Digital Biology

L. Ridgway Scott
University of Chicago

Ariel Fernández
Rice University

Release 0.4
DO NOT DISTRIBUTE

May 10, 2012

Contents

Chapter 1

How is biology digital?

Conventional views of biology have been qualitative, not quantitative. There is no periodic table of biological traits. Although some fundamental operations of nature appear continuous (analog) at first, we will see that there is a discrete (quantum) interpretation that is essential in explaining function. We will see that protein biophysics can be understood in terms that are used to describe digital computers. We will use this terminology to differentiate between analog systems (such as the connection between a computer and its monitor) and digital systems, which talk in zeros and ones. Fundamental in this dichotomy is the notion of a switch that can change behavior discretely from one state to another. Not surprisingly, biology utilizes many switches at a fundamental level.

Biology can be viewed as an information system. As a simple example, we are biological entities communicating via this book. More to the point, many types of signaling in biological systems involve interactions between proteins and ligands.[1] A type of physical baton-passing is used to communicate information. But there are too many examples of information processing in biology to stop here to enumerate them. What is of interest here is to understand how certain biological systems (involving proteins) function as digital information systems despite the fact that the underlying processes are analog in nature.

There are many areas of research that fit under the rubric of quantitative biology. The entire field of biology is becoming more quantitative. We primarily study proteins and their interactions because of the **protein bottleneck**. There is tremendous biotic diveresity, and corresponding genomic diversity, but the number of protein structures appears to be quite small according to existing measures, such as the SCOP classification (Section 5.1.2). Moreover, the observable diversity of the macrobiological world is evident. Thus a central question in quantitative biology is how such diversity gets transferred through proteins, which would appear to be a bottleneck according to current understanding.

It should not be strange that continuous behaviors can result in quantum components. Quantum mechanics explains continuous wave behavior in terms of discrete (atomistic) behavior. Proteins do this at a much larger length scale, but the principle is similar. Proteins are often involved in signaling and function in a discrete (or digital, or quantized) way. In addition, proteins are discrete

---

[1]A ligand is anything that binds to something. We provide a glossary of terms like this in Chapter 19 rather than defining them in the text.

building blocks of larger systems, such as viruses and cells. How they bind together (e.g., in a virus capsid) is also deterministic (repeatable) and precise. But the underlying chemical/physical mechanisms used are fundamentally continuous (analog).

Digital circuits on computer chips are also based on continuous mechanisms, namely electrical currents in wires and electronic components. The analogy with our topic is hopefully apparent. A book by Mead and Conway [289] written at the end of the 1970's transformed computer architecture by emphasizing design rules that simplified the task of converting a fundamentally analog behavior into one that was digital and predictable. We seek to do something analogous here, but we are not in a position to define rules for nature to follow. Rather, we seek to understand how some of the predictable, discrete behaviors of proteins can be explained as if certain methodologies were being used.

The benefits of finding simple rules to explain complicated chemical properties are profound. The octet rule (Section 2.1) for electron shell completion allowed rapid prediction of molecule formulation by simple counting [325]. Resonance theory (Section 13.1) describes general bonding patterns as a combination of simple bonds (e.g, single and double bonds) [324]. The discrete behavior of DNA elucidated by Crick, Franklin, Watson, Wilkins and others [157, 413, 418] initiated the molecular biology revolution. Our objective here is to provide an introduction to some basic properties of protein-ligand interactions with the hope of stimulating further study of the discrete nature of molecular interactions in biology.

Much of biology is about describing significant differences between things of interest that may look similar at first glance, as well as identifying similarities among things which look superficially different. Here the main focus will be on differences between proteins. For example, we would like to identify the difference between proteins that support animal life and those that are toxic. Such differences in chemistry can be quite small. The small difference between methanol and ethanol is a well known example. Similarly, propylene glycol is a methylated form of ethylene glycol. The latter is toxic to animals but the former is not, and both are effective as anti-freeze. We will see that such changes via a single methylation alone are able to have profound effects on the behavior of anti-cancer drugs [138, 119, 144]. We will explain simple rules that provided guidance in making such modifications. The key issue is that methylation changes the dielectric effect of water locally.

The only force of interest in biochemistry is the electric force. Electrical gradients in proteins are among the largest known in nature. Moreover, we are primarily interested in proteins operating in an aqueous, and thus dielectric, environment. We will devote significant space to the dielectric effect in subsequent chapters, but the main point to know for now is that a dielectric medium shields (diminishes) the effect of electric charges. The dielectric properties of water are among the strongest in nature, and indeed water can be viewed as hostile to proteins. This leads to an interesting contention that we address in more detail in Section 2.2.2, but for now we depict these as opposing arrows in Figure 1.1.

One way to envisage the dielectric effect and protein charges is to imagine a harbor in fog. The red and green buoy lights correspond to the positive and negative charges of a protein, and the fog is the dielectric medium that tends to shield (obscure) the charges. Fog can be dispersed locally by some atmospheric change, and the lights will be suddenly more visible. In proteins, hydrophobic groups tend to reduce the fog of the dielectric. Their intervention helps to make decisions quicker

Large electrostatic gradients   Screening by dielectric effect

Modulation of dielectric strength by hydrophobic effect
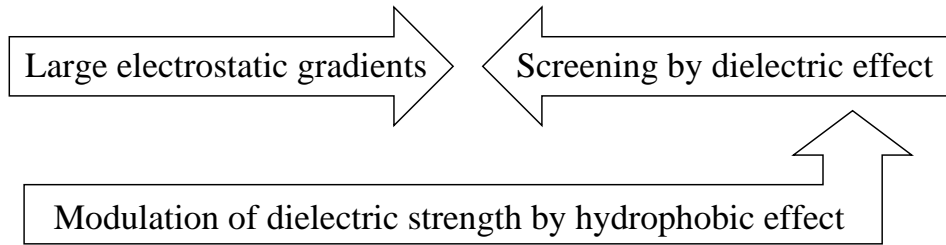
Figure 1.1: Three competing effects that determine protein behavior. These conspire to weaken interactive forces, making biological relationships more tenuous and amenable to mutation.

and more decisive. The delicate balance of interactions leads to effective swithces which are both sensitive and energetically efficient. Just how this happens is a major focus of the book.

Not only is the dielectric coefficient of water remarkably large, but it is also capable of being strongly modulated in ways that are still being unveiled. In particular, hydrophobic effects modulate the dielectric properties of water [100]. Proteins are an amazing assembly of hydrophobic, hydrophilic and amphiphilic side chains. Moreover, the charge variation on proteins is so large that it is hard to make an analogy on larger scales, and the variation in hydrophobicity is equally extreme. Hydrophobic mediation of the dielectric properties of water appears to have significant impact on protein function. Thus we are faced with a series of counterbalancing and extreme properties, depicted in Figure 1.1, that must be comprehended in order to see how proteins are functioning at a biophysical level.

Our take home message is that the modulation of the dielectric properties of water by the hydrophobic parts of proteins is an essential aspect of molecular chemistry that needs to be considered carefully. Typical representations of proteins show only physical location, basic bonds and individual charges. Adding a way of viewing the modulation of the dielectric environment is of course complex. We review one effective technique that utilizes a representation which signals the effect of the dielectric modulation on hydrogen bonds. Similar techniques can be applied to other bonds as well. But this is an area where further innovation will be needed.

Although the book will present many positive results, it will also highlight many difficiencies in the field. We explain the lack of a model of hydrogen bonds, even though their importance in biology is unquestioned. Similarly, a nanoscale model of water is lacking, although the role of water in biology is pervasive and complex. Further developments at the level of basic physics and chemistry in areas like these will have a substantial benefit for the understanding of biological function.

This book is not a typical introduction to a well developed field in which all the main results are already standard. Rather it is an invitation to join in the study of the detailed mechanisms of protein interactions. We expect this to require many hands. Our intention here is to help stimulate in particular study of some more mathematical questions, many of which we leave open. To quote Mead and Conway [289], "And thus the period of exploration begins."

1.1 Exercises

Exercise 1.1 *Pour cooking oil into a glass of water and stir it vigorously until the oil is well dispersed. Now wait and watch as the oil droplets coalesce. Do the individual droplets retain any sort of discrete form? Or does the hydrophobic force just create a blob in the end?*

Exercise 1.2 *Acquire a pair of polarized sunglasses and observe objects just below the surface of a body of water both with and without the sunglasses. Do these observations while facing the sun, when it is at a low angle with respect to the water surface. You should observe that the 'glare' is greatly reduced by the polarizing lenses. Also make the same observations when the sun is overhead, and when looking in a direction away from the sun when it is at a low angle.*

Exercise 1.3 *Quantum-mechanical computations suffer from the 'curse of dimensionality' because each additional electron adds another three dimensions to the problem. Thus a problem with k electrons requires the solution of a partial differential equation in $\mathbb{R}^{3k}$. If we require a Cartesian discretization with m grid intervals per dimension, then the resulting problem requires m^{3k} words of memory to store the discrete representation. Compare this with the number of atoms in the observable universe. Assuming we could somehow make a computer using all of these atoms with each atom providing storage for one of the m^{3k} words of memory required for the discrete representation, determine how large a value of k could be used. Try values of $m = 3$ and $m = 10$.*

Exercise 1.4 *Pour salt into a glass of water and watch what happens to the salt. Take a small amount out and put it under a microscope to see if the picture stays the same.*

Exercise 1.5 *Pour salt into a glass of water and stir it until it dissolves. Now also add some oil to the water and stir it until small droplets form. Look at the surface of the oil droplets and see if you can see salt crystals that have reformed due to the change in electrostatic environment there. This might best be done on a slide beneath a microscope objective.*

Chapter 2

Digital rules for proteins

Digital rules are a hallmark of a mature science. This book attempts to present such rules for interactions involving proteins. We begin with a sketch of some of the main ideas that the book will cover. This is not an outline but rather is a narrative that introduces the main goals and challenges to be addressed, and gives a glimpse of some of the major advances.

We describe some challenging features of modeling the interactions of proteins in biological systems as well as opportunities to be addressed in the future. This is meant to provide some orientation, but it is also meant to be a disclaimer. That is, we disclose what we see as limitations of standard approaches which have forced us to adopt new strategies. There may well be other approaches that will be even more successful in the future.

2.1   Digital nature of molecules

We begin by illustrating what we mean by digital, or discrete, behavior in analog, or continuous, systems. This gives us an opportunity to review some basic concepts from chemistry. The basic entities of chemistry are molecules, and the building blocks of molecules are atoms. We begin by looking at digital rules for atoms, and then move to molecules.

2.1.1   Digital nature of atoms

Atoms can be characterized by the number of electrons, protons and neutrons of which they are composed. Some atoms of primary interest in protein biochemistry are listed in Table 2.1, and this table expresses digial rules at different scales: subatomic, atomic, and molecular. There are other subatomic digital rules that we will generally take for granted here, like Pauli's electron spin exclusion principle and Hund's rule about electron orbital occupation.

Several rules are encoded in Table 2.1. The first rule is used to reduce the number of columns: the number of protons always equals the number of electrons (the net charge is zero). A second rule is that the typical number of neutrons in the dominant isotope is nearly the same as the number of protons. But the most important rule is the **octet rule**: the number of the electrons in the outer shell plus the number (listed in the 'lacking' column) of electrons contributed by atoms covalently

bonded to it is always eight (except for hydrogen), up to Chlorine, and then it is eighteen for the larger atoms (for even larger atoms, the magic number is thirty-two, but those atoms do not concern us). This simple rule facilitates the determination of molecular bond formation.

Another rule contained in Table 2.1 is that the atomic mass is very close to the number of protons plus the number of neutrons (in the specified unit, the Dalton). The mass of the proton and neutron are approximately the same, and the rest mass of an electron is less than 0.0006 times this size. The units of atomic mass are discussed in Section 17.2.1, but for now the main point is that the mass of the proton and neutron are about one in the standard unit for atomic measurements, the Dalton. For reference, we list in Table 2.2 these masses in more familiar units.

As we see in Table 2.1, the number of neutrons can vary. We have listed what is known as the dominant isotope, followed (in parentheses) by the other possible stable isotopes (involving the listed numbers of neutrons). Neutrons add mass but not charge. Various isotopes are important in certain contexts; a hydrogen atom with an extra neutron is called deuterium. Atoms occur naturally in different isotopic forms, and the atomic mass reflects this natural variation. Otherwise, the atomic mass would be essentially the sum of the numbers of protons and neutrons, with a small correction for the electronic mass, as well as another correction that we will discuss shortly. For Chlorine, about a quarter of the atoms have 20 neutrons, and thus the atomic mass is about halfway between integer values. The given atomic masses are themselves only averages, and any particular set of atoms will vary in composition slightly; see the Periodic Table in [325] for more details.

We might expect that the atomic mass of a pure isotope would be given by

$$m \approx \mu(p, n) = p(m_p + m_e) + nm_n, \tag{2.1}$$

where m_p, m_e, and m_n are the mass of the proton, electron and neutron, respectively, and p is the number of protons (and electrons) and n is the number of neutrons. We list in Table 2.2 the masses of the proton, neutron and electron in familiar units (10^{-27} grams), as well as the standard unit of atomic mass, the Dalton, in these units. However, we see that for Carbon-12 (for which the number of neutrons is six), the formula (2.1) would predict that

$$m \approx \mu(6, 6) = 6(m_p + m_e + m_n) = 20.091 \times 10^{-27} \text{ grams} = 12.0989 \text{ Daltons.} \tag{2.2}$$

However, it turns out that the definition of the Dalton is exactly one twelfth of the mass of Carbon-12. Thus the mass of the component parts is greater than the mass of the atom. The difference in mass corresponds to a difference in energy ($E = mc^2$), and the atom represents a lower energy configuration than the separated constituents. For reference, we give in Table 2.3 the ratios between the measured atomic mass and the prediction in (2.1) for a few atoms for which there is only one stable isotope, in addition to the ratio for Carbon-12. See Exercise 2.1 regarding similar computations for atoms with more complex isotopic combinations. Note that the mass of the neutron is 1.0087 Dalton, and the mass of the proton is 1.0073 Dalton.

