.

Max Born postulated a radical reinterpretation of the quantum theory; different quantum states of electron distributions in the vicinity of an atomic nucleus are characterised by different probability distributions. These probability distributions are the fundamental quantities predicted by theory, and intrinsic uncertainties in their values are written in to the laws of nature (Heisenberg’s uncertainty principle). The Schrödinger equation predicts the probability that an electron is found at a certain position. The full information about a bound electron inside an atom is now thought to be described by four quantum numbers; \( n \) (the principle quantum number related to the energy), \( l \) (the total angular momentum), \( m \) (the \( z \) component of the angular momentum) and \( s \) (the angular spin momentum). Based on these four quantum numbers, combined with the symmetry of the electrons’ wave functions, Pauli deduced his exclusion principle, which is that ‘no two electrons can be associated with the same nucleus and have precisely the same values of all four of the quantum numbers’. This principle stops matter collapsing in on itself, since electrons repel each other to form shells such that each electron has a different permutation of the four quantum numbers. This stability is reflected in the repulsive hard-core interactions experienced by all atoms at short distances, i.e. the excluded volume potential.

The quantum theory also explains the formation of molecules. Neighbouring atoms share or transfer electrons to create more energetically stable quantised electronic orbital structures. The geometries of atomic orbitals are classified as \( s \) (spherical), \( p \) (double lobed), \( d \) (double lobe threading a doughnut) and \( f \) (double lobe threading two doughnuts), Figure 1.2. Molecular orbitals tend to be even more complicated, since they require hybridisation of neighbouring atomic orbitals.

The prediction of chemical bonding patterns tends to be a job for the intuition of a good synthetic chemist with many years of experience, or an extremely hard \textit{ab initio} quantum mechanical calculation for a computer. Chemical bonding can be classified under the broad (and often overlapping) headings of \textit{ionic}, \textit{covalent}, \textit{metallic} and \textit{hydrogen}.
Ionic bonding occurs when atoms ionise to form electrolytes with the classic example being table salt, NaCl. An ionisation energy is associated with the movement of an electron from one atom to another, so their electronic structures are more energetically favourable.

When the difference in electronegativity is small between two atoms, they may form molecules through covalent bonding, e.g. H₂, HF, H₂O or metallic bonding, e.g. Na, K, Fe. Covalent bonding requires electrons to be shared and they are closely localised between the two bonded atoms. Metallic bonding involves delocalised electronic wave functions that allow rapid mobility of electrons through the crystalline lattices.

The development of a good intuition for the relationship between molecular structures and chemical formulae requires years of experience, but it is instructive to ask a simple question on tetrahedral bonding to start to build some understanding. It is useful to ask why the bonding pattern in methane (CH₄) is symmetrical while water (H₂O) is not (Figure 1.3). The answer is found in the quantum mechanics, because carbon has two fewer electrons than oxygen. In water 2s and 2p orbitals of oxygen hybridise (four orbitals in total); two electrons bond with two Hs, and the other two electrons become lone pairs. The lone pairs are negatively charged and attract the H atoms from neighbouring water molecules (hydrogen bonding). Water thus has a distorted tetrahedral structure due to the different interactions between the lone pairs and the hybridised orbitals. In contrast methane’s carbon atom forms covalent bonds with four hydrogen atoms in a perfectly symmetrical tetrahedral structure.
Molecules assemble together to form different phases of matter (gas, liquid, solid, etc.), but retain their individual identities at the atomic scale. The exact phase adopted is determined by intermolecular forces between the molecules and they are weaker than the intramolecular forces already considered (ionic bonding, covalent bonding, etc.).

The most common intermolecular force is that of van der Waals. This can be thought of as a default interaction that occurs between all molecules and dominates the interactions if all the other forces are switched off, e.g. with liquid helium at low temperatures. In fact van der Waals forces correspond to a family of three or more types of force. Debye, Keesom, and van der Waals interactions are the principle subclassifications, but higher-order multipolar interactions also occur (Chapter 4). In general, van der Waals forces arise from instantaneous stochastic dipole moments associated with the motions of individual electrons. The instantaneous distribution of electrons can influence those of surrounding atoms, and the net outcome can be a weakly positive attraction. In biology, this force can explain the miraculous manner in which geckos can crawl up windows and flies can stick to ceilings. It is, however, not a universal mechanism of adhesion, e.g. tropical frogs use capillary forces to hold themselves onto surfaces, which in turn depends on the humidity.

Unlike covalent bonding, the van der Waals force does not have a unique direction, it cannot be saturated (the number of atoms or molecules involved is based on geometrical conditions, not on the electronic structure of the orbitals), it depends on the sample geometry and requires quantum electrodynamics (QED) calculations for a quantitative treatment. Although the derivations are complicated, analytic QED solutions do exist for standard geometries, e.g. see Adrian Parsegian’s book, ‘van der Waals forces in biology’, for more details (also Chapter 4).

The importance of hydrogen bonding in biology cannot be more emphasised (Figure 1.4, Section 1.7). All living organisms contain a large amount of water, and water has its unique properties due to hydrogen bonding. For example, water expands when it freezes, which helped life’s origins in the sea immeasurably. Water is also important for the self-assembly of many biomolecules. Indeed from a certain perspective, humans are predominantly structured self-assembled water. There is a nice analogy with a wobbly children’s party jelly that is 98% water and 2% protein, but appears (at the low frequencies people are familiar with) to be a solid. The water component in human cells is slightly lower at ~70% w/w, but the jelly analogy helps the development
of an intuition of the cell’s varied dynamic properties at different time scales (its viscoelasticity). At long time scales cells appear to be solid, whereas at short time scales they are liquid-like.