The digital description of an atom is to be contrasted with the analog description of the Schrödinger equation. This equation describes the electron distribution, which is the key determinant of atomic interaction. It is a continuum equation predicting the electronic distribution at all points in space, and there is a separate three-dimensional space at the least for each electron,

Atom	Symbol	+/−	neutrons	outer	lacking	ion	mass	radius
Hydrogen	H	1	0 (1)	1	1	+1	1.008	1.20
Carbon	C	6	6 (7)	4	4		12.01	1.70
Nitrogen	N	7	7 (8)	5	3		14.007	1.55
Oxygen	O	8	8 (9,10)	6	2	-2	15.9994	1.52
Fluorine	F	9	10	7	1	-1	18.998	1.47
Sodium	Na	11	12	1	7	+1	22.9898	2.27
Magnesium	Mg	12	12 (13,14)	2	6	+2	24.305	1.73
Phosphorus	P	15	16	5	3		30.974	1.80
Sulfur	S	16	16 (17,18,20)	6	2	-2	32.065	1.80
Chlorine	Cl	17	18 (20)	7	1	-1	35.4527	1.75
Potassium	K	19	20 (22)	1	17	+1	39.098	2.75
Calcium	Ca	20	20 (22-24)	2	16	+2	40.08	2.00
Iron	Fe	26	30 (28,31,32)	8	10	+3	55.845	1.10*
Copper	Cu	29	34 (36)	11	7	+1;2	63.55	1.40*
Zinc	Zn	30	34 (36-38,40)	12	6	+2	64.4	1.39*
Selenium	Se	34	46 (40,42-44)	16	2		78.96	1.90
Iodine	I	53	74	17	1	-1	126.90	1.98

Table 2.1: Subset of the periodic table. The column ' + /−' denotes the number of protons and electrons in the atom. The column 'outer' is the number of electrons in the outer shell. The column 'lacking' is the number of electrons needed to complete the outer shell. The column 'ion' gives the charge of the atomic ion (copper has two possible forms). The column 'mass' give the atomic mass in Daltons (see Chapter 17 for details), reflecting the naturally occurring isotopic distribution. The column 'radius' lists the 'mean' van der Waals radius [48], with the exceptions marked by *'s taken from various web sites.

proton	neutron	electron	Dalton
1672.622	1674.927	0.910938	1660.539

Table 2.2: Masses of atomic constituents, as well as the Dalton, listed in units of 10^{-27} grams. The Dalton is the standard unit of mass for atomic descriptions.

Hydrogen	Carbon-12	Fluorine	Sodium	Phosphorus	Iron	Iodine
1.000036	1.0083	1.0084	1.0087	1.0091	1.0095	1.0091

Table 2.3: Ratios of the atomic masses of different atoms to the mass predicted by formula (2.2). The isotopic fraction for Hydrogen was taken to be 0.015% Deuterium, and the isotopic fractions for Iron were taken to be (28) 5.8%, (30) 91.72%, (31) 2.2%, and (32) 0.28%, where the numbers in parentheses indicate the number of neutrons.

and in some cases for the protons as well. Even if it were simple to solve this equation (which it is not), it would be difficult to determine simple facts from such a representation. We are forced to consider effects on this level in many cases, but operating at the atomic level has clear advantages.

There are other simple rules in chemistry that allow prediction of bond formation, such as the electronegativity scale (Section 8.2.1) and the resonance principle (Section 13.1). The electronegativity scale allows the determination of polarity of molecules (Section 8.2). The resonance principle states that observed states of molecular bonds are often a simple convex combination of two elementary states. For example, a benzene ring can be thought of as being made of alternating single and double bonds, whereas in reality each bond is closely approximated by a convex combination of these two bonds. The resonance principle may be thought of as a Galerkin approximation to solutions of the Schrödinger equation.

The size data for atoms listed in the 'radius' column in Table 2.1 provides another way to distinguish between different atoms in a simple way, although the lengths do not correspond to a tangible boundary. If we could look at atoms at this level, the nucleus would be a tiny dot, and the electrons would be a fuzzy cloud, extending beyond the stated radius with a certain (nonzero) probability (cf. Figure 1 in [48]). Rather the length corresponds to the size of an 'exclusion zone' to give an idea of the size of a region where other atoms would not (typically) be found. A similar notion of length will be discussed in Section 5.4 regarding the size of atomic groups that form proteins. The length variation may not seem extreme, but the corresponding volume variation is over an order of magnitude.

For simplicity, Table 2.4 gives the relevant volumes for boxes of various sizes, ranging from 1.2Å to 2.8Å on a side. Thus in Table 2.1 we see that nitrogen and oxygen each have a volume twice that of hydrogen, and carbon has nearly three times the volume of hydrogen. Curiously, despite their similarity, potassium has a volume more than twice that of calcium. Moreover, the size relation in Table 2.1 seems to be going in the wrong direction. The size of atoms is a *decreasing* function of the number of electrons in the outer shell (for reference, the radius of Lithium is 1.82Å). When a shell becomes filled, the atom size jumps, but it then decreases as the new shell becomes populated. Of

r	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8
r^3	1.7	2.2	2.7	3.4	4.1	4.9	5.8	6.9	9.3	10.6	12.2	13.8	15.6	17.6	19.7	22.0

Table 2.4: Relation between volume and length in the range of lengths relevant for atoms.

course, when the number of electrons in the outer shell is increasing, so is the number of positive charges in the nucleus, and this means the nucleus has an increasing ability to pull the electrons closer.

We have described simple rules for atoms and their interactions both because they will be used extensively in the following but more importantly because they form a model of the type of rules we will attempt to establish for proteins. We will provide an introduction to proteins starting in Chapter 4, but we now give a description of the type of rules for protein interactions that we will establish.

2.1.2   Ionic rules

Certain ions are key to life in humans. There many important ionic effects that are specific to individual ions and are not easily explained by simple rules [245]. However, the charge of many ions can be deduced from a model that fits the data in Table 2.1 quite well for most, but not all, ions. The rule is that the atomic ion charge c is related to the number of electrons ℓ lacking in the outer shell and the total number of electrons O in a complete outer shell by

$$c = \begin{cases} O - \ell & \text{postive ions} \\ -\ell & \text{negative ions.} \end{cases} \tag{2.3}$$

Thus the rules simply measure the distance to the nearest complete shell. However these rules do not apply universally, and among the cases in Table 2.1 it works only when $|c| \leq 2$. Copper has two different ionic states; iron is an outlier. Nitrogen forms a complex ion N_3^-, and phosphorous forms an ionic complex with oxygen as well as with other atoms. Selenium forms different ionic complexes with oxygen.

2.1.3   Carbon/hydrogen rules

Another example of a simple rule at the molecular level relates to the common occurence of hydrogens bonded to carbons. These occur so commonly that they are often only implied in graphical representations. For example, a benzene ring might be written as a simple hexagon, without any labeling. The chemical formula for benzene is C_6H_6. Implicit in the graphical representation is that a carbon is located at each vertex of the hexagon, and that each carbon is bonded to a hydrogen. The rules apply to a wide variety of atoms, such as GAL and NAG (Section 4.7.3).
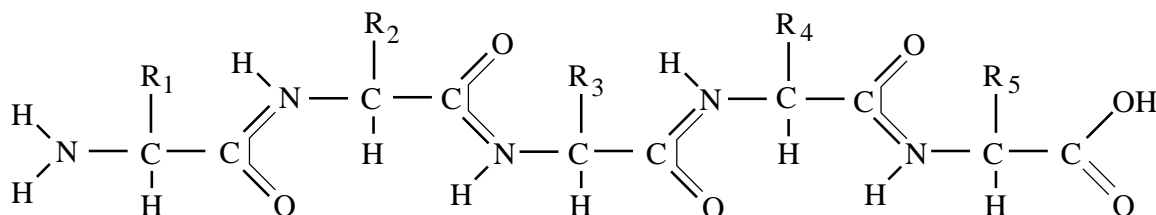
Figure 2.1: Schematic representation of a peptide sequence with five units (all in the trans conformation). The double bonds (double lines) between C and N denote the transition point between one peptide unit and the next. The residues are numbered in increasing order starting at the N-terminal end.

2.2 Digital nature of proteins

The digital and deterministic nature of protein function is implied by the fact that their structure is encoded by a discrete mechanism, DNA. This linear description is then translated into protein, a small fragment of which is depicted in Figure 2.1. But this linear representation of proteins belies the functional geometry of a protein, which is three dimensional. It is the latter representation that appears at first to be anything but a discrete widget. However, we will see that it is so. There are post-translational events (Section 4.3) which modify proteins and make their behavior more complex, but it is clear that nature works hard to make proteins in the same way every time.

What is striking about the fact that proteins act in quantized ways is the significant role played by hydrophobic effects (Section 2.5) in most protein-ligand interactions. Such interactions account not only for the formation of protein complexes, but also for signaling and enzymatic processes. But the hydrophobic effect is essentially nonspecific. Thus its role in a discrete system is intriguing.

We will see that it is possible to quantify the effect of hydrophobicity in discrete ways. The concept of *wrapping* (see Chapter 8) yields such a description, and we show that this can effect many important phenomena, including protein binding (Chapter 7) and the flexibility of the peptide bond (Chapter 13). In these two examples, we will see behaviors that are essentially digital in nature that can be predicted based on quantitative measures. In particular, both of these effects yield what can be viewed as switches, things that can be turned on and off. There are other examples of switches in protein behavior, and even the fundamental hydrogen bond can be viewed in this way to a certain extent.

We will also study other features of protein systems that can be described by simple quantized rules. We will consider different types of bonds that can be formed between proteins. These are all based on electronic interactions, and thus we will study extensively different types of electronic interactions from a mathematical point of view, including van der Waals forces.

The effect of electronic interactions can be substantially modified by the dielectric effect. This is modulated by hydrophobicity, so we now discuss the main concepts related to hydrophobicity.

2.2.1   Hydrophobicity and hydrophilicity

One major objective of the book is to clarify the effects of hydrophobicity in particular cases, including its role in protein-ligand interactions and other important phenomena. A primary effect of hydrophobicity is to modify the dielectric effect of water. Hydrophobic molecular groups can reduce the dielectric effect locally, which in turn enhances certain nearby bonds. Particular atomic groups in protein sidechains will be identified as being hydrophobic.

In one sense, hydrophilicity is the exact opposite of hydrophobicity. The latter means to repel water, and the former means to attract water. We will identify certain atom groups in protein sidechains as hydrophilic. However, one of the key points that we will emphasize is that hydrophilicity should not be thought of as a counterbalance to hydrophobicity. In particular, a hydrophilic group of atoms does not have a simple role of reversing the effect of hydrophobicity on the dielectric environment. Hydrophilic groups are always polar, which means that they represent an imbalance in charge. Thus they contribute to modifying the electric environment. The dielectric effect tends to dampen the effect of electric charges, so hydrophobic groups can have the effect of removing the damper on the charges of hydrophilic groups. Thus the effects of hydrophobicity and hydrophilicity are orthogonal in general and not opposite on some linear scale. In fact, hydrophobic groups can enhance the effect of a polar group, so they can correlate in just the opposite way from what the words seem to mean.

2.2.2   Solvation $\neq$ salvation

The life of a protein in water is largely a struggle for the survival of its hydrogen bonds. The hydrogen bond (cf. Chapter 6) is the primary determinant of the structure of proteins. But water molecules are readily available to replace the structural hydrogen bonds with hydrogen bonds to themselves; indeed this is a significant part of how proteins are broken down and recycled. We certainly cannot live without water [320], but proteins must struggle to live with it [123, 259].

Proteins are the fabric of life, playing diverse roles as building blocks, messengers, molecular machines, energy-providers, antagonists, and more. Proteins are initiated as a sequence of amino acids, forming a linear structure. They coil into a three-dimensional structure largely by forming hydrogen bonds. Without these bonds, there would be no structure, and there would be no function. The linear structure of amino acid sequences is entropically more favorable than the bound state, but the hydrogen bonds make the three-dimensional structure energetically favorable.

Water, often called the matrix of life [158], is one of the best makers of hydrogen bonds in nature. Each water molecule can form hydrogen bonds with four other molecules and frequently does so. Surprisingly, the exact bonding structure of liquid water is still under discussion [2, 377, 398, 417], but it is clear that water molecules can form complex bond structures with other water molecules. For example, water ice can take the form of a perfect lattice with all possible hydrogen bonds satisfied.

But water is equally happy to bind to available sites on proteins instead of bonding with other water molecules. The ends of certain side chains of amino acids look very much like water to a water molecule. But more importantly, the protein backbone hydrogen bonds can be replaced by

hydrogen bonds with water, and this can disrupt the protein structure. This can easily lead to the break-up of a protein if water is allowed to attack enough of the protein's hydrogen bonds.

The primary strategy for protecting hydrogen bonds is to bury them in the core of a protein. But this goes only so far, and inevitably there are hydrogen bonds formed near the surface of a protein. And our understanding of the role of proteins with extensive non-core regions is growing rapidly. The exposed hydrogen bonds are more potentially interactive with water. These are the ones that are most vulnerable to water attack.

Amino acids differ widely in the hydrophobic composition of their side chains (Section 4.1.1). Simply counting carbonaceous groups (e.g., CH_n for $n = 0, 1, 2$ or 3) in the side chains shows a striking range, from zero (glycine) to nine (tryptophan). Most of the carbonaceous groups are non-polar and thus hydrophobic. Having the right amino acid side chains surrounding, or **wrapping**, an exposed hydrogen bond can lead to the exclusion of water, and having the wrong ones can make the bond very vulnerable. The concept of wrapping an electrostatic bond by nonpolar groups is analogous to wrapping live electrical wires by non-conducting tape.

We refer to the under-protected hydrogen bonds, which are not sufficiently wrapped by carbonaceous groups, as **dehydrons** (Section 2.5.2) to simplify terminology. The name derives from the fact that these hydrogen bonds benefit energetically from being dehydrated.

2.2.3   Energetic ambivalence

One could imagine a world in which all hydrogen bonds were fully protected. However, this would be a very rigid world. Biology appears to prefer to live at the edge of stability. Thus it is not surprising that new modes of interactions would become more prevalent in biology than in other areas of physics. For example, it has been recently observed that exposed hydrogen bonds appear to be sites of protein-protein interactions [142]. Thus what at first appears to be a weakness in proteins is in fact an opportunity.

One could define an **epidiorthotric force** as one that is associated with the repair of defects. The grain of sand in an oyster that leads to a pearl can be described as an epidiorthotric stimulant. Similarly, snow flakes and rain drops tend to form around small specs of dust. Such forces also have analogies in personal, social and political interactions where forces based on detrimental circumstances cause a beneficial outcome. A couple who stay together because they do not want to be alone provides such an example. The defect of an under-protected hydrogen bond gives rise to just such an epidiorthotric force. The action of this force is indirect, so it takes some explaining.

An under-protected hydrogen bond would be much stronger if water were removed from its vicinity. The benefit can be understood first by saying that it is the result of removing a threat of attack (or the intermittent encounter of water forming hydrogen bonds with it). But there is an even more subtle (but mathematically quantifiable) effect due to the change in dielectric environment when water is removed, or even just structured, in the neighborhood. The dielectric constant of water is about eighty times that of the vacuum. Changing the dielectric environment near an under-protected hydrogen bond makes the bond substantially stronger.

If the removal of water from an under-protected hydrogen bond is energetically favorable, then this means there is a force associated with attracting something that would exclude water. Indeed,

one can measure such a force, and it agrees with what would be predicted by calculating the change energy due to the change in dielectric (Section 9.1). You can think of this force as being somewhat like the way that adhesive tape works. Part of the force results from the removal of air between the tape and the surface, leaving atmospheric pressure holding it on. However, the analogy only goes so far in that there is an enhancement of electrical energy associated with the removal of water. For sticky tape, this would correspond to increasing the mass of the air molecules in the vicinity of the tape, by a factor of 80, without increasing their volume!

Thus the epidiorthotric force associated with water-removal from an under-protected hydrogen bond provides a mechanism to bind proteins together. This is a particular type of hydrophobic effect, because wrapping the bond with hydrophobic groups provides protection from water. It is intriguing that it arises from a defect which provides an opportunity to interact.

2.3 Pchemomics

The term "omics" refers to the use of biological data-bases to extract new knowledge by large-scale statistical surveys. The term "cheminformatics" is an accepted moniker for the interaction of informatics and chemistry, so there is some precedent for combining terms like pchem (a.k.a., physical chemistry) with a term like 'omics.' We do not suggest the adoption of the (unpronounceable) term pchemomics, but it serves to suggest the particular techniques being combined in a unique way. An example of pchemomics is the early study of the hydrogen bond [242]. Indeed, the original study of the structure of the peptide bond (see section 8-4 of [324]) used such an approach. But pchemomics involves a two way interaction with data. In addition to providing a way to learn new properties in physical chemistry, it also involves using physical chemistry to look at standard data in new ways.

The Protein Data-Base (**PDB**) provides three-dimensional structures that yield continuing opportunities for proteomics discoveries. Using the perspective of physical chemistry in datamining in the PDB, some simple laws about protein families were determined by studying patterns of under-wrapped hydrogen bonds [130]. We examine just one such result in Section 2.3.4; many other results in physical chemistry can be likewise explored.

A simple view of the PDB only gives a representation suitable for Lagrangian mechanics (or perhaps just statics). If we keep in mind which atom groups are charged, we begin to see an electrostatic view of proteins, and standard protein viewers will highlight the differently charged groups. But the dielectric effect of the solvent is left to the imagination. And the crucial role of the modulation of the dielectric effect by hydrophobic groups is also missing. Adding such views of proteins involve a type of physical chemistry lens.

When you do look at proteins by considering the effect of wrapping by hydrophobic groups, you see many new things that may be interpreted in ways that are common in bioinformatics. One striking observation is that there is a simple correlation between the number of under-wrapped hydrogen bonds and evolutionary trends. Figure 2.2 depicts the number of dehydrons found in the protein myoglobin (or its analog) in various species [142]. Similar trends are seen with other proteins in Table 11.3.

The number of under-wrapped hydrogen bonds appears to be evolving (increasingly), providing increasing opportunities for interaction in advanced species. This provides additional understanding
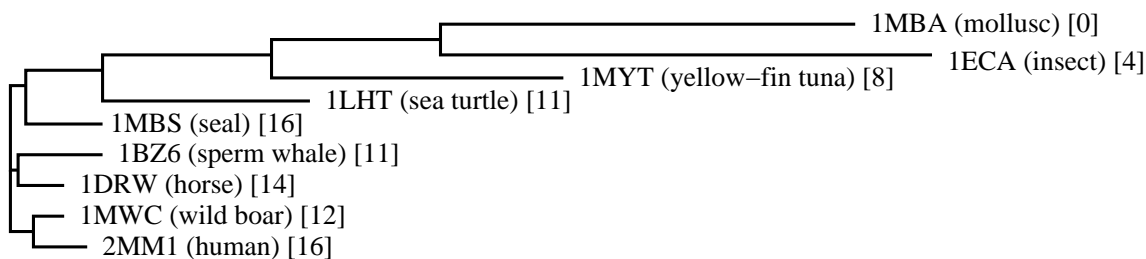
Figure 2.2: Number of dehydrons [shown in square brackets] in the protein myoglobin found in various species [142], which are presented in an evolutionary tree determined by sequence alignment distances.

of how higher species may have differentiated function without dramatically increasing the number of genes which code for proteins.

It is also significant that under-wrapped hydrogen bonds appear to be conserved more than other parts of proteins. But since the number of under-wrapped hydrogen bonds is growing, we should say that once they appear they tend to be conserved [142].

Given our understanding of what it means to be under-wrapped, it is not surprising that under-wrapped hydrogen bonds would appear more often in regions of proteins that are themselves not well structured. NORS (NO Recognizable Structure) regions [203] in proteins are large (at least seventy consecutive amino acids) sections which form neither α-helices or β-sheets. These appear more frequently among interactive proteins. Correspondingly, studies [145] have shown a strong correlation between the number of under-wrapped hydrogen bonds and interactivity.

A full understanding of wrapping and the related force associated with under-wrapping requires tools from physical chemistry. Interactions between physical chemistry and "omics" will offer further insights into biological systems. Indeed, precise modeling of water even by explicit solvent methods is still a challenge. Only recently have models begun to predict the temperature behavior of the density of liquid water [263]. This means that for very subtle issues one must still be careful about even all-atom simulations. The mysteries of water continue to confront us. But its role in biology will always be central.

2.3.1   A new tool?

Since we are seeking to answer new types of research questions, it may be comforting to know that there is a powerful tool that is being used. The combination of data mining and physical chemistry is not new, but its usefulness is far from exhausted. Moreover, it is not so common to see these utilized in conjunction with more conventional techniques of applied mathematics, as we do here. Thus we take a moment to reflect on the foundations of the basic concepts that make up what we refer to as pchemomics.

Typical datamining in bioinformatics uses more discrete information, whereas the PDB uses continuous variables to encode chemical properties. The need for physical chemistry in biology has long been recognized. In the book [393], the following quote is featured:

> The exact and definite determination of life phenomena which are common to plants and animals is only one side of the physiological problem of today. The other side is the construction of a mental picture of the constitution of living matter from these general qualities. In the portion of our work *we need the aid of physical chemistry.*

The emphasis at the end was added as an aid to the eye. These words were written by Jacques Loeb in "The biological problems of today: physiology" which appeared in the journal *Science* in volume 7, pages 154–156, in 1897. So our theme is not so new, but the domain of physical chemistry has advanced substantially in the last century, so there continues to be an important role for it to play in modern biology.

2.3.2 Data mining definition

It is useful to reflect on the nature of **data mining**, since this is a relatively new term. It is a term from the information age, so it is suitable to look for a definition on the Web. According to WHATIS.COM,

> Data mining is sorting through data to identify patterns and establish relationships.
>
> Data mining parameters include:
>
> - Association - $\boxed{\text{looking}}$ for patterns where one event is connected to another event
> - Sequence or path analysis - $\boxed{\text{looking}}$ for patterns where one event leads to another later event
> - Classification - $\boxed{\text{looking}}$ for new patterns (May result in a change in the way the data is organized but that's ok)
> - Clustering - finding and $\boxed{\text{visually}}$ documenting groups of facts not previously known

Our conclusion? Data mining involves **looking** at data (the $\boxed{\text{boxes}}$ have been added for emphasis). If data mining is looking at data, then **what type of lens do we use?**

2.3.3 Data mining lens

There are many ways to look at the same biological data. In the field of data mining, this might be called using different *filters* on the data. However, it is not common to look at the same data with many different filters, so we prefer the different metaphor of a lens. It could be a telescope, a microscope, polarized sunglasses, or just a good pair of reading glasses.

All proteins have chemical representations, e.g., the protein

$$C_{400}H_{620}N_{100}O_{120}P_1S_1.$$
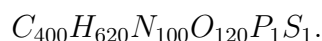
In the early research on proteins [393], discovering such formulæ was a major step. But a much bigger step came with the realization that proteins are composed of sequences of amino acids. This

allowed proteins to be described by alphabetic sequences, and the descriptions come in different forms: DNA, RNA, amino acid sequences. One can think of these from a linguistic perspective, and indeed this has been a productive approach [212].

The function of DNA is largely to store sequence information, but proteins operate as three-dimensional widgets. Not all proteins have a stable three-dimensional representation, but most biologically relevant proteins function via three-dimensional structures. Indeed, random proteins would be expected not to form stable three-dimensional structures [102]. The PDB is a curated database of such structures that provides a starting point to study protein function from a physical chemistry perspective.

But structure alone does not explain how proteins function. Physical chemistry can both simplify our picture of a protein and also allow function to be more easily interpreted. In particular, we will emphasize the role of the modulation of the dielectric environment by hydrophobic effects. We describe a simple way this can be done to illustrate the effect on individual electronic entities, such as bonds. But there is need for better lenses to look at such complex effects.

2.3.4   Hydrogen bonds are orientation-dependent

The hydrogen bond provides a good starting example of the use of "pchem" data mining to reveal its properties. Figure 6 of [242] shows clearly both the radial and the angular dependence of the hydrogen bond. Similar evidence is found in later papers; Figure 3 in [394] suggests that hydrogen bonds are stronger when they are both shorter and better aligned. However, the precise relationships between angle and distance can depend on the context, being different in different types of protein structure [30]. Figure 8 of [430] shows a similar relationship between the angle of the hydrogen bond and its distance, derived using protein data. The data in that figure is consistent with a conical restriction on the region of influence of the bond. More recently, the orientation dependence of the hydrogen bond has been revisited. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes [300]. All of these invovled datamining.

An alternative method for modeling hydrogen bonds is to study their energetics via quantum mechanical calculations and to interpolate the resulting energy surfaces [309, 373]. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations has also been reported [238]. Despite the inherent interest in hydrogen bonds, a general model of them has not yet been developed. In particular, hydrogen bonds do not appear as primary bonds in molecular dynamics simulations. Due to the primary importance of the hydrogen bond in protein structure, we will review what is known and not known in Chapter 6.

2.3.5   What is an answer?

Before we begin to ask questions in earnest, we need to talk about what sort of answers we might expect. In high-school algebra, an answer takes the form of a number, or a small set of numbers. In calculus, the answer is often a function. Here, we will often find that the answer is statistical in nature. There appear to be few absolutes in biology, so a probability distribution of what to expect

is the best we can hope for.

A probability distribution provides a way to give answers that combine the types of answers you get with high-school algebra and those you get with calculus. An answer that is a number is a Dirac δ-function, whereas a function corresponds to a measure that is absolutely continuous. This added level of sophistication is especially helpful in a subject where it seems almost anything can happen with some degree of probability.

Mathematics tells us that it is a good idea to have metrics for the space of answers that we expect. Metrics on probability distributions are not commonly discussed. We suggest in Exercise 7.2 one application of metrics on probability distributions to our subject.

In classical physics, problems were often considered solved only when names for the functions involved could be determined. This paradigm is extremely robust and useful. When the names are familiar, they suggest general properties (exponential versus sinusoidal), and they provide a simple algorithm to compute specific values for particular instances. The programming language Fortran was designed specifically to facilitate the evaluation of expressions such as

$$\sin(\log(\tan(\cos(J_1(e^x + \sqrt{\pi x}))))). \tag{2.4}$$

Unfortunately, the classical paradigm is limited by our ability to absorb new names. While the names in (2.4) are familiar to many who have studied Calculus, the list required in practice includes less well known Bessel functions, Hankel functions, elliptic functions, theta functions, zeta functions, and so on. Moreover, it may be that each new problem requires a new name, in which case the paradigm fails; it is only successful if it provides an abstraction that allows the simplification of the answer. Moreover, strict adherence to this paradigm causes an unnecessary impediment from a computational point of view. All that we may care about is the asymptotic form of a function, or particular values in a certain range, i.e., a plot, or just the point at which it has a minimum.

The newer computational paradigm is not to associate names to solutions, but rather to associate standard algorithms to problems that can be used to provide the information required to understand the mechanism being studied. For example, we may be content if we can specify a well-posed differential equation to be solved to determine numerical values of a function. Thus we might say that the equation $u' = u$ is a sufficient description of the exponential function. When we discuss quantum mechanics, we will adopt this point of view.

2.4   Multiscale models

But why don't we just write down a mathematical model and use it to simulate protein dynamics? This is a reasonable question, and we attempt here to show why such an approach at the moment would not be productive. The difficulty is the particular multiscale aspect of the problem: the temporal scales are huge but the spatial scales overlap, as depicted in Figure 2.3. Of course, existing models are useful in limited contexts. However, we will explain limitations in two such models that must be addressed in order to use them on more challenging simulations.

Models for many systems have components which operate at different scales [198]. Scale separation often simplifies the interactions among the different scales. The differences often occur in
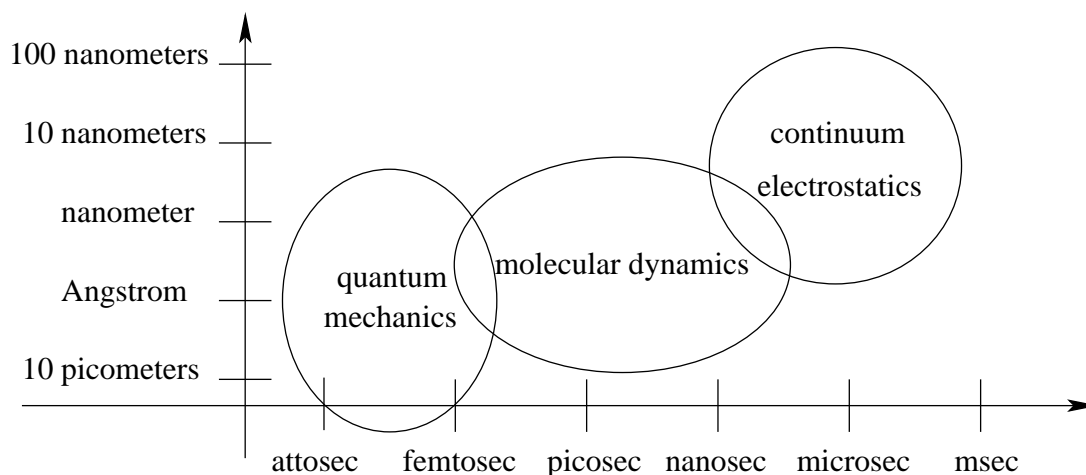
Figure 2.3: Cartoon of spatial and temporal scales of biomolecular models. See the text for more details.

both spatial and temporal scales. Scale separation often simplifies the study of complex systems by allowing each scale to be studied independently, with only weak interactions among the different scales. However, when there is a lack of scale separation, interactions among the scales become more difficult to model.

There are three models of importance in protein biochemistry. The different spatial and temporal scales for these models are depicted in Figure 2.3. The smallest and fastest scale is that of quantum chemistry. The model involves continuous variables, partial differential equations and functions as solutions.

The molecular scale is more discrete, described only by the positions of different atoms in space, perhaps as a function of time. The time scale of molecular dynamics is much longer than the quantum scale. But the length scale is comparable with the quantum scale. For example, the Ångstrom can be used effectively to describe both without involving very large or very small numbers.

Finally, the electric properties of proteins are mediated by the dielectric behavior of water in a way that is suitable for a continuum model [94, 365]. But again the length scale is not much bigger than the molecular scale. Many solvated systems are accurately represented using a system in which the size of the solvation layer is the same as the protein dimension. On the other hand, dielectric models are inherently time independent, representing a 'mean field' approximation. Thus there is no natural time scale for the continuum dielectric model, but we have depicted in Figure 2.3 the time scale for so-called Brownian dynamics models which are based on a continuum dielectric model [278].