Hydrogen bonding in water has a number of subtle secondary interactions that are important in the determination of its effects. One phenomenon is the hydrophobic effect that is primarily due to entropy. Consider a fat chain surrounded by water molecules. The normal hydrogen-bonded network in liquid water is disrupted and the water molecules become structured due to the loss of mobility as they orient themselves away from the fat chains. There is consequently a penalty in the free-energy term due to the loss of entropy (\(-TS\) in \(F = U - TS\), the Helmholtz free energy, Section 3.4). Entropy will be considered in more detail in Chapter 3. In a hand-waving manner, entropy is a measure of the randomness of a system, or equivalently the number of accessible states to a system. The hydrophobic effect is the molecular origin of why oil and water do not mix on the macroscale.

In general, all the forces between different biomolecules, subcellular compartments, cells, organisms and nanoscaled particles are of interest to biological physics. These interactions lead to relatively weak forces compared to chemical bonds, but help explain important physical and biological phenomena such as phase transitions and self-assembly. Such coarse-grained interactions over nanometer length scales are called mesoscopic forces, and will be considered in more detail in Chapter 4.

1.2 Chirality

Chirality is a symmetry operation in which molecules do not superpose with their mirror images (Figure 1.5), e.g. right-handed B DNA molecules (that exist naturally) do not superpose with left-handed B DNA molecules (an artificial construct). The occurrence of chirality can be determined for many organic molecules directly by inspection of their molecular structure. A carbon atom with four different groups attached to it can immediately be deduced to be chiral (Figure 1.5c) and this chirality could in turn affect the macroscopic behaviour of the material. All amino acids have chiral centres, except glycine, and so too do the nucleic acids, DNA and RNA, many carbohydrates and many lipids. This implies a fundamental molecular cause for the macroscopic chirality observed in biological materials, although spontaneous chiral symmetry breaking is also possible (in this case both chiralities occur on average in equal amounts).

Chiral interactions between molecules often dramatically perturb the phase of matter formed, e.g. the crystalline or liquid-crystalline phase adopted will become twisted with characteristic orientational defect structures (Chapter 6). Furthermore, chiral phases of matter often have unusual optical properties when they interact with photons (used commercially in liquid-crystalline displays with synthetic molecules). Ciliaal chirality in human embryos is thought to lead to the partial left/right asymmetry of the body parts in humans, e.g. the heart is on the left in most individuals.

1.3 Proteins

Polymers consist of a large number of identical subunits (monomers) connected together with covalent bonds. A protein is a special type of polymer; in a protein there are up to twenty different amino acids (Figure 1.6) that can function as monomers and all the monomers are connected together with identical peptide linkages (C–N bonds, Figure 1.7). Only twenty amino acids occur in nature that are used to create proteins in eukaryotic cells, although three additional protein forming amino acids occur in bacteria and over 140 nonprotein-forming amino acids are known. A particularly persuasive evolutionary argument of why only twenty amino acids exist in proteins in eukaryotic cells is still lacking. The best partial explanation is that twenty is enough, and reflects the common evolutionary origins of all life on the planet Earth. Synthetic chemists have made new amino acids
and connected them together to form novel synthetic proteins, so there is in principle no clear chemical barrier
that restricts the number of possibilities (which is explored in the field of synthetic biology), but natural life
only uses twenty (or twenty three if bacterial life is included in the list).

The twenty eukaryotic protein forming amino acids can be placed in different families dependent on the
chemistry of their different side groups. Five of the amino acids form a group with lipophilic (fat-liking) side
chains: glycine, alanine, valine, leucine, and isoleucine. Proline is a unique circular amino acid that is given its
own separate classification. There are three amino acids with aromatic side chains: phenylalanine, tryptophan,
and tyrosine. Sulfur is in the side chains of two amino acids: cysteine and methionine. Two amino acids have
hydroxyl (neutral) groups making them water loving: serine and threonine. Three amino acids all have very
polar positively charged side groups: lysine, arginine and histidine. Two amino acids form a family with acidic
negatively charged side groups and they are joined by two corresponding neutral counterparts that have a
similar chemistry: aspartate, glutamate, asparagine, and glutamine. More generally, the protein forming amino
acids can be separated into three principle families; hydrophobic (they hate water), polar (they like water) and
carged (they like water and are charged when they are incorporated into proteins).

The linkages between amino acids all have the same chemistry and basic geometry (Figure 1.7), which
greatly simplifies their classification on the atomic scale. The peptide linkage that connects all amino
acids together consists of a carbon atom attached to a nitrogen atom through a single covalent bond. The
condensation of two α-amino acids to form a dipeptide is shown in Figure 1.8. There is only one way that

Figure 1.5  Chirality (‘handedness’) occurs when an object cannot be superposed with its mirror image. (a) A human hand and (c) a carbon molecule with four different substituents are chiral objects. (b) A flat letter ‘G’ is achiral in three dimensions, since it can be lifted up and overlaid with its mirror image. (c) Such chiral carbon atoms are found in many organic molecules and give rise to their macroscopic chiral phases, e.g. amino acids in the α helices of proteins can form twisted cholesteric liquid-crystalline phases. The letters A, B, D, and E denote arbitrary distinct chemical substituents, whereas C is a carbon atom.
Figure 1.6  The chemical structure of the twenty amino acids that form proteins in eukaryotic cells.
two amino acids can be connected in proteins, the peptide linkage, and the chemical formula for four amino acids connected in line is

\[-\text{CR}_1\text{HCONH-}\text{CR}_2\text{HCONH-}\text{CR}_3\text{HCONH-}\text{CR}_4\text{HCONH-}\]

where the Rs are the groups that differentiate the 20 different amino acids and the hyphens are the peptide linkages. The peptide linkage has a directionality and in general -A-B- is usually different from -B-A-. Peptides are thus conventionally written with the N-ending peptide (the unbound amine terminus) on the left and the C-ending peptide (the unbound carboxyl terminus) on the right.