The lack of physical scale separation, linked with the extreme time separation, in biological systems is the root of some of the key challenges in modeling them. Note that the temporal scales in Figure 2.3 cover fifteen orders of magnitude whereas the spatial scales cover only three or four orders of magnitude. Many biological effects take place over a time scale measured in seconds, but there may be key ingredients which are determined at a quantum level. This makes it imperative

to develop simplified rules of engagement to help sort out behaviors, as we attempt to do here.

We do not give a complete introduction to quantum models, but we do include some material so that we can discuss some relevant issues of interest. For example, molecular-level models utilize force fields that can be determined from quantum models, and this is an area where we can predict significant developments in the future. The hydration structure around certain amino acid residues is complex and something that begs further study. But this may require water models which are currently under development, and these models may require further examination at the quantum level.

Multi-scale models are most interesting and challenging when there is significant information flow between levels. One of the most intriguing examples is the effect of the electric field on the flexibility of the peptide bond [122]. The electric field is determined by effects at the largest scale and causes a change in behaviors at the smallest scale, forcing a re-structuring of the molecular model (Chapter 13).

The Schrödinger equation is a well-accepted model for quantum chemistry. However, it is too detailed for use as a numerical model for large systems. Molecular dynamics models are used routinely to simulate protein dynamics, but there are two drawbacks. On the one hand, there are some limitations in the basic theoretical foundations of the model, such as the proper force fields to be used, so the predictions may not be fully accurate (cf. Section 13.4). On the other hand, they are still complicated enough that sufficiently long-time simulations, required for biological accuracy, are often prohibitive [12]. Electrostatic models hope to capture the expected impact of dielectric solvation, but there are limitations here as well. The dielectric coefficient of water is orders of magnitude larger than what would be found inside a large protein. This is a very large jump in a coefficient in a continuum model, and it is prudent to be cautious about any model with such large changes. It is clear that in the neighborhood of the jump in the coefficient, a more complex model might be required [365].

We can anticipate improvements to the models used for biochemical simulation, and we hope that these will contribute to improved computational techniques in molecular biology. In addition to improvements in models at the various scales, we also anticipate advances in linking models at different scales. While this is an extremely difficult problem, significant advances are being made [103, 416].

2.5 Hydrophobic interactions

Hydrophobic interactions are sometimes said to be more important than even the hydrogen bond [221]. Although not completely understood, the **hydrophobic force** [41] derives from the **hydrophobic effect** [392]. This effect is one of the central topics of our study. However, the hydrophobic effect has many manifestations in protein behavior.

The hydrophobic effect [41, 392] was proposed in the 1950's [393] as a major contributor to protein structure. However, it is only recently that the detailed nature of hydrophobic forces have been understood. Indeed, the dehydron can be viewed as a particular type of hydrophobic effect.

In two recent papers, further understanding of hydrophobic forces have been provided [99, 100]. It was seen that the role of hydrophobic modulation of solvent dielectric is critical to the hydrophobic
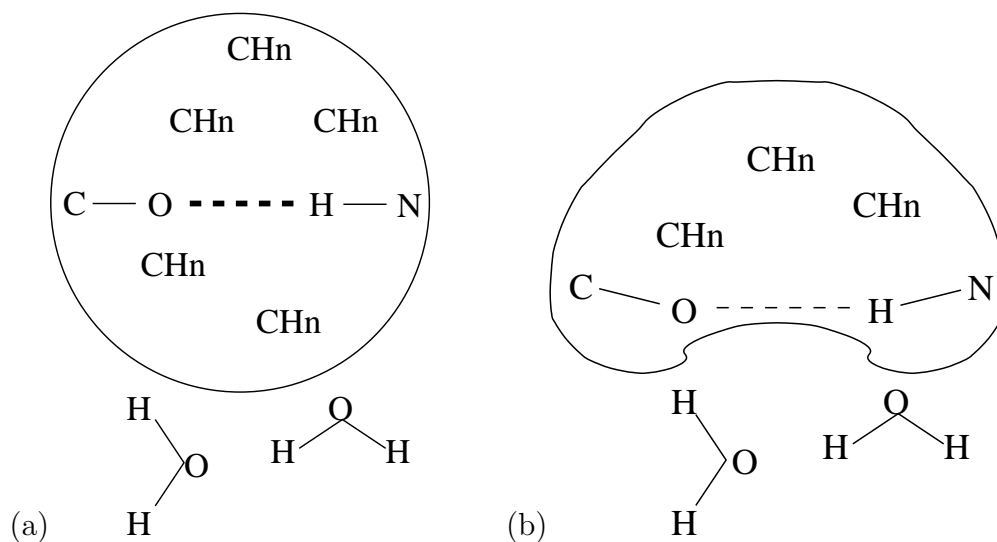
Figure 2.4: (a) Well wrapped hydrogen bond (b) Underwrapped hydrogen bond.

force [99].

There is a simple view of how hydrophobic forces work. There are certain molecules that are hydrophobic (cf. Section 2.5.2 and Chapter 8), meaning that they repel water. Regions of proteins that have many such molecules, e.g., a protein with a large number of hydrophobic residues on a part of its surface, would tend to prefer association with another such surface to reduce the frustration of having two water-hydrophobe interfaces. It is this simple effect that makes cooking oil form a single blob in water even after it has been dispersed by vigorous stirring.

More precisely, the argument is that the elimination of two hydrophobic surfaces with a water interface is energetically favorable. One could also argue by considering volume changes (cf. Section 5.4) since hydrophobic side chains take up more volume in water. Recent results show how a hydrophobic force can arise through a complex interaction between polarizable (e.g., hydrophobic) molecules and (polar) water molecules [99, 100]. These arguments are compelling, but they suggest a nonspecific interaction. Indeed, hydrophobic attraction leads to nonspecific binding [144].

But there are other kinds of hydrophobic effects as well. We will show that hydrophobicity plays a central role in a number of electrostatic forces by modulating the dielectric effect of water. In addition, water removal can affect the local polar environment, which can modify the nature of covalent bonds.

2.5.1 Solvent mediation of electric forces

Some bonds become substantially altered in the presence of water. We have already noted that certain ionic bonds (in table salt) are easily disrupted by water. The main bond holding proteins together is the hydrogen bond, and this bond is extremely susceptible to alteration by water interaction since water molecules can each make four hydrogen bonds themselves. So protein survival depends on keeping the hydrogen bond dry in water [123].
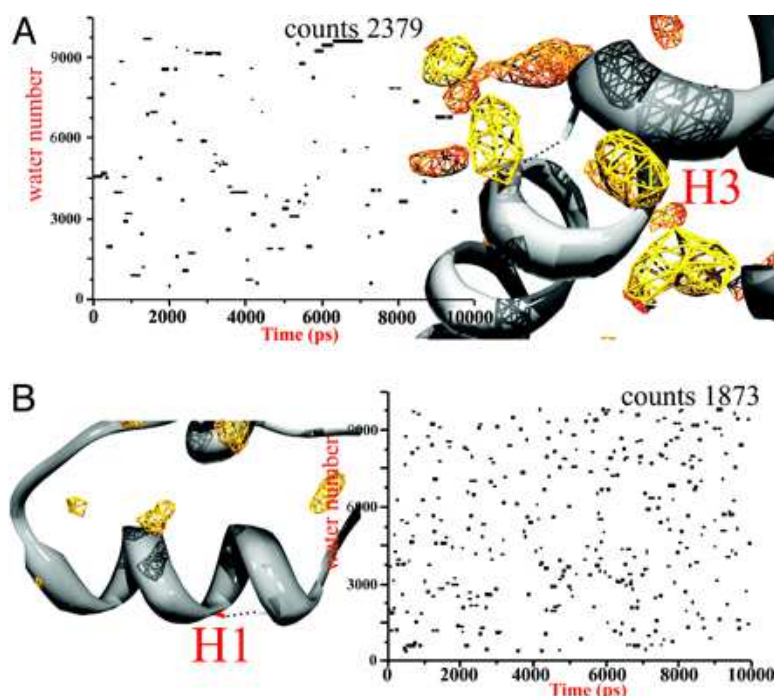
Figure 2.5: Dynamics of water near hydrogen bonds, reproduced from Fig. 5 in [91]. (A) Hydrogen bond (H3) is well wrapped. (B) Hydrogen bond (H1) is underwrapped.

Another type of solvent effect that occurs on the quantum level is the rigidity of the peptide bond (Chapter 13) which requires an external field to select one of two resonant states. Such a field can be due to hydrogen bonds (see Section 5.1 and Chapter 6) formed by backbone amide or carbonyl groups, either with other backbone or sidechain groups, or with water. In some situtations, water removal can cause a switch in the resonance state to a flexible mode [122].

Another example of a change of electrical properties resulting from differences in the water environment involves a more gross change. Proteins which penetrate a cell membrane go from a fully solvated environment to one that is largely solvent-free (inside the membrane). This can be related to a large-scale change in the secondary structure of the protein conformation that has implications for drug delivery [144].

More generally, solvent mediation can alter any electrostatic force via dielectric effects (Chapter 16). Changes in dielectric properties of the environment can have a substantial impact on any electrical property. Much of our study will be related to the dielectric effect and its modulation by hydrophobic groups. But rather than try to address this by introducing a precise model (see Chapter 16), we prefer to introduce the concept by example. We thus begin by looking at one particular example of hydrophobic modulation of the dielectric behavior of water around hydrogen bonds.

2.5.2   Dehydrons

In [142], a quantifiable structural motif, called **dehydron**, was shown to be central to protein-ligand interactions. A dehydron is a defectively 'wrapped' hydrogen bond in a molecular structure whose electrostatic energy is highly sensitive to water exclusion by a third party. Such pre-formed, but underprotected, hydrogen bonds are effectively adhesive, since water removal from their vicinity contributes to their strength and stability, and thus they attract partners that make them more viable (see Section 2.5.5 and Chapter 9).

A review of protein structure and the role of hydrogen bonds will be presented in Chapter 4. The concept of 'wrapping' of a hydrogen bond is based on the hydrophobic effect [41, 392], and its role in modulating the dielectric effect (Chapter 16). At the simplest level, wrapping occurs when sufficient nonpolar groups (CH_n, $n = 1, 2, 3$) are clustered in the vicinity of intramolecular hydrogen bonds, protecting them by excluding surrounding water [140]. The concept of wrapping of a hydrogen bond is depicted informally in Figure 2.4. A well wrapped hydrogen bond, Figure 2.4(a), is surrounded by CH_n groups on all sides, and water is kept away from the hydrogen bond formed between the C-O group of one peptide and the N-H group of another peptide (Section 4.1). An underwrapped hydrogen bond, Figure 2.4(b), allows a closer approach by water to the hydrogen bond, and this tends to disrupt the bond, allowing the distance between the groups to increase and the bond to weaken.

It is possible to identify dehydrons as **under wrapped hydrogen bonds** (**UWHB**) by simply counting the number of hydrophobic side chains in the vicinity of a hydrogen bond. This approach is reviewed in Section 8.3. More accurately, a count of all (nonpolar) carbonaceous groups gives a more refined estimate (Section 8.4). However, it is possible to go further and quantify a force associated with dehydrons which provides a more refined measure of the effect geometry [142] of the wrappers (Section 8.5).

We have already seen in Figure 2.2 that dehydrons are a sensitive measure of protein differences. At the structural level, a significant correlation can be established between dehydrons and sites for protein complexation (Chapter 8). The HIV-1 capsid protein P24 complexed with antibody FAB25.3 provides a dramatic example, as shown in Figure 2 of [142] and in a cartoon in Figure 11.1.

2.5.3   Dynamics of dehydrons

The extent of wrapping changes the nature of hydrogen bond [91] and the structure of nearby water [133]. Hydrogen bonds that are not protected from water do not persist [91]. Figure 5 of [91] shows the striking difference of water residence times for well wrapped and underwrapped hydrogen bonds. Private communication with the authors of [91] have confirmed that there is a marked difference as well in the fluctuations of the hydrogen bonds themselves. Under wrapped hydrogen bond lengths are larger (on average) than well wrapped hydrogen bonds. More strikingly, the distributions of bond lengths as shown in Figure 2.6 are quite different, confirming our prediction based on Figure 2.4 that the coupling of the hydrogen bond characteristics with the water environment would be different.

The H-bond R208–E212 depicted in Fig. 5(A) [91] is well wrapped whereas V189–T193 depicted in Fig. 5(B) is a dehydron (see Fig 3a in [135] page 6448). Well-wrapped hydrogen bonds are visited by fewer water molecules but have longer-lasting water interactions (due to the structuring effect
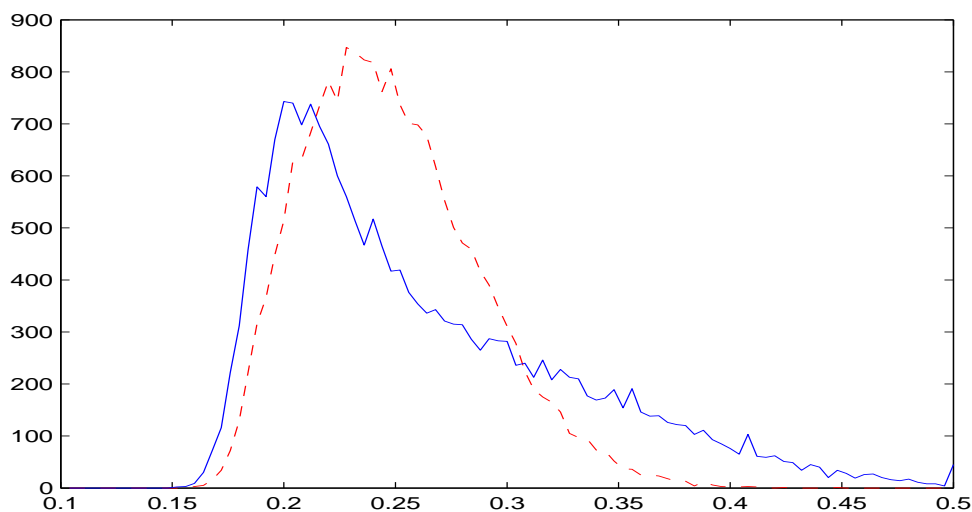
Figure 2.6: Distribution of bond lengths for two hydrogen bonds formed in a structure of the sheep prion [91]. The horizontal axis is measured in nanometers, whereas the vertical axis represents numbers of occurrences taken from a simulation with 20,000 data points with bin widths of 0.1 Ångstrom. The distribution for the well-wrapped hydrogen bond (H3) has a smaller mean value but a longer (exponential) tail, whereas the distribution for the underwrapped hydrogen bond (H1) has a larger mean but Gaussian tail.
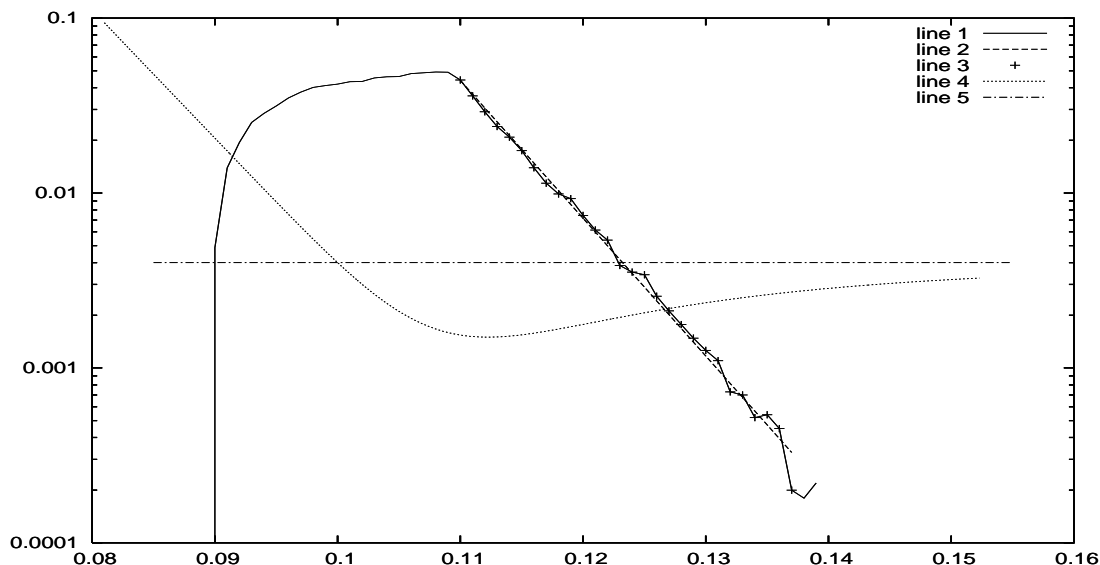
Figure 2.7: Simulation of a random walk with a restoring force. Shown is the distribution of values x_i defined in (2.5) for 10^5 time steps i, starting with $x_1 = 0.1$, scaled by a factor of 10^{-3}. Also shown is a graph of $\phi + 0.03$ where ϕ is the potential (2.5). The dot-dashed horizontal line provides a reference axis to facilitate seeing where ϕ is positive and negative. The $+$'s indicate the part of the distribution exhibiting an exponential decay; the dashed line is a least-squares fit to the logarithm of these distribution values. The distribution has been scaled by a factor of 10^{-3} so that it fits on the same plot with ϕ.

of the hydrophobes), whereas the behavior of dehydrons is more like that of bulk water: frequent re-bonding with different water molecules [91].

The long residence time of waters around a well-wrapped hydrogen bond would seem to have two contributing factors. On the one hand, the water environment is structured by the hydrophobic barrier, so the waters have reduced options for mobility: once trapped they tend to stay. But also, the polar effect of the hydrogen bond which attracts the water is more stable, thus making the attraction of water more stable. With a dehydron, both of these effects go in the opposite direction. First of all, water is more free to move in the direction of the hydrogen bond. Secondly, the fluctuation of the amide and carbonyls comprising the hydrogen bond contribute to a fluctuating electrostatic environment. The bond can switch from the state depicted in Figure 2.4(b) when water is near, to one more like that depicted in Figure 2.4(a) if water molecules move temporarily away. More precisely, the interaction of the bond strength and the local water environment becomes a strongly coupled system for an underwrapped hydrogen bond, leading to increased fluctuations. For a well wrapped hydrogen bonds, the bond strength and water environment are less strongly coupled.

The distance distribution for under-wrapped hydrogen bonds can be interpreted as reflecting a strong coupling with the thermal fluctuations of the solvent. Thus we see a Boltzmann-type distribution for the under-wrapped hydrogen bond distances in Figure 2.6. It is natural to expect
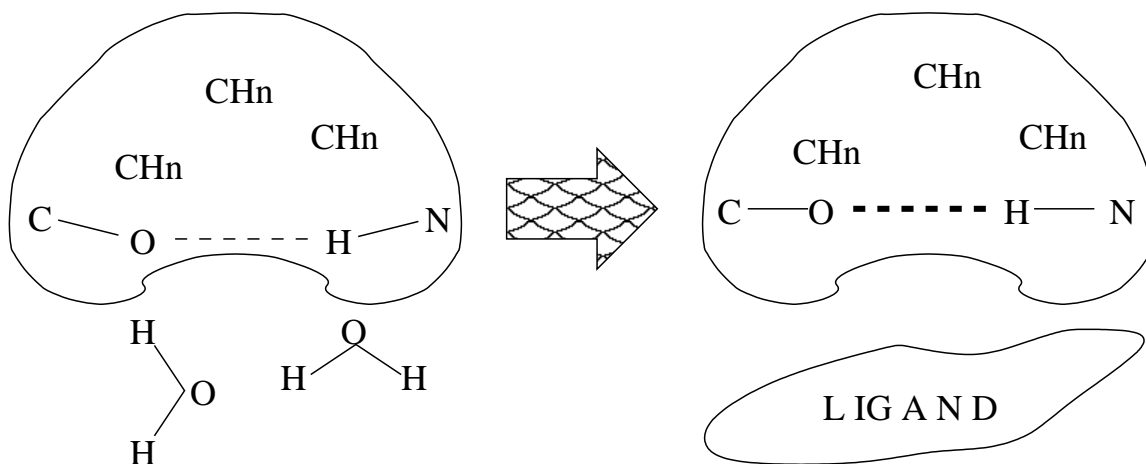
Figure 2.8: Cartoon showing dehydration due to ligand binding and the resulting strengthening of an underwrapped hydrogen bond.

the mean distances in this case to be larger than the mean distances for the underwrapped case, but the tails of the distribution are at first more confusing. The distribution in the underwrapped case exhibit a Gaussian-like tail (that is, exponential of the distance squared), whereas the well-wrapped case decays more slowly, like a simple exponential. (See Figure 5.13 for a comparison of these distributions.) Thus the well-wrapped hydrogen bond is sustaining much larger deviations, even though the typical deviation is much smaller than in the underwrapped case. To explain how this might occur, we turn to a simulation with a simple model.

2.5.4   Simulated dynamics

The data in Figure 2.6 can be interpreted via a simulation which is depicted in Figure 2.7. This figure records the distribution of positions for a random walk subject to a restoring force defined by

$$x_{i+1} = x_i + \Delta t(f_i + \phi(x_i)) \tag{2.5}$$

with f_i drawn randomly from a uniform distribution on $[-0.5, 0.5]$, and with ϕ being a standard Lennard-Jones potential

$$\phi(x) = (0.1/x)^{12} - (0.1/x)^6. \tag{2.6}$$

The particular time step used in Figure 2.7 is $\Delta t = 0.02$; the simulation was initiated with $x_1 = 0.1$ and carried out for 10^5 steps.

  The simulation (2.5) represents a system that is forced randomly with a restoring force back to the stationary point $x = 0.1$, quantified by the potential ϕ in (2.6). Such a system exhibits a distribution with an exponential decay, as verified in Figure 2.7 by comparison with a least-squares fit of the logarithm of the data to a straight line.

2.5.5 Stickiness of dehydrons

Desolvation of an underwrapped hydrogen bond can occur when a ligand binds nearby, as depicted in Figure 2.8. The removal of water lowers the dielectric and correspondingly strengthens the hydrogen bond. The resulting change in energy due to the binding effectively means that there is a mechanical force of attraction for a dehydron. This applies that dehydrons are sticky. These concepts are explained in more detail in Chapter 9, and the experimental support for the force of attraction for dehydrons is reviewed as well.

Since dehydrons are sticky, they must be binding sites. Thus we see in Chapter 7 that they are likely to be found at protein-ligand interfaces. In Chapter 11 we explore several types of protein-ligand interfaces in which dehydrons play a decisive role. Finally, we review in Chapter 14 the role of dehydrons in drug design.

2.6 Dehydron switch

The strength and stability of hydrogen bonds depend on many factors: the distances between donor and acceptor and other constituents, the angles formed by the constituents, and the local dielectric environment. While we cannot formally quantify the effect of these factors, we can imagine that they combine to form a single variable that describes the 'quality' of the hydrogen bond. Then the stability and strength of the hydrogen bond depend in some monotonic way on this quality variable.

As wrapping is varied, the quality of a hydrogen bond can vary in a highly nonlinear way, as depicted in Figure 2.9. When a hydrogen bond is extremely underwrapped, the addition of one wrapper could have only a small effect, especially if it appears in a region already wrapped. At the other extreme, additional wrappers will have diminishing effects on a fully protected hydrogen bond. In the middle, additional wrappers may have the most effect, leading to a sensitive switch.

Binding events can cause the quality of a hydrogen bond to change significantly, due to several factors. Wrapping can increase the quality, and unwrapping can

Figure 2.9: Cartoon depicting the capability of a dehydron to switch on and off when a ligand binds. The binding event may increase wrapping, in which case the hydrogen bond becomes stronger and more stable, or it may decrease wrapping, in which case the hydrogen bond becomes weaker and less stable.

decrease the quality. In addition, large scale structural changes can occur that may alter the distances and angles that determine the quality of a hydrogen bond. However, even without such gross motion in a protein, the change in the dielectric environment (which has the capability to change by two orders of magnitude) can effect the quality significantly. The former increases the enthalpy of the hydrogen bond, and thus causes dehydrons to be sticky in many cases.

Unwrapping a dehydron can effectively switch off the hydrogen bond, eliminating a constraint on the protein. This has the potential to increase the entropy of the overall system because removing the constraint can increase the degrees of freedom of motion. Thus both effects have the potential to increase free energy upon association.

The variables that determine the quality of a hydrogen bond are not independent. As a hydrogen bond becomes underwrapped, it becomes weaker due just to the change in local dielectric, and this can mean that the structural determinants of the hydrogen bond can change as well. There may be a torque on the hydrogen bond in the original state that is balanced by the hydrogen bond. As the bond becomes weaker, the torque has more effect and the constituents of the hydrogen bond can move. This has the effect of further weakening the hydrogen bond. Thus the change in dielectric may get amplified by structural changes; the resulting change in quality may be drammatic.

We will discuss experimental evidence in Section 8.1 that a single nonpolar carbonaceous group can have a measurable effect on the strength and stability of a hydrogen bond. However, these results concern formation of basic protein structure (helices) in an isolated solvent environment. In more complex protein environments, the effect of small amounts of wrapping may have less effect due to environmental constraints. These effects can lead to the more drammatic switch behavior depicted in Figure 2.9.

2.7 Exercises

Exercise 2.1 *Compare (2.1) with the atomic mass of atoms not listed in Table 2.3. Consult appropriate tables to find out the fraction of different isotopes that occur naturally.*

Exercise 2.2 *Download a PDB file for a protein and compute the distance distribution between sequential C_α carbons. What is the mean of the distribution? Compare this with the data in the figure at the top of page 282 in [324]. Look at PDB file 3SIL and explain why some distances are so small (nearly zero). Look at PDB file 1E08 and explain why one value is so much smaller than the others.*

Exercise 2.3 *Download a PDB file for a protein and compute the distance distribution between C_α carbons separated in sequence by k. That is, the sequential neighbors have $k = 1$. How does the mean distance vary as a function of k? Compare the distributions for $k = 3$ and $k = 4$; which has C_α carbons closer together?*

Exercise 2.4 *Download a PDB file for a protein and compute the N-O distance distribution between the N's and O's that are nearest in sequence (cf. Figure 2.1). That is, for each residue i, there is an O_i and an N_i. Plot the distributions for the distances $|O_i - N_i|$ and $|O_i - N_{i+1}|$.*

Exercise 2.5 *Download a PDB file for a protein and compute the N-O distance distribution between all pairs of carbonyl and amide groups in the peptide bonds (cf. Figure 2.1). Is there part of the distribution that corresponds to pairs forming local interactions (e.g., a hydrogen bond)? Explain your reasoning. Pick a cut-off distance R and plot the distribution of sequence separation k for all all N-O pairs whose distance is less than R. Experiment with R between 3 and 7 Å to see how the distribution changes.*

Exercise 2.6 *Download a PDB file for a protein and compute the N-O distance distribution between each carbonyl and and its nearest amide group with which is not covalently bonded in the peptide bonds (cf. Figure 2.1). That is, for each C-O group, find the nearest N and record the N-O distance. Is there part of the distribution that corresponds to pairs forming local interactions (e.g., a hydrogen bond)? Explain your reasoning. Also record the sequence distance k between the N's and O's and plot a two-dimensional distributions of k versus physical distance d. (Hint: exclude the N's and O's that are near neighbors in the peptide bond backbone.)*

Chapter 3

Electrostatic forces

The only force of significance in biochemistry is the electric force. However, it appears in many guises, often modulated by indirection, or induction. Chemistry has classified different regimes of electronic forces by cataloging **bonds** between different atoms. In terrestrial biology, water plays a dominant role as a dielectric that modulates different types of electronic interactions. Some bonds are more easily affected by water than others.

Here we briefly outline the main types of electronic forces as they relate to biology, and especially to proteins and other molecular structures. There are so many books that could be used as a reference that it is hard to play favorites. But the books by Pauling [324, 325] are still natural references.

The order of forces, or bonds [340], that we consider is significant. First of all, they are presented in order of strength, starting with the strongest. This order also correlates directly with the directness of interaction of the electrons and protons, from the intertwining of covalent bonds to indirect, induced interactions. Finally, the order is also reflective of the effect of solvent interaction to some extent, in that the dielectric effect of solvent is increasingly important for the weaker bonds.

3.1 Direct bonds

The strongest bonds can be viewed as the direct interactions of positive and negative charges, or at least distributions of charge.

3.1.1 Covalent bonds

These are the strong bonds of chemistry, and they play a role in proteins, DNA, RNA and other molecules of interest. However, their role in biology is generally static; they rarely break. They form the backbones of proteins, DNA, and RNA and support the essential linear structure of these macromolecules. Single lines represent single bonds and double (parallel) lines represent double bonds, as depicted in examples in Figure 3.1. The geometry for single and double bonds is often different. Double bonds often confer planar geometry, as shown in Figure 3.1(a). Certain atoms with only single bonds often confer a tetrahedral geometry, as shown in Figure 3.1(b). The letter R
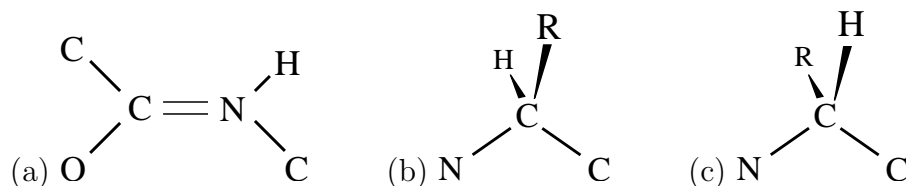
Figure 3.1: Single and double bonds. The double bond (a) often confers a planar geometry to the atoms; all six atoms in (a) are in the plane of the page. The upper-left and lower-right carbons represent C_α carbons in a peptide sequence, and would each have three additional single bonds (not shown). (b) Tetrahedral arrangement of atoms around the central (C_α) carbon in the basic (L-form) peptide unit. The Nitrogen and two Carbons are in the plane of the page, with the Hydrogen lying below the plane, away from the viewer, and the residue R (see text) lying above the plane, toward the viewer. The lower Nitrogen and Carbon would each be double-bonded (not shown) to a Carbon and Nitrogen (respectively) in a peptide sequence, as in (a). (c) The D-form of the peptide base, which occurs (naturally) only infrequently in proteins [295].

in Figure 3.1(b) stands for 'residue' which connotes a complex of from one to eighteen atoms which determine the different amino acid constituents of proteins.