Although the chemistry of peptide linkages is fairly simple, to relate the primary sequence of amino acids (the combination of amino acids along the chain) to the resultant three-dimensional structure of the proteins is very complicated and predominantly remains an unsolved problem. To describe protein structure in more detail it is useful to describe motifs of secondary structure that occur in their morphology. The motifs include alpha helices, beta sheets and beta barrels (Figure 1.9). The first beta sheet and alpha helical structures were suggested by William Astbury during the 1930s. He also proposed that adjacent proteins were held together in hair and wool by hydrogen bonds. The first refined models of protein structure required more quantitative

---

**Figure 1.7** All of the amino acids have the same primitive structure and are connected with the same peptide linkage through C–N bonds (O, N, C, and H indicate oxygen, nitrogen, carbon and hydrogen respectively). R is a pendant side group that provides the amino acid with its identity, i.e. proline, glycine, etc (Figure 1.6). Defining the rotational angles (ψ, φ) for each amino acid gives a reasonably compact description of the peptide backbone conformation (it leads to Ramachandran maps of protein conformation).

**Figure 1.8** A condensation reaction between two peptides can create a peptide linkage and a water molecule.
analysis of X-ray diffraction patterns and the exact geometries of peptide bonds were calculated by Linus Pauling and Robert Corey in 1951.

Standard alpha helices in proteins are right-handed (amino acid carbon atoms are often chiral, since they are attached to four different groups, Figure 1.5c) and the pitch (repeat distance) is 3.6 amino acid groups (Figure 1.9a). The helical chirality can also be observed on the next larger length scale in a coiled coil structure, a helix of helices, although the pitch is less well defined on the helical superstructure that exists on the largest length scale. A standard molecular protein alpha helix is 0.54 nm in width. Each amino acid is displaced at ~100° to its predecessor along the chain and there is a small 0.15 nm displacement along the axis of the helix. The alpha helix structure is stabilised by hydrogen bonds between neighbouring amino acids (N–H groups interact with C=O groups). Typically an amino acid has hydrogen bonds that interact most strongly with its neighbour four steps along the chain.

Peptide chains can fold back on themselves with one strand adjacent to the next, in the form of beta sheets. In a similar manner to alpha helices, the structures are stabilised by hydrogen bonds between neighbouring amino acids, principally the N–H and C=O groups. The pleating of chains causes the distance between peptides to be ~0.6 nm along a single chain. The distance between adjacent chains is ~0.5 nm. The carbon atoms in beta sheet peptides are chiral, so the sheets have a chiral twist to them. Peptide chains have a directionality (N terminus versus the C terminus), which is indicated by an arrow on a diagram (Figure 1.9b). Adjacent beta sheet strands
can be arranged antiparallel, parallel or in combinations of the two. The antiparallel arrangement (alternating N terminus directions) is slightly more energetically stable than the other possibilities.

The full three-dimensional tertiary structure of a protein usually takes the form of compact globular morphologies (the globular proteins) or long extended conformations (fibrous proteins, Figure 1.10). Globular morphologies typically consist of a number of secondary motifs combined with more disordered regions of peptide. Quaternary globular structures are also possible where two or more whole protein chains are assembled side by side into a superstructure, e.g. haemoglobin is an assembly of four individual globular chains.

Charge interactions are very important for the determination of the conformation of biological polymers. The degree of charge on a polyacid or polybase (protein, nucleic acid, etc.) is determined by the pH of a solution, i.e. the concentration of hydrogen ions. Water has an ability to dissociate into oppositely charged ions, and this process depends on temperature,

\[ \text{H}_2\text{O} \rightleftharpoons \text{H}^+ + \text{OH}^- \] (1.1)

The product of the hydrogen and hydroxide ion concentrations formed from the dissociation of water is a constant \((K_w)\) at room temperature,

\[ c_{\text{H}^+}c_{\text{OH}^-} = 1 \times 10^{-14} \text{M}^2 = K_w \] (1.2)

where \(c_{\text{H}^+}\) and \(c_{\text{OH}^-}\) are the concentrations of hydrogen and hydroxide ions, respectively. Addition of acids and bases perturbs the equilibrium dissociation process of water and the acid/base equilibrium phenomena involved are a cornerstone of the physical chemistry of aqueous solutions. Due to the vast range of possible hydrogen ion \((\text{H}^+)\) concentrations typically encountered in aqueous solutions, it is normal to use a logarithmic scale (pH) to quantify them. The pH is defined as the negative logarithm (base 10!) of the hydrogen ion concentration,

\[ \text{pH} = -\log c_{\text{H}^+} \] (1.3)

Typical values of pH range from 6.5–8 in physiological intracellular conditions. pHs in the range 1–2 are strongly acidic and pHs in the range 12–13 are strongly basic.

When an acid (HA) dissociates in solution it is possible to define an equilibrium constant \((K_a)\) for the dissociation of its hydrogen ions \((\text{H}^+)\),

\[ \text{HA} \rightleftharpoons \text{H}^+ + \text{A}^- \]
\[ \text{HA} \xrightleftharpoons{\text{H}^+ + \text{A}^-} K_a = \frac{c_{\text{H}^+} c_{\text{A}^-}}{c_{\text{HA}}} \] (1.4)

where \(c_{\text{H}^+}\), \(c_{\text{A}^-}\) and \(c_{\text{HA}}\) are the concentrations of hydrogen ion, acid ion and acid respectively. Since the hydrogen ion concentration follows a logarithmic scale it is natural to define the dissociation constant on a logarithmic scale (pK\(_a\)) as well,

\[ pK_a = -\log K_a \] (1.5)

The logarithm of both sides of equation (1.4) can be taken to give a relationship between the pH and the pK\(_a\) value,

\[ \text{pH} = pK_a + \log \left( \frac{c_{\text{A}^-}}{c_{\text{HA}}} \right) \] (1.6)

where \(c_{\text{A}^-}\) and \(c_{\text{HA}}\) are the concentrations of the conjugate base (e.g. A\(^-\)) and acid (e.g. HA), respectively. This equation enables the degree of dissociation of an acid (or base) to be calculated and it is named after Henderson and Hasselbalch. Thus, from a knowledge of the pH of a solution and the pK\(_a\) value of an acid or basic group the charge fraction on a protein can be found to a first approximation. The propensity of the amino acids to dissociate in water is described in Table 1.1. In contradiction to what their name might imply, only amino acids with acidic or basic side groups are actually charged when incorporated into proteins. These charged amino acids are arginine, aspartic acid, cysteine, glutamic acid, histidine, lysine and tyrosine.

Another important interaction between amino acids is determined by the degree to which they are able to form hydrogen bonds with the surrounding water molecules. This amino acid hydrophobicity (the amount they

**Table 1.1** Fundamental physical properties of the twenty amino acids that form proteins in eukaryotic cells. [Ref. Data adapted from C.K. Matthews, K.E. van Holde, K.G. Ahem, Biochemistry, 3rd edition, Prentice Hall, 1999.]

<table>
<thead>
<tr>
<th>Name</th>
<th>pK(_a) value of side chain</th>
<th>Mass of residue (Da)</th>
<th>Occurrence in natural proteins (% mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Neutral</td>
<td>71</td>
<td>9.0</td>
</tr>
<tr>
<td>Arginine</td>
<td>12.5</td>
<td>156</td>
<td>4.7</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Neutral</td>
<td>114</td>
<td>4.4</td>
</tr>
<tr>
<td>Aspartic acid</td>
<td>3.9</td>
<td>115</td>
<td>5.5</td>
</tr>
<tr>
<td>Cysteine</td>
<td>8.3</td>
<td>103</td>
<td>2.8</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Neutral</td>
<td>128</td>
<td>3.9</td>
</tr>
<tr>
<td>Glutamic acid</td>
<td>4.2</td>
<td>129</td>
<td>6.2</td>
</tr>
<tr>
<td>Glycine</td>
<td>Neutral</td>
<td>57</td>
<td>7.5</td>
</tr>
<tr>
<td>Histidine</td>
<td>6.0</td>
<td>137</td>
<td>2.1</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Neutral</td>
<td>113</td>
<td>4.6</td>
</tr>
<tr>
<td>Leucine</td>
<td>Neutral</td>
<td>113</td>
<td>7.5</td>
</tr>
<tr>
<td>Lysine</td>
<td>10.0</td>
<td>128</td>
<td>7.0</td>
</tr>
<tr>
<td>Methionine</td>
<td>Neutral</td>
<td>131</td>
<td>1.7</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Neutral</td>
<td>147</td>
<td>3.5</td>
</tr>
<tr>
<td>Proline</td>
<td>Neutral</td>
<td>97</td>
<td>4.6</td>
</tr>
<tr>
<td>Serine</td>
<td>Neutral</td>
<td>87</td>
<td>7.1</td>
</tr>
<tr>
<td>Threonine</td>
<td>Neutral</td>
<td>101</td>
<td>6.0</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Neutral</td>
<td>186</td>
<td>1.1</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>10.1</td>
<td>163</td>
<td>3.5</td>
</tr>
<tr>
<td>Valine</td>
<td>Neutral</td>
<td>99</td>
<td>6.9</td>
</tr>
</tbody>
</table>
dislike water) is an important driving force for the conformation of proteins. Crucially it leads to the compact conformation of globular proteins, as the hydrophobic groups are buried in the centre of the globule to avoid contact with the surrounding water.

Covalent interactions are possible between adjacent amino acids and can produce solid protein aggregates (Figures 1.10 and 1.11). For example, disulfide linkages are possible in proteins that contain cysteine and these disulfide bonds form the strong interprotein linkages that are found in fibrous proteins, e.g. keratins in hair.

The internal secondary structures of protein chains (α helices and β sheets) are stabilised by hydrogen bonds between adjacent atoms in peptide groups along the main chain. The important structural proteins such as keratins, collagens (Figure 1.11), silks (Figure 1.10), arthropod cuticle matrices, elastins (Figure 1.12), resilin and abductin are formed from a combination of both intermolecular disulfide and hydrogen bonds.

Some examples of the globular structures adopted by proteins are shown in Figure 1.13. Globular proteins can be denatured in a folding/unfolding transition. Typically the complete denaturation transition is a first-order thermodynamic phase change with an associated latent heat (thermal energy is absorbed during the transition). The unfolding process involves an extremely complex sequence of molecular origami transitions. There are a vast number of possible molecular configurations \(2^{N-1}\) for an \(N\) residue protein, with
the very restrictive assumption that each peptide linkage only has two possible conformations; clearly a lower limit on the calculation) during the reverse process of protein folding, when the globular protein is constructed from its primary sequence by the cell, and thus frustration can be important. At first sight it appears a certainty that protein molecules will become trapped in an intermediate state and never reach their correctly folded form, since the process appears to require astronomically long time scales. This is called Levinthal’s paradox, the process by which natural globular proteins manage to find their native state among the billions of possibilities in a finite time. The current explanation of protein folding is that there is a funnel of energy states that guides the kinetics of folding across the complex energy landscape that has a huge number of degrees of freedom (Figure 1.14).

There are two main types of interchain interaction between different proteins in solution; those in which the native state remains largely unperturbed in processes such as protein crystallisation and the formation of filaments in sheets and tapes, and those interactions that lead to a loss of conformation, e.g. heat set gels (e.g. table jelly, and boiled eggs) and amyloid fibres (e.g. Alzheimer’s disease, and Bovine Spongiform Encephalopathy).

A wide range of protein structures with atomic resolution are known from X-ray crystallography experiments. This data and the associated three-dimensional visualisation software are freely available on the internet and are well worth a quick investigation (search online for the ‘protein data bank’).

1.4 Lipids

Cells are divided into a series of subsections and compartments by membranes that are formed predominantly from lipids (Chapter 2). The other main roles for lipids are as energy-storage compounds and for cell signalling in messenger molecules such as steroid hormones (oestrogen and testosterone).

Lipids are amphiphilic; the head groups like water (and hate fat) and the tails like fat (and hate water). This amphiphilicity drives the spontaneous self-assembly of the molecules into membranous morphologies.