One covalent bond of significant note that is not involved in defining the backbone is the disulfide bond (or disulfide bridge) between two cysteine sidechains (Section 4.2.2) in proteins. Further examples are shown in Figures 4.3–4.4 for aminoacid sidechains and Figure 13.1 for the peptide bond.

Covalent bonds involve the direct sharing of electrons from two different atoms, as required by the octet rule mentioned in Section 2.1. Such bonds are not easily broken, and they typically survive immersion in water. The octet rule [324, 325] allows the prediction of covalent bond formation through counting of electrons in the outer-most shell (see Table 2.1) of each atom. Explaining further such simple rules for other types of bonds is one of the major goals of this work.

Although covalent bonds are not easily broken, their character can be modified by external influences. The most important covalent bond in proteins is the peptide bond (Figure 13.1) formed between amino acids as they polymerize. This bond involves several atoms that are typically planar in the common form of the peptide bond. But if the external electrical environment changes, as it can if the amide and carbonyl groups lose hydrogen bond partners, the bond can bend. We review this effect in Chapter 13.

3.1.2 Ionic bonds/salt bridges

Ionic bonds occur in many situations of biological interest, but it is of particular interest due to its role in what is called a *salt bridge* (Section 4.2.1). Such an ionic bond occurs between oppositely charged side chains in a protein, as indicated in Figure 3.2. Ionic bonds involve the direct attraction of electrons in one molecule to the positive charge of another.

The potential for the electrostatic interaction between two charged molecules, separated by a
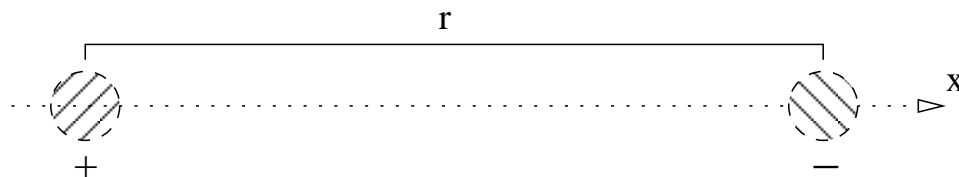
Figure 3.2: Salt bridges are ionic bonds.

distance r, is (see Section 3.2)

$$V(r) = z_1 z_2 r^{-1}, \tag{3.1}$$

where z_i is the charge on the i-th molecule. For two molecules with equal but opposite charges, say, $z_1 = 1$ and $z_2 = -1$, the potential is $-r^{-1}$.

We will see that different (noncovalent) bonds are characterized by the exponent of r in their interaction potential. For potentials of the form r^{-n}, we can say that the bonds with smaller n are more long range, since $r^{-n} >> r^{-m}$ for $n < m$ and r large. The ionic bond is thus the one with the longest range of influence.

In addition to being long range, ionic bonds are often stronger as well. For all bonds of attraction which are of the form r^{-n}, there would be infinite attraction at $r = 0$. However, there is always some other (electrostatic) force of repulsion that keeps the entities from coalescing. Thus the form of the attractive force is not sufficient to tell us the strength of the bond. However, ionic bonds are often quite strong as well as being long range, second only to covalent bonds in strength.

Although ionic bonds are relatively strong and have a long-range influence, they are also easily disrupted by water, as a simple experiment with table salt introduced into a glass of water will easily show. Salt forms a stable crystal when dry, but when wet it happily dissolves into a sea of separated ions. The source of attraction between the sodium and chloride ions in salt is the ionic bond.

3.1.3   Hydrogen bonds

Although weaker than covalent and ionic bonds, hydrogen bonds play a central role in biology. They bind complementary DNA and RNA strands in a duplex structure, and they secure the three-dimensional structure of proteins. However, they are also easily disrupted by water, which is the best hydrogen bond maker in nature.

First suggested in 1920, hydrogen bonds were not fully accepted until after 1944 [393]. The detailed structure of hydrogen bonds in biology is still being investigated [238, 300, 377, 398, 417]. Most of the hydrogen bonds of interest to us involve a hydrogen that is covalently bonded to a heavy atom X and is noncovalently bonded to a nearby heavy atom Y. Typically the heavy atoms X and Y are N, O, or S in protein systems, e.g., NH - - O or OH - - S, etc.; see Table 6.2 for a list. The bond OH – O describes the hydrogen bond between two water molecules.

The special nature of the hydrogen bond stems in part from the mismatch in size and charge compared to the other so-called 'heavy' atoms. Carbon is the next smallest atom of major biological

interest, with six times as many electrons and protons. The mismatch with nitrogen and oxygen is even greater. Hydrogen bonds will be discussed in more detail in Chapter 6.

3.1.4 Cation-π interactions

Aromatic residues (phenylalanine, tyrosine and tryptophan: see Section 4.4.5) are generally described as hydrophobic, due to the nonpolar quality of the carbon groups making up their large rings. But their carbon rings have a secondary aspect which *is* polar, in that there is a small negative charge distribution on each side of the plane formed by the rings [88, 164, 431]. This large distribution of negative charge can directly attract the positive charges of cations (e.g., arginine and lysine).

Cation-π interactions will be discussed in more detail in Chapter 12.

3.2 Charge-force relationship

We want to talk about the interaction energy (and force) between two charged groups. The units of charge and energy are not the same, and so we need to introduce a conversion factor to allow this.

Suppose we have a charge z at the origin in space $\mathbb{R}^3$. This induces an electric field $\mathbf{e}$ in all of space, and the relationship between the two is

$$\varepsilon \nabla \cdot \mathbf{e} = z\delta, \tag{3.2}$$

where ε is the permittivity and δ denotes the Dirac delta-function. Here, $\nabla \cdot \mathbf{e} = \sum_{i=1}^3 e_{i,i}$ is the divergence operator applied to a vector function $\mathbf{e}$ with components e_i; we have used the 'comma' notation to indicate the partial derivative with respect to the i-th variable. The concept of the Dirac delta-function is complex but well known: the expression (3.2) means that for any smooth function ϕ that vanishes outside a bounded set

$$-\varepsilon \int_{\mathbb{R}^3} \mathbf{e}(x) \cdot \nabla \phi(x) \, dx = z\phi(0). \tag{3.3}$$

Going between (3.3) and (3.2) is just integration by parts, except that $\nabla \cdot \mathbf{e}$ is not regular enough for this to be justified in a simple way. Thus (3.3) is taken as definition of (3.2).

When the medium is a vacuum, ε is the permittivity of free space, ε_0. When we write the expression (3.1), we have in mind the permittivity of free space. In other media (e.g., water) the value of ε is much larger. This quantity measures the strength of the dielectric enviroment. We can now see one example of the lack of duality between hydrophobic and hydrophilic groups mentioned in Section 2.2.1. Hydrophobicity affects the coefficient ε in (3.2), whereas hydrophilic groups would contribute to the right-hand side in the equation.

The exact value for ε_0 depends on the units (Chapter 17) chosen for charge, space, time, etc. The electric field $\mathbf{e}$ does not have units of force. If there is no other charge in the field, no force will be felt. The resulting force on a second charge z' is proportional to the amount of that charge:
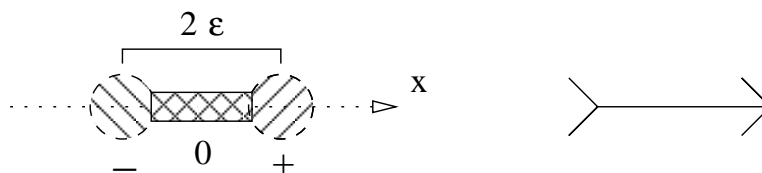
Figure 3.3: Left: configuration of a dipole consisting of a pair of charged molecules at $(\pm\epsilon, 0, 0)$. Right: abstract representation of a dipole as a vector.

$z'\mathbf{e}$. So the electric field $\mathbf{e}$ has units of force per unit of charge, whereas z has units of charge. The coefficient ε provides the change of units required by the relation (3.2), cf. (17.2).

The electric field $\mathbf{e}$ can be written as (minus) the gradient of a potential

$$\mathbf{e} = -\nabla V, \tag{3.4}$$

and therefore the potential is related to the charge by

$$-\varepsilon \Delta V = z\delta, \tag{3.5}$$

where we again have to invoke an interpretation like (3.3) to make proper sense of (3.5). It is not too difficult to verify that a solution to (3.5) is

$$V(r) = \frac{z}{4\pi\varepsilon r}. \tag{3.6}$$

Note that the units of V are energy per unit charge. We can have a simple representation of the relationship between charge z and its electric potential

$$V(r) = \frac{z}{r}, \tag{3.7}$$

provided we choose the units (Chapter 17) appropriately so that $\varepsilon = 1/4\pi$. The resulting potential energy of a pair of charges z_1 and z_2 is thus given by (3.1).

We will make this simplification in much of our discussion, but it should be remembered that there is an implicit constant proportional to the permittivity in the denominator. In particular, we see that a larger permittivity leads to a smaller potential and related force.

3.3   Interactions involving dipoles

A **dipole** is an abstract concept based on a collection of charges, e.g., in a molecule. The simplest example is given by two molecules of opposite charge that are fused together, e.g., by covalent bonds, as depicted in Figure 3.3. Water is a sea of such dipoles, oriented somewhat arbitrarily in response to the ambient electric field. The ability of the water dipole to orient in response to an external field is the basis for the dielectric response.

Mathematically, we imagine that the two charged molecules are placed on the x-axis with the $\pm$ charges at the positions $(\pm\epsilon, 0, 0)$, as shown on the left-hand side of Figure 3.3. We will see that
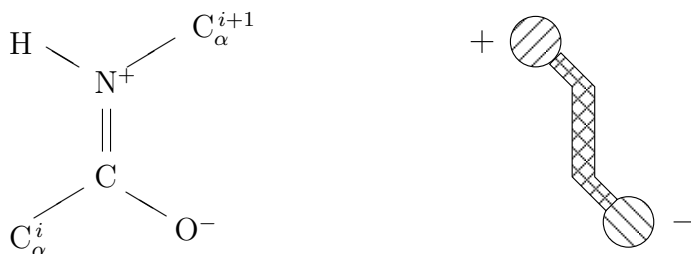
Figure 3.4: The dipole of the peptide backbone (in the 'trans' configuration, cf. Figure 4.2). On the left is the chemical description, and the right shows a cartoon of the dipole. The double bond between the central carbon and nitrogen keeps the peptide bond planar. There is a positive charge center near the hydrogen and a negative charge center near the oxygen.

it will be possible (for ϵ small) to think of a dipole as just a vector, as depicted on the right-hand side of Figure 3.3.

   We will see many examples of dipoles. One very important one is in the peptide backbone shown in Figure 3.4.

   Many interactions can be modeled as dipole-dipole interactions, e.g., between water molecules. More generally, the use of partial charges (cf. Table 12.1) represents many interactions as dipole-dipole interactions. Forces between molecules with fixed dipoles are often called Keesom forces [161]. For simplicity, we consider dipoles consisting of the same charges of opposite signs, separated by a distance 2ϵ. If the charges have unit value, then the dipole strength $\mu = 2\epsilon$. Interacting dipoles have two orientations which produce no torque on each other.

3.3.1   Single-file dipole-dipole interactions

In the single-file orientation, the base dipole has a unit positive charge at $(\epsilon, 0, 0)$ and a unit negative charge at $(-\epsilon, 0, 0)$; the other dipole is displaced on the x-axis at a distance r: a unit positive charge at $(r + \epsilon, 0, 0)$ and a unit negative charge at $(r - \epsilon, 0, 0)$ (cf. Figure 3.5). Since the potential for two charges is the sum of the individual potentials (that is we assume linear additivity), the potential due to the base dipole at a distance $r >> \epsilon$ along the x-axis is

$$
\begin{aligned}
V(r) &= \frac{1}{r - \epsilon} - \frac{1}{r + \epsilon} = \frac{(r + \epsilon) - (r - \epsilon)}{(r - \epsilon)(r + \epsilon)} \\
&= \frac{2\epsilon}{(r - \epsilon)(r + \epsilon)} = \frac{2\epsilon}{r^2 - \epsilon^2} \approx 2\epsilon r^{-2} = \mu r^{-2},
\end{aligned}
\tag{3.8}
$$

where $\mu = 2\epsilon$ is the dipole strength.

   We use the expression $f(r) \approx g(r)$ to mean that the expression $f(r)$ is a good approximation to $g(r)$. More precisely, in this case we mean that the two expressions are asymptotically equal for
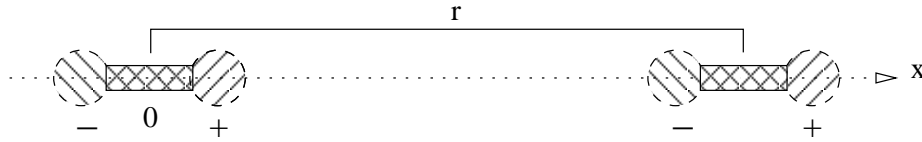
Figure 3.5: Single-file dipole-dipole configuration consisting of two pairs of molecules with charges ± 1 at $(\pm\epsilon, 0, 0)$ and $(r \pm \epsilon, 0, 0)$.

large r, that is, that

$$\lim_{r \to \infty} g(r)/f(r) = 1. \tag{3.9}$$

In (3.8), $f(r) = 1/(r^2 - \epsilon^2)$ and $g(r) = r^{-2}$, so that $g(r)/f(r) = 1 - \epsilon^2/r^2$, and thus (3.9) follows. Moreover, we can get a quantitative sense of the approximation: the approximation in (3.8) is 99% accurate for $r \geq 10\epsilon$, and even 75% accurate for $r \geq 2\epsilon$.

In the field of the dipole (3.8), the potential energy of a single charge on the x-axis at a distance r is thus μr^{-2}, for a charge of $+1$, and $-\mu r^{-2}$, for a charge of -1. In particular, we see that the charge-dipole interaction has a potential one order lower (r^{-2}) than a charge-charge interaction (r^{-1}). The charge-dipole interaction is very important, but we defer a full discussion of it until Section 10.3.1.