There are four principle families of lipids, fatty acids with one or two tails (including carboxylic acids of the form RCOOH, where R is a long hydrocarbon chain), and steroids and phospholipids, where two fatty acids are
linked to a glycerol backbone (Figure 1.15). The type of polar head group differentiates the particular species of naturally occurring lipid. Cholesterol is a member of the steroid family and these compounds are often found in membrane structures. Glycolipids also occur in membranes and in these molecules the phosphate group on a phospholipid is replaced by a sugar residue. Glycolipids have important roles in cell signalling and the immune system. For example, these molecules are an important factor for the determination of blood cell compatibility during a blood transfusion.

Fatty acids are some of the simplest lipids and often consist of 16 to 18 carbon atoms arranged in a chain, e.g. oleic acid (a free fatty acid, found in olive oil). This chain can contain single or double covalent bonds and has nonpolar C–H bonds except for at the end group e.g. CH₃(CH₂)₇CH=CH(CH₂)₇COOH. Fatty acids are stored long term in the form of triacylglycerols.

Phospholipids are the main components of eukaryotic cell membranes. They consist of two fatty acids chains joined together with a polar head group and are thus amphiphilic. Examples include phosphatidylcholine (a pulmonary surfactant) and phosphatidylethanolamine (a constituent of the membranes of nervous tissue).

1.5 Nucleic Acids

The ‘central dogma of biochemistry’ according to F.C. Crick is illustrated in Figure 1.16a. DNA contains the basic blueprint that guides the construction of the vast majority of living organisms. To implement this
Figure 1.15  Space-filling models of some lipid molecules typically encountered in biology (a) fatty acid with one tail (oleate), (b) phospholipid and (c) cholesterol.

Figure 1.16  (a) The central dogma of molecular biology considers the transcription and translation of DNA. DNA is transcribed to form a messenger RNA (mRNA) chain and this information is translated into a protein sequence. The dotted line indicates reverse transcription that can happen through the action of retroviruses, e.g. HIV or leukaemia virus. DNA is also duplicated from a DNA template (replication). (b) RNA polymerase is required for transcription, whereas a ribosome is used for translation.
blueprint, cells need to transcribe the DNA to RNA and this structural information is subsequently translated into proteins that use specialised protein factories (the ribosomes). The resultant proteins can then be used to catalyse specific chemical reactions (enzymes) or be used as building materials to construct cells.

This simple biochemical scheme for the transfer of information has powerful implications. DNA can now be altered systematically using recombinant DNA technology and then placed inside a living cell, to hijack the cell’s mechanisms for translation. The proteins that are subsequently formed can be tailor-made by the genetic engineer to fulfil a specific function, e.g. bacteria can be used to form fibrous proteins that are used to create biodegradable plastics.

Modern biochemistry experiments show that some minor corrections are required to the central dogma. Some viruses (retroviruses) are able to reverse the direction of information transfer in the transcription step (Figure 1.16a). RNA is translated into DNA that resides long term in the infected organism’s genome. This is a reason for the high infection rates for lentiviruses such as HIV and leukemia virus. It does however imply an efficient mechanism for the genetic modification of organisms can be obtained using genetically modified lentiviruses, and this is now a standard technique in molecular biology.

The monomers of DNA are made of a sugar, an organic base, and a phosphate group (Figure 1.17). There are only four organic bases that naturally occur in DNA and these are thymine, cytosine, adenine, and guanine (T, C, A, and G). The sequence of bases in each strand in a double helix contains the genetic code. The base

Figure 1.17  (a) The four DNA bases Adenine (A), Thymine (T), Guanine (G) and Cytosine (C) are constructed from carbon, nitrogen, oxygen and hydrogen atoms. (b) The generic chemical structure of the base of a nucleic acid consists of a phosphate group, a sugar and a base.
pairs in each strand of double-helical DNA are complementary, \( A \) has an affinity for \( T \) (they form three hydrogen bonds) and \( G \) for \( C \) (they form two hydrogen bonds) (Figure 1.18). The interaction between the base pairs is driven by the geometry of the hydrogen-bonding sites. Thus, each strand of the DNA helix contains an identical copy of the genetic information to its complementary strand and replication can occur by separation and resynthesis of two additional chains on each of the two original double-helical strands. Methylation of cytosine and adenine in DNA is common after replication and is thought to be an important factor in gene regulation.

There is a major groove and a minor groove on the biologically active A and B forms of the DNA double helix. The individual polynucleotide chains have a sense of direction, in addition to their individuality (a complex nucleotide sequence). DNA replication \textit{in vivo} is conducted by a combination of the DNA polymerases (I, II and III).

The formation of helical secondary structures in DNA drastically increases the rigidity of each separate chain and is called a \textit{helix-coil transition}. DNA in its double helical form can store torsional energy, since the monomers are not free to rotate (like a telephone cable). The ends of a DNA molecule can be joined together in a circle to form a compact supercoiled structure, that often occurs in bacteria, and their behaviour presents a series of fascinating questions in both statistical mechanics and topological analysis (Chapter 10).
DNA has a wide variety of structural possibilities (Table 1.2, Figure 1.19). There are three standard types of double helix labelled A, B, and Z that occur in solid fibres and are deduced from fibre diffraction experiments that average over a huge number of base pair conformations. Typically, DNA in solution has a structure intermediate between A and B, dependent on the chain sequence and the exact aqueous environment. An increase in the level of hydration tends to increase the number of B-type base pairs in an aqueous double helix. Z-type DNA is favoured in some extreme nonphysiological conditions and has little importance for naturally occurring cells.

There are a number of local structural modifications to the helical structure that are dependent on the specific chemistry of the individual strands and occur in addition to the globally averaged A, B and Z classifications. The kink is a sudden bend in the axis of the double helix which is important for complexation in the nucleosome. The loop contains a rupture of hydrogen bonds over several base pairs and the separation of two nucleotide chains produces loops of various sizes. During the process of DNA transcription RNA polymerase is bound to DNA to form a loop. In the breathing of a double helix a temporary break in the hydrogen bonds is caused by a rapid partial rotation of one base pair, which makes the hydrogen atoms in the NH groups accessible and enables them to be exchanged with neighbouring protons in the presence of a catalyst. The Cruciform structure is formed in the presence of self-complementary palindromic sequences separated by several base pairs. Hydrophobic molecules (e.g. DNA active drugs) can be intercalated in DNA, i.e. slipped between two base pairs. Helices that contain three or four nucleic acid strands are also possible with DNA, but do not occur in vivo. Real double helices experience thermal fluctuations on the nano scale, so there are breathing modes of the base pairs as well as flexural modes of the chains’ backbone. Furthermore the local ordering depends on the base sequence to some extent and phonons (quantised lattice vibrations) travel through the structure (not to mention the possibility of solitons due to nonlinear interactions). The DNA chain is highly charged and interacts Coulombically with itself, neighbouring chains, and a fluctuating ion cloud that contains a large number of associated counterions. There is clearly a great deal of molecular diversity in the structure of DNA chains and it represents some knotty structural problems.

DNA has interesting features with respect to its polymer physics. The persistence length ($l_p$, a measure of the chains’ flexibility) of DNA is on the order of 60 nm for E-coli (which depends on ionic strength), it can have millions of monomers in its sequence and a correspondingly gigantic contour length ($L$) (for humans $L$ is 95 mm!). The large size of DNA has a number of important consequences; fluorescently labelled DNA is visible with an optical microscope and the cell has to solve a tricky packaging problem in vivo (it uses chromosomes) to fit the DNA inside the nucleus of a cell, which is at most a few micrometers in diameter (Chapter 18).

DNA and RNA are both made from nucleic acids and their chemistry is closely related. Ribonucleic acids have the same base pairs except thymine is replaced by uracil (U). DNA can form double helices, whereas RNA predominantly does not. RNA is chemically less stable, which implies an improved evolutionary fitness for long-term data storage in DNA, rather than in RNA.

### Table 1.2

**Averaged structural parameters of polynucleotide helices calculated from X-ray fibre diffraction patterns.**

<table>
<thead>
<tr>
<th>Property</th>
<th>A form</th>
<th>B form</th>
<th>Z form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direction of helix rotation</td>
<td>Right</td>
<td>Right</td>
<td>Left</td>
</tr>
<tr>
<td>Number of residues per turn</td>
<td>11</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Rotation per residue (°)</td>
<td>33°</td>
<td>36°</td>
<td>30°</td>
</tr>
<tr>
<td>Rise in helix per residue (nm)</td>
<td>0.26</td>
<td>0.34</td>
<td>0.37</td>
</tr>
<tr>
<td>Pitch of helix (nm)</td>
<td>2.8</td>
<td>3.4</td>
<td>4.5</td>
</tr>
</tbody>
</table>
The range of applications for DNA analysis and manipulation in molecular biology is huge, with possibilities such as transgenic organisms, genetic analysis of disease, and a molecular understanding of evolution (Chapter 2).

1.6 Carbohydrates

Historically, advances in carbohydrate research have been overshadowed by developments in protein and nucleic acid science. This has in part been due to the difficulties in the analysis of the structure of carbohydrates and the extremely large number of chemical structures that exist naturally.

Figure 1.19 Molecular models of A-, B- and Z-type double helical structures of DNA. A- and B-type helical structures typically occur in biological systems. Z-DNA helical structures crystallise under extreme nonphysiological conditions.
There are two important glucose polymers that occur in plants (the first and second most important biopolymers in terms of biomass on the Earth) that are differentiated by the linkage between the monomers; cellulose and amylopectin. Cellulose is a very rigid polymer and has both nematic liquid-crystalline and semicrystalline phases. It is used widely by plants as a structural material. The straight chain formed by the $\beta(1\rightarrow4)$ linkage between glucose molecules is optimal for the construction of fibres, since it gives them a high tensile strength (Figures 1.20 and 1.21) in the chain direction and reasonable strength perpendicular to the chain due to the substantial intrachain hydrogen bonding in sheet-like structures. Amylose and its branched form, amylopectin (starch), are used in plants to store energy and often this material adopts smectic liquid-crystalline phases (Figure 1.22). Starches form the principle component of mankind’s food sources. In amylose the glucose molecules are connected together with $\alpha(1\rightarrow4)$ linkages. $\alpha$-Linkages between glucose molecules are well suited to the formation of an accessible sugar store, as they are flexible and can be easily degraded by enzymes. Amylopectins are formed from amyloses with additional $\alpha(1\rightarrow6)$ flexible branched linkages between glucose molecules (Figure 1.23) and contain both double-helical and amorphous structures. Amylopectin is stored in a hierarchical liquid-crystalline architecture. Glycogen is an analogous amorphous hyperbranched glucose polymer similar to amylopectin that is used in animals as an energy store.

**Figure 1.20** Sheet-like structures formed in cellulosic materials. The $\beta(1\rightarrow4)$ linkages between glucose molecules induce extended structures and the chains are linked together with hydrogen bonds.

**Figure 1.21** The hierarchical structure of cellulose found in plant cell walls. Polymeric cellulose chains are combined into microfibrils that form the walls of plant cells.
Figure 1.22  Four length scales are important in the hierarchical structure of starch, a glucose storage material found in plants; (a) the whole granule morphology (~μm), and the growth rings (~100 nm), (b) the crystalline and amorphous lamellae (~9 nm), and (c) the molecular structure of the amylopectin (~Å). [Reproduced with permission from T.A. Waigh, PhD Thesis, University of Cambridge, 1996.]