The combined potential energy of two opposite charges in the field generated by a dipole is given by the difference of terms of the form (3.8). In this way, we derive the potential energy of a dipole, e.g., a positive charge at $(r+\epsilon, 0, 0)$ and a negative charge at $(r-\epsilon, 0, 0)$, as the sum of the potential energies of the two charges in the field of the other dipole:

$$\frac{\mu}{(r+\epsilon)^2} - \frac{\mu}{(r-\epsilon)^2}. \tag{3.10}$$

Considering two such charges as a combined unit allows us to estimate the potential energy of two dipoles as

$$
\begin{aligned}
\frac{\mu}{(r+\epsilon)^2} - \frac{\mu}{(r-\epsilon)^2} &= -\mu \frac{(r+\epsilon)^2 - (r-\epsilon)^2}{(r+\epsilon)^2(r-\epsilon)^2} \\
&= -\mu \frac{4r\epsilon}{(r+\epsilon)^2(r-\epsilon)^2} \approx -4\mu\epsilon r^{-3} = -2\mu^2 r^{-3}.
\end{aligned}
\tag{3.11}
$$

The negative sign indicates that there is an attraction between the two dipoles in the configuration Figure 3.5.

The electric force field $\mathbf{F}$ is the negative gradient of the potential ∇V. For V defined by (3.8), only the x-component of ∇V is non-zero along the x-axis, by symmetry. Differentiating (3.8), we find that for $r \gg \epsilon$ along the x-axis,

$$
\begin{aligned}
F_x(r, 0, 0) &= -(r-\epsilon)^{-2} + (r+\epsilon)^{-2} \\
&= \frac{-(r+\epsilon)^2 + (r-\epsilon)^2}{(r-\epsilon)^2(r+\epsilon)^2} = \frac{-4\epsilon r}{(r-\epsilon)^2(r+\epsilon)^2} \\
&\approx -4\epsilon r^{-3} = -2\mu r^{-3}.
\end{aligned}
\tag{3.12}
$$
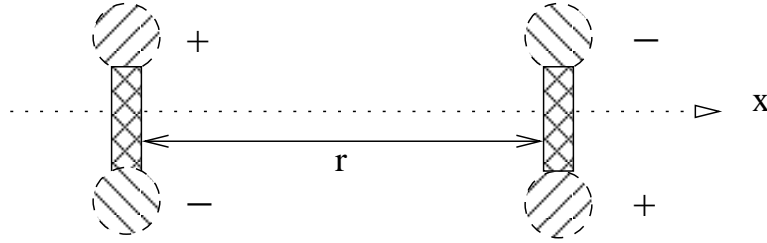
Figure 3.6: Parallel dipole-dipole configuration with ± 1 charges at $(0, \pm\epsilon, 0)$ and $(r, \mp\epsilon, 0)$.

The attractive force experienced by a dipole displaced on the x-axis at a distance r, with a positive charge at $(r + \epsilon, 0, 0)$ and a negative charge at $(r - \epsilon, 0, 0)$, is thus (asymptotically)

$$
\begin{aligned}
-\frac{2\mu}{(r+\epsilon)^3} + \frac{2\mu}{(r-\epsilon)^3} &= 2\mu \frac{(r+\epsilon)^3 - (r-\epsilon)^3}{(r+\epsilon)^3(r-\epsilon)^3} \\
&= 2\mu \frac{6r^2\epsilon + 2\epsilon^3}{(r+\epsilon)^3(r-\epsilon)^3} \approx 6\mu^2 r^{-4},
\end{aligned}
\tag{3.13}
$$

which is equal to the derivative of the potential (3.11) as we would expect.

3.3.2 Parallel dipole-dipole interactions

In the parallel orientation, the base dipole has a positive charge at $(0, \epsilon, 0)$ and a negative charge at $(0, -\epsilon, 0)$; the other dipole is displaced on the x-axis at a distance r: a positive charge at $(r, -\epsilon, 0)$ and a negative charge at $(r, +\epsilon, 0)$ (cf. Figure 3.6).

The potential in the (x, y)-plane due to the base dipole at a distance r along the x-axis is

$$
V(x, y) = \frac{1}{\sqrt{(y-\epsilon)^2 + x^2}} - \frac{1}{\sqrt{(y+\epsilon)^2 + x^2}}
\tag{3.14}
$$

The potential energy of a dipole displaced on the x-axis at a distance r, with a positive charge at $(r, -\epsilon, 0)$ and a negative charge at $(r, \epsilon, 0)$, is thus

$$
\begin{aligned}
\left(\frac{1}{\sqrt{(2\epsilon)^2 + r^2}} - \frac{1}{r} \right) - \left(\frac{1}{r} - \frac{1}{\sqrt{(2\epsilon)^2 + r^2}} \right) &= -2 \left(\frac{1}{r} - \frac{1}{\sqrt{(2\epsilon)^2 + r^2}} \right) \\
&= -2 \frac{\sqrt{(2\epsilon)^2 + r^2} - r}{r\sqrt{(2\epsilon)^2 + r^2}} = -2 \frac{\sqrt{(2\epsilon/r)^2 + 1} - 1}{r\sqrt{(2\epsilon/r)^2 + 1}} \\
&\approx -\frac{(2\epsilon/r)^2}{r} = -\mu^2 r^{-3}.
\end{aligned}
\tag{3.15}
$$

Thus the potential energy of the parallel orientation is only half of the single-file orientation.

The potential $V(x, y)$ in (3.14) vanishes when $y = 0$. Therefore, its derivative along the x-axis also vanishes: $\frac{\partial V}{\partial x}(r, 0) = 0$. However, this does not mean that there is no attractive force between
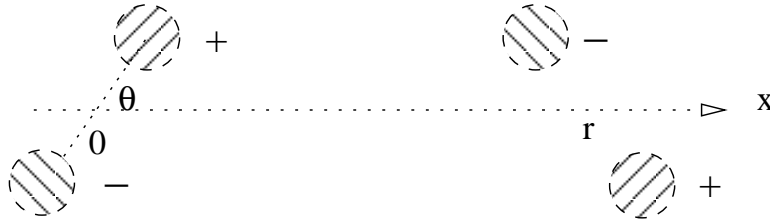
Figure 3.7: General θ-dependent dipole-dipole configuration. The connective material between the charge centers has been omitted for simplicity.

the dipoles, since (by symmetry) $\frac{\partial V}{\partial x}(r, \pm\epsilon) = \pm f(\epsilon, r)$. Thus the attractive force is equal to $2f(\epsilon, r)$. For completeness, we compute the expression $f(\epsilon, r)$:

$$\frac{\partial V}{\partial x}(x, y) = \frac{-x}{((y - \epsilon)^2 + x^2)^{3/2}} + \frac{x}{((y + \epsilon)^2 + x^2)^{3/2}} \tag{3.16}$$

for general y. Choosing $y = \pm\epsilon$, (3.16) simplifies to

$$\frac{\partial V}{\partial x}(r, \pm\epsilon) = \mp r^{-2} \pm \frac{r}{((2\epsilon)^2 + r^2)^{3/2}} = \mp r^{-2}\left(1 - \frac{1}{((\mu/r)^2 + 1)^{3/2}}\right)$$
$$= \mp \frac{((\mu/r)^2 + 1)^{3/2} - 1}{r^2 \left((\mu/r)^2 + 1\right)^{3/2}} \approx \mp \frac{3\mu^2}{2r^4}, \tag{3.17}$$

for large r/ϵ. The net force of the field (3.17) on the two oppositely charged particles on the right side of Figure 3.6 is thus $3\mu^2 r^{-4}$, consistent with what we would find by differentiating (3.15) with respect to r.

The electric force field in the direction of the second dipole (that is, the y-axis) is

$$\frac{\partial V}{\partial y}(r, y) = \frac{\epsilon - y}{((y - \epsilon)^2 + r^2)^{3/2}} + \frac{\epsilon + y}{((y + \epsilon)^2 + r^2)^{3/2}}. \tag{3.18}$$

At a distance $r \gg \epsilon$ along the x-axis, this simplifies to

$$\frac{\partial V}{\partial y}(r, \pm\epsilon) = \frac{\mu}{(\mu^2 + r^2)^{3/2}} \approx \mu r^{-3}, \tag{3.19}$$

for large r/ϵ. Although this appears to be a force in the direction of the dipole, the opposite charges on the dipole on the right side of Figure 3.6 cancel this effect. So there is no net force on the dipole in the direction of the y-axis.

3.3.3   Dipole stability

Only the single-file dipole orientation is stable with respect to perturbations. This can be seen as follows. Suppose the dipoles are arranged along the x-axis as above but that they are both tilted
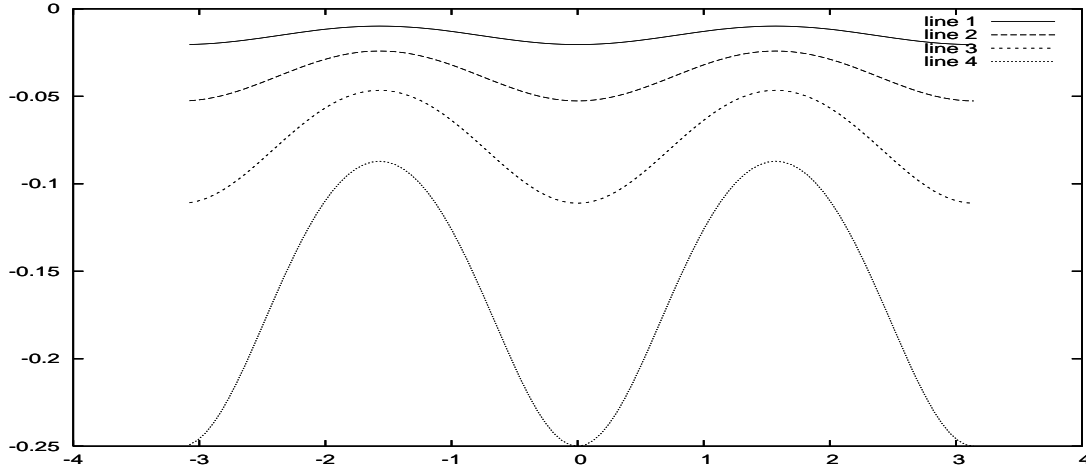
Figure 3.8: Potential energy variation $v(\rho, \theta)$ as defined in (3.23) (vertical axis) of dipoles as a function of θ (horizontal axis) for the configurations shown in Figure 3.7 for $\rho = 0.02$ (top), $0.05, 0.1, 0.2$ (bottom), where ρ is defined in (3.22).

away from the x-axis at an angle θ, as shown in Figure 3.7. The connective material between the charge centers has been omitted for simplicity, and will be omitted in future drawings. Define θ so that $\theta = 0$ (and $\theta = \pi$) is the single-file dipole configuration and $\theta = \pi/2$ is the parallel configuration. Thus one dipole has a positive charge at $\epsilon(\cos \theta, \sin \theta, 0)$ and a negative charge at $-\epsilon(\cos \theta, \sin \theta, 0)$. The other dipole is displaced on the x-axis at a distance r: a positive charge at $(r + \epsilon \cos \theta, -\epsilon \sin \theta, 0)$ and a negative charge at $(r - \epsilon \cos \theta, \epsilon \sin \theta, 0)$.

The potential at the point $(x, y, 0)$ due to the rotated base dipole is

$$V(x, y) = \frac{1}{\sqrt{(x - \epsilon \cos \theta)^2 + (y - \epsilon \sin \theta)^2}} - \frac{1}{\sqrt{(x + \epsilon \cos \theta)^2 + (y + \epsilon \sin \theta)^2}} \tag{3.20}$$

Therefore the potential energy of the second rotated dipole, with a positive charge at $(r + \epsilon \cos \theta, -\epsilon \sin \theta, 0)$ and a negative charge at $(r - \epsilon \cos \theta, \epsilon \sin \theta, 0)$, is thus

$$
\begin{aligned}
V(r, \theta) &= \frac{1}{\sqrt{r^2 + (2\epsilon \sin \theta)^2}} - \frac{1}{r + 2\epsilon \cos \theta} - \left(\frac{1}{r - 2\epsilon \cos \theta} - \frac{1}{\sqrt{r^2 + (2\epsilon \sin \theta)^2}} \right) \\
&= \frac{2}{\sqrt{r^2 + (2\epsilon \sin \theta)^2}} - \frac{1}{r + 2\epsilon \cos \theta} - \frac{1}{r - 2\epsilon \cos \theta} \\
&= \frac{2}{\sqrt{r^2 + (2\epsilon \sin \theta)^2}} - \frac{2r}{r^2 - (2\epsilon \cos \theta)^2} \\
&= \frac{2}{r} \left(\frac{1}{\sqrt{1 + \rho \sin^2 \theta}} - \frac{1}{1 - \rho \cos^2 \theta} \right) := \frac{2}{r} v(\rho, \theta),
\end{aligned}
\tag{3.21}
$$

where the (nondimensional) parameter ρ is defined by

$$\rho = (2\epsilon/r)^2. \tag{3.22}$$
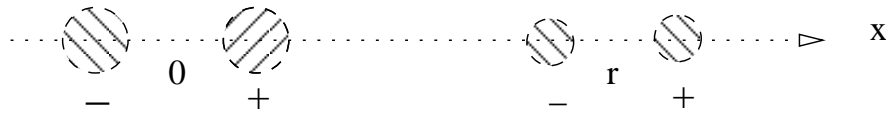
Figure 3.9: Single-file dipole-dipole configuration with different dipole strengths.

name	constituents
Keesom	two fixed dipoles
Debye	one fixed dipole and one induced dipole
London dispersion	two induced dipoles

Table 3.1: Forces involving dipoles.

This expression

$$v(\rho, \theta) = \frac{1}{\sqrt{1 + \rho \sin^2 \theta}} - \frac{1}{1 - \rho \cos^2 \theta} \tag{3.23}$$

in (3.21) has a minimum when $\theta = 0$ and a maximum when $\theta = \pi/2$. A plot of v in (3.23) is shown in Figure 3.8 for various values of ρ. When ρ is small, the expression (3.23) tends to the limit

$$
\begin{aligned}
v(\rho, \theta) &\approx \frac{1}{1 + \frac{1}{2}\rho \sin^2 \theta} - \frac{1}{1 - \rho \cos^2 \theta} \\
&\approx \left(1 - \tfrac{1}{2}\rho \sin^2 \theta\right) - \left(1 + \rho \cos^2 \theta\right) = -\tfrac{1}{2}\rho \left(1 + \cos^2 \theta\right).
\end{aligned}
\tag{3.24}
$$

Of course, what we have presented is only an indication of the stability and energy minimum of the single-file dipole configuration. We leave a complete proof as Exercise 3.5.

3.3.4 Different dipoles

So far, we considered dipoles with identical charges and charge distributions (separations). Here we consider a single-file configuration as in Figure 3.5, but with the dipole on the right consisting of charges $\pm q$ separated by a distance δ, as depicted in Figure 3.9. We consider the potential energy of the right-hand dipole in the potential field (3.8) of the left dipole. Similar to (3.11), we find

$$\frac{\mu q}{(r + \delta)^2} - \frac{\mu q}{(r - \delta)^2} = -\mu q \frac{4r\delta}{(r + \delta)^2 (r - \delta)^2} \approx -4\mu q \delta r^{-3} = -2\mu\nu r^{-3}, \tag{3.25}$$

where $\nu = 2q\delta$ is the strength of the dipole on the right. Notice that the expression (3.25) is symmetric in the two dipole strengths μ and ν.