Figure 1.23  The branched primary structure found for amylopectin in starch. Both α(1-4) (linear regions) and α(1-6) (branched regions) flexible linkages occur between glucose monomers.
Chitin is another structural polysaccharide that forms the exoskeleton of crustaceans and insects. It is similar in its functionality to cellulose, it is a very rigid polymer and has a cholesteric liquid-crystalline phase. It must be emphasised that the increased complexity of linkages between sugar molecules, compared with nucleic acids or proteins, provides a high density mechanism for encoding information. A sugar molecule can be polymerised in a large number of ways e.g. the six carbons of a glucose molecule could each be polymerised that provides an additional $6^{N-1}$ arrangements for a carbohydrate compared with a protein of equivalent length ($N$) in which there is only one possible mechanism for the connection of amino acids, the peptide linkage. These additional possibilities for information storage with carbohydrates are used naturally in a range of immune response mechanisms and signalling pathways.

Pectins are extracellular plant polysaccharides that form gums (used in jams), and similarly algins can be extracted from sea weed. Both are widely used in the food industry. Hyaluronic acid is a long negatively charged semiflexible polyelectrolyte and occurs in a number of roles in animals. For example, it is found as a component of cartilage (a biological shock absorber) and as a lubricant in synovial joints.

1.7 Water

Water is a unique polar solvent and its properties have a vast impact on the behaviour of biological molecules (Figure 1.24). Water has a high dipole moment of $6.11 \times 10^{-30}$ C m, a quadrupole moment of $1.87 \times 10^{-39}$ C m$^2$ and a mean volume polarisability of $1.44 \times 10^{-30}$ m$^3$.

Water exists in a series of crystalline states at subzero temperature or elevated pressures. The structure of ice formed in ambient conditions has unusual cavities in its structure due to the directional nature of hydrogen bonds and it is consequently less dense than liquid water at its freezing point. The polarity of the OH bonds formed in water drives the association into dimers, trimers, etc. (Figure 1.14) and causes a complex many-body problem for the description of water in both liquid and solid condensed phases.

Antifreeze proteins have been designed through evolution to have an alpha-helical dipole moment that disrupts the ability of the surrounding water to crystallise through the disruption of the hydrogen-bonded network structure. These antifreeze molecules are often found in organisms that exist in subzero temperatures, e.g. arctic fish and plants. Genetically engineered fish antifreeze proteins are added to ice cream to increase its shelf life and improve its consistency.

Images of biological processes can be created in vivo using the technique of nuclear magnetic resonance that depends on the mobility of water to create the images (Section 19.9). This powerful noninvasive method allows water to be viewed in a range of biological processes, e.g. in the cerebral activity of living organisms.

Figure 1.24  The geometry of a single water molecule that becomes tetrahedral once hydrogen bonded to other water molecules in ice crystals (Figure 1.4). The oxygen molecule is shown in red and the hydrogen molecule is shown in green.
Even at very low volume fractions water can act as a plasticiser, to change solid biopolymers between glassy and nonglassy states. This ingress of water can act as a switch, to trigger cellular activity in plant seeds, and such dehydrated cellular organisms can remain dormant for many thousands of years e.g. seeds found entombed in Egyptian pyramids can still be made to successfully germinate.

A wide range of time scales ($10^{-18}$–$10^{3}$ s) of water are important to understand its biological function (Figure 1.25). The range of time scales includes such features as the elastic collisions of water at ultrafast times ($\sim 10^{-15}$ s) to the macroscopic hydrodynamic processes observed in blood flow at much slower times ($\sim$ s) (Chapter 14).

The physical interaction between water and other molecules is of key importance in the determination of the properties of biological cells. Water is a polar molecule (the hydrogen–oxygen covalent bonds are dipolar, since they are readily polarised). The hydrogen atoms have a slight positive charge and the oxygen atoms have a slight negative charge. Water molecules can form hydrogen bonds due to the interaction energy of the dipoles and they also interact with other charged molecules, e.g. charged ions.

As a result of water’s high dielectric permittivity ions and polar molecules are readily soluble, i.e. the attractive electrostatic interactions are readily screened by the charged water molecules. Furthermore, uncharged molecules that can form hydrogen bonds often dissolve in water. Together, both sets of molecules are called hydrophilic. Nonpolar molecules tend to avoid contact with water in solution and tend to closely associate with one another (hydrophobic). It is possible for a single molecule to have both hydrophobic and hydrophilic regions, and these schizophrenic molecules play a crucial role in the determination of many biological structures, e.g. the shape of globular proteins (hydrophobic sections are tucked away in the centres) and the structure of membranes (hydrophobic regions are placed back to back in a bilayer) are determined by hydrophobicity.

1.8 Proteoglycans and Glycoproteins

Proteoglycans (long carbohydrate molecules attached to short proteins) and glycoproteins (short carbohydrate molecules attached to relatively long proteins) are constructed from a mixture of protein and carbohydrate molecules (glycosoaminoglycans). In common with carbohydrates, proteoglycans and glycoproteins exhibit extreme structural and chemical heterogeneity. Furthermore, the challenges presented to crystallography by their noncrystallinity, means that a full picture of their biological function is still being developed. There still

\[\text{Figure 1.25} \quad \text{The range of time scales that determine the physical properties of liquid water shown on a logarithmic scale. [Reproduced with permission from The Structure and Properties of Water by D. Eisenberg & W. Kauzmann (2005). By permission of Oxford University Press.]}\]
remains a ‘here lies beasties’ sign on the glycoprotein area of the map of molecular biology. This is both metaphorically and technically correct, since a vast range of micro-organisms spend their lives swimming through mucins inside human stomachs and intestines.

Many proteoglycans and glycoproteins used in the extracellular matrix (regions outside the cells that are often carefully regulated) have bottle-brush morphologies (Figures 1.26 and 1.27). An example of a sophisticated proteoglycan architecture is aggrecan and it consists of a bottle-brush of bottle-brushes (Figure 1.26). These materials have a very large viscosity and are used to dissipate energy in collagenous cartilage composites and to reduce friction in synovial joints. An example of an extracellular glycoprotein is the mucin (MUC5AC) found in the stomach of mammals. These molecules experience telechelic (either end) associations to form thick viscoelastic gels that protect the stomach lining from autodigestion (Figure 1.27). In addition to the secreted gelling mucins, many genes in the mammalian genome express mucins that exist tethered to the surfaces of epithelial cells. These provide steric protection for the cellular membranes (polymeric brushes) and their malfunction is implicated in some respiratory diseases, e.g. chronic obstructive pulmonary disorder or cystic fibrosis.

Other examples of glycoproteins occur in enzymes (Ribonuclease B), storage proteins (egg white), blood clots (fibrin) and antibodies (Human IgG).

### 1.9 Viruses

Viruses are intracellular parasites; biological entities that multiply by the invasion of cells. Often viruses are simple molecular constructions that consist of a single nucleic acid chain covered with a repeated pattern of proteins that form a coat. In addition to aspects related to their biological role in disease, viruses have attracted a great deal of attention from biophysicists for their physical properties. Viruses often self-assemble into well-defined monodisperse geometrical shapes (rods and polyhedra) (Figure 1.28), from their constituent components. Such materials have proven ideal model systems for the examination of the phase behaviour of charged colloids and lyotropic liquid crystals (Chapter 6), and in terms of the self-assembly of their native structure in solution (Chapter 8).
Figure 1.27  (a) Mammalian stomach mucin contains a series of sticky telechelic bottle-brushes. The dumbbell monomers assemble end-on-end to form symmetric dimers that subsequently polymerise to form a network structure. Crosslinking/branching of fibres occurs at low pHe and creates thick viscoelastic gels. Peptide regions are shown in blue and carbohydrates are shown in red. (b) Many mucin genes code for epithelial mucins that form membrane-tethered coats on epithelial cells (surface-tethered polymeric brushes). Peptide regions are shown in green and carbohydrates are shown in red.

Figure 1.28  Schematic diagram of a range of virus structures; rod-like (TMV), asymmetric (bacteriophage), and icosohedral (picorna).
Spherical (icosahedral) type viruses have well-defined symmetries and the rules for their tessellation (the recipes they use to fill geometrical space) are now mathematically well described. Many of these viruses are important for human health, e.g. polio virus.

Tobacco mosaic virus (TMV) is a much-studied fibrous virus that infects tobacco plants. TMV forms long extended structures, which can be self-assembled in vitro by the correct choice of physical conditions (pH, temperature, etc.) and the viruses so formed can then go on to infect tobacco plants (Section 8.2).

1.10 Other Molecules

The inorganic ions of the cell include sodium (Na\(^+\)), potassium (K\(^+\)), magnesium (Mg\(^{2+}\)), calcium (Ca\(^{2+}\)), chloride (Cl\(^-\)), hydrogen carbonate (HCO\(_3\)^-), and phosphate (HPO\(_4^{2-}\)). These ions only constitute a small fraction of the cell’s weight with less than 1% of the cell’s mass, but they play a critical role in cell function and a huge amount of the human body’s energy consumption is spent pumping them into and out of cells with membrane proteins. Inorganic ions are particularly important in electrically active cells (e.g. neurons), since they give rise to action potentials used in signalling.

Adenosine diphosphate (ADP) and adenosine triphosphate (ATP) are the ‘currency of energy’ in many biochemical processes. One molecule can release around 20 \(kT\) \((8.2 \times 10^{-20} \text{ J})\) of biochemically useful energy. The energy is stored by the addition of an extra phosphate in the ATP structure and can be released when it is metabolised into ADP (Figure 1.29). Phosphate groups are highly charged and electrostatic energy is liberated once the end phosphate group of the row of three in ATP is chopped off. Guanosine triphosphate (GTP), hydrogen ions and sodium ions are also used as fuels in cells, but they occur less frequently.

There is a vast range of other biomolecules (polyphenols, vitamins, etc.) that have not been covered in this short introductory section and the reader should refer to a specialised biochemistry textbook for details, e.g. Berg, Tymoczko and Stryer.

Suggested Reading

If you can only read one book, then try:


### Tutorial Questions 1

1.1 Make a list of ten biological molecules and give a disease associated with the absence/malfunction of each, e.g. insulin and diabetes.

1.2 Metals occur in a range of biological processes and form a key component of the structure in a number of biological molecules. Make a list of some of the biological molecules in which metal atoms occur.

1.3 Explain how peptide chains find their unique folded structure. Calculate the number of permutations for a chain of 200 amino acids if it is assumed there are two possible conformational states for each amino acid monomer.

1.4 Describe the practical barriers that impede our understanding of the structure of carbohydrates and their complexes (glycoproteins, glycolipids, etc.).

1.5 The primary sequence of proteins can be predicted from the genome. The human genome project was completed in 2003. Explain why there still are many unsolved problems in the molecular biology of human cells.

1.6 Cancers are genetic diseases that affect DNA. Name some different carcinogenic materials. Mechanistically explain the difference between a cancerous and noncancerous genetic mutation.

1.7 A DNA chain has a molecular weight of $4 \times 10^8$ and the average monomer molecular weight of a nucleic acid is 660 Da. For an A-type helix there are 11 residues per helical pitch, and the translation per residue is 2.6 Å. For a B-type helix there are 10 residues per helical pitch, and the translation per residue 3.4 Å. For a Z-type helix there are 12 residues per helical pitch and the translation per residue is 3.7 Å. Calculate the length in cm of a duplex DNA chain if it is in the A, B and Z helical forms. Find the average size of the nucleus in a mammalian cell, e.g. from a Google search. Explain how the cell manages to accommodate the DNA chain in its nucleus.

1.8 Suppose that a micelle is isolated that contains a single protein that normally exists as a transmembrane molecule. Describe how the lipid and protein would be expected to be arranged on the surface of the micelle.

1.9 Calculate the pH of a 0.2 M solution of the amino acid arginine if its $pK_a$ value is 12.5.