3.4 van der Waals forces

Many of the electric forces we consider are induced rather than direct. The best known of these are called van der Waals forces, although this term covers a range of forces known by other names.

atom	ρ	ϵ	$V(\rho/2)$	D	κ
C (aliphatic)	1.85	0.12	476	1.54	83.1
O	1.60	0.20	794	1.48	33.2
H	1.00	0.02	79	0.74	104.2
N	1.75	0.16	635	1.45	38.4
P	2.10	0.20	794	1.87	51.3
S	2.00	0.20	794	1.81	50.9

Table 3.2: Lennard-Jones parameters from Amber for various atoms involving the van der Waals radius ρ measured in Ångstroms and energy (well depth) ϵ in kcal/mol. For comparison, covalent bond lengths D and strengths [324] κ are given in kcal/mol, together with the repulsion potential energy $V(\rho)$ at the van der Waals radius ρ.

Keesom forces, which we covered in Section 3.3, are often included in this group, but we will see that there is a qualitative difference in behavior. One prominent web site went as far as to say "all intermolecular attractions are known collectively as van der Waals forces" but this seems a bit extreme. To clarify names and usages, we summarize in Table 3.1 the different terms used and what they refer to.

We cover van der Waals forces in detail here to clarify that they are electrostatic in nature, and not some new or different type of force. As such they are susceptible to modulation by solvent dielectric behavior.

Debye forces and London dispersion forces [161] involve induced dipole-dipole interactions, which we will study using the results derived in Section 3.5. The most significant example is the London dispersion force [161] which results from both dipoles being induced.

3.4.1 Lennard-Jones potentials

The van der Waals interactions are often modeled via the Lennard-Jones potential

$$V(r) := \epsilon \left(\left(\frac{\rho}{r} \right)^{12} - 2 \left(\frac{\rho}{r} \right)^{6} \right). \tag{3.26}$$

The attractive potential r^{-6} is a precise result of the interaction of a fixed dipole and an induced dipole, which we derive in Section 3.5. In Section 3.5.2, we consider the self-induction of two dipoles. The repulsive term r^{-12} is a convenient model, whereas other terms are more accurate [86].

The minimum of V is at $r = \rho$, with $V(\rho) = -\epsilon$, so we can think of the well depth ϵ as giving the energy scale. The parameter ρ is called the **van der Waals radius**, and can be defined as the separation distance at which the force of attraction and repulsion cancel [48]. Typical values for these parameters, from the Amber force field, are shown in Table 3.2. Note that $V(\rho/1.2) \approx -3V(\rho)$, and $V(\rho/2) = -3968V(\rho)$, so the repulsion is quite strong in this model.

The maximal attractive force occurs when $r/\rho = \sqrt[6]{13/7}$ (Exercise 3.8).

molecule	α_0	α_1	α_2	α_3
benzene		10.66	10.66	4.01
methane	2.62	2.62	2.62	2.62
water	1.49			

Table 3.3: Experimental [10] and derived [402] values for polarizabilities of some molecules. Units are Å^3.

Figure 3.10: The Debye force: a dipole (left) inducing a dipole in a polarizable molecule (right). The upper configuration (a) shows the dipole and polarizable molecule well separated, and the lower configuration (b) shows them closer, with the molecule on the right now polarized.

3.5 Induced dipoles

Dipoles can be induced in two ways. Fixed dipoles, such as water molecules, induce a dipole in any polarizable material. Such interactions give rise to what are frequently called Debye forces [161]. More subtly, two polarizable molecules can induce dipoles in each other, via what are called London dispersion forces [161].

Essentially all materials are polarizable. This just means that the distributions of electrons can be distorted by an electric field. Table 3.3 gives some typical values of polarizability.

3.5.1 Debye forces

If a polarizable molecule is subjected to an electric field of strength $\mathbf{F}$, then it is reasonable to expect that an induced dipole μ_i will result, given by

$$\mu_i \approx \alpha \mathbf{F} \tag{3.27}$$

for small $\mathbf{F}$, where α is the polarizability. This is depicted visually in Figure 3.10, where the upper configuration (a) shows the dipole and polarizable molecule well separated, and the lower configuration (b) shows them closer, with the molecule on the right now polarized.

In general, the electric field $\mathbf{F}$ is a vector and the polarization α is a tensor (or matrix). Also, note that a dipole is a vector quantity: it has a magnitude and direction. In our previous discussion, we

(a)

(b)

Figure 3.11: Mutual polarization of two molecules. The upper configuration (a) shows the polarizable molecules well separated, and the lower configuration (b) shows them closer, with the molecules now visibly polarized.

considered only the magnitude, but the direction was implicit (the line connecting the two charges). For simplicity, we assume here that α can be represented as a scalar (times the identity matrix), that is, that the polarizability is isotropic.

We can approximate a polarized molecule as a simple dipole with positive and negative charges $\pm q$ displaced by a distance δ, as depicted in Figure 3.9. This takes some justification, but it will be addressed in Chapter 10. There is ambiguity in the representation in that only the product $q\delta$ matters: $\mu = q\delta$.

We derived in (3.12) that the electric force field due to a fixed dipole μ_f has magnitude

$$F_x = 2\mu_f r^{-3}, \tag{3.28}$$

where the x-axis connects the two charges of the fixed dipole. We assume that the molecule whose dipole is being induced also lies on this axis. By combining (3.27) and (3.28), we conclude that the strength of the induced dipole is

$$\mu_i \approx 2\alpha\mu_f r^{-3}. \tag{3.29}$$

From (3.25), we know that the potential energy of the two dipoles is

$$V(r) \approx -2\mu_f\mu_i r^{-3} \approx -4\alpha\mu_f^2 r^{-6}, \tag{3.30}$$

in agreement with the Lennard-Jones model in (3.26).

3.5.2 London dispersion forces

Suppose now that we start with two nonpolar, but polarizable, molecules that are well separated. Due to the long range interaction (correlation) of the electron distributions of the two molecules, they can become polarized. This property is related to what is known as **entanglement**. To get a sense of what might happen, suppose one of them polarizes first so that it becomes the dipole depicted on the left in Figure 3.10. Then as it approaches the other molecule, it induces a dipole in it. But what if the molecules are identical? Then the induced dipole is the same as the 'fixed'

dipole that was in the case of the Debye force: $\mu_i = \mu_f$. Thus there is only one μ in the discussion now. This situation is depicted in Figure 3.11.

The dipole μ is induced by the electric field of the other dipole, so that again $\mu \approx \alpha\mathbf{F}$ where $\mathbf{F}$ is the electric field strength and α is the polarizability. The electric strength of the field $\mathbf{F}$ is again given by (3.28): the x-component of $\mathbf{F}$ is given by $F_x = 2\mu r^{-3}$. But now the electric field strength and the dipole strength are coupled in a new way, and it is not simple to solve this system.

The expression (3.25) remains valid for at least a component of the potential energy of the induced dipoles:

$$V \approx -2\mu^2 r^{-6}. \tag{3.31}$$

But how big is the induced dipole μ in expression (3.31)? We saw in (3.29) that the dipole induced by a fixed dipole has a magnitude that is asymptotic to r^{-3}. If such an asymptotic behavior were to hold in the case of doubly induced dipoles, it would lead to an expression for the potential energy of the induced dipoles of the form

$$V \approx -cr^{-12}, \tag{3.32}$$

which is quite different from the (attractive part of the) Lennard-Jones model. We will see that the correct asymptotic behavior for the energy of interaction for the induced dipoles themselves is smaller than r^{-6}. However, a new phenomenon emerges which forces us to expand our molecular view to include a distributed description of the electronic distribution for each molecule. In Section 3.5.3 we show how a dipole can interact with any matter due to the fact that the positive and negative charges are not co-located. However, we will also see (Section 3.5.3) that this does not provide the expected r^{-6} behavior of the potential energy, but this requires a more detailed argument. Before we discard our simplistic molecular view, let us look in more detail at these ideas.

Let us review the arguments used to estimate the magnitude of the dipole induced by a fixed dipole to see where it fails for doubly induced dipoles. It is reasonable to assume that the dipole strength is a monotone function of the induced field $\mathbf{F}$; we used the *ansazt* that $\mu \approx \alpha\mathbf{F}$ for small $\mathbf{F}$ in the derivation of the r^{-6} dependence of V for a dipole induced by a fixed dipole. But since $\mathbf{F}$ depends on r, so must μ depend on r, and this would mean that our expression (3.31) would not be a complete description of the asymptotic behavior of V as a function of r, and it would imply that the behavior of $\mathbf{F} = \nabla V$ would go to zero faster than r^{-3}. This would imply that $\mu \approx \alpha\mathbf{F}$ would be even smaller. In fact, if we iterate the argument, we would never converge on a finite power of r. Let us analyze the argument in more detail.

We used two key equations, namely (3.27) and (3.28), in deriving the expression for V for a dipole induced by a fixed dipole. If we now assume that $\mu = \mu_i = \mu_f$, the two equations $\mu = \alpha\mathbf{F}$ (in scalar form, $\mu = \alpha F_x$) and $F_x = 2\mu r^{-3}$ can be solved to find

$$r = \sqrt[3]{2\alpha}. \tag{3.33}$$

Note that this is dimensionally correct, since the units of the polarizability α are the same as volume. Thus using the two equations (3.27) and (3.28) together with the simplification $\mu = \mu_i = \mu_f$ determines a particular value of r, in contradiction to our derivation of an expression valid for various values of r.
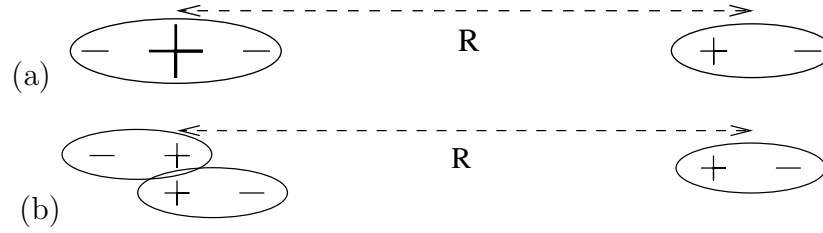
Figure 3.12: (a)Interaction between a dipole (right) and a neutral group (left). The neutrally charged group on the left has charges $+2c$ at the origin and $-c$ located at $(\pm\delta, 0, 0)$, and the dipole on the right has charges ± 1 at $(R \mp \frac{1}{2}\mu, 0, 0)$. (b) Equivalent representation of the interaction between a dipole and a neutral group, in which the neutral group in (a) is written as two dipoles.

To achieve a correct description of the van der Waals force, a quantum mechanics description of the atomic interactions is needed. It can be shown that

$$\mu \approx c_1 r^{-3}, \tag{3.34}$$

where an expression for the constant c_1 can be made explicit. In addition, one can also demonstrate that the full potential energy of the doubly-induced dipole pair has potential energy

$$V(r) \approx c_2 r^{-6}, \tag{3.35}$$

where an expression for the constant c_2 can also be made explicit.

3.5.3 Dipole-neutral interactions

Suppose we have two charge groups as indicated in Figure 3.12(a). The exact positions of the charges are as follows. We assume that the dipole on the right consists of charges ± 1 located at $(R \mp \frac{1}{2}\mu, 0, 0)$, and the neutrally charged group on the left has charges $-c$ located at $(\pm\delta, 0, 0)$ and $+2c$ at the origin. Then the potential energy of this systems is $cV(\mu, \delta)$ where

$$
\begin{aligned}
V(\mu, \delta) =& \frac{2}{R - \frac{1}{2}\mu} + \frac{1}{R + \frac{1}{2}\mu - \delta} + \frac{1}{R + \frac{1}{2}\mu + \delta} - \frac{2}{R + \frac{1}{2}\mu} - \frac{1}{R - \frac{1}{2}\mu - \delta} - \frac{1}{R - \frac{1}{2}\mu + \delta} \\
=& \frac{2\mu}{R^2 - \frac{1}{4}\mu^2} - \frac{\mu}{R^2 - (\frac{1}{2}\mu - \delta)^2} - \frac{\mu}{R^2 - (\frac{1}{2}\mu + \delta)^2} \\
=& \frac{2\mu}{R^2 - \frac{1}{4}\mu^2} - \frac{\mu}{R^2 - \frac{1}{4}\mu^2 + \mu\delta - \delta^2} - \frac{\mu}{R^2 - \frac{1}{4}\mu^2 - \mu\delta - \delta^2} \\
=& \frac{2\mu}{R^2 - \frac{1}{4}\mu^2} - \frac{2\mu(R^2 - \frac{1}{4}\mu^2 - \delta^2)}{(R^2 - \frac{1}{4}\mu^2 + \mu\delta - \delta^2)(R^2 - \frac{1}{4}\mu^2 - \mu\delta - \delta^2)} \\
=& \frac{2\mu}{\rho + \delta^2} - \frac{2\mu\rho}{\rho^2 - \mu^2\delta^2},
\end{aligned}
\tag{3.36}
$$

where $\rho = R^2 - \frac{1}{4}\mu^2 - \delta^2$. Therefore

$$
\begin{aligned}
V(\mu, \delta) &= \frac{2\mu}{\rho + \delta^2} - \frac{2\mu\rho}{\rho^2 - \mu^2\delta^2} = \frac{2\mu(\rho^2 - \mu^2\delta^2) - 2\mu\rho(\rho + \delta^2)}{(\rho + \delta^2)(\rho^2 - \mu^2\delta^2)} \\
&= -\frac{2\mu^3\delta^2 + 2\mu\rho\delta^2}{(\rho + \delta^2)(\rho^2 - \mu^2\delta^2)} \approx -\frac{2\mu\delta^2}{R^4},
\end{aligned}
\tag{3.37}
$$

for R large. To check that this result is correct, we could also think of this system as involving two dipoles on the left, of opposite sign and displaced by a distance δ, as depicted in Figure 3.12(b). Since the dipole-dipole interaction is of order R^{-3}, the difference of two such interactions should be of the order R^{-4}.

In Section 10.5, we will see that Figure 3.12 represents the interaction between a dipole and a quadrupole. This is significant because all atoms have distributed charges that appear to some extent like Figure 3.12. That is, there is a positively charged nucleus surrounded by a negatively charged cloud of electrons. The results of this section, and in particular (3.37), show that there is a natural attraction between a dipole and such a distributed charge. We can think of this as being a type of van der Waals force.

3.6 Exercises

Exercise 3.1 *Show that the approximation in (3.8) is 96% accurate for $r \geq 5\epsilon$.*

Exercise 3.2 *Prove that (3.11) is still correct if we use the exact form in (3.8) instead of the approximation $V(r) \approx \mu r^{-2}$.*

Exercise 3.3 *Prove that (3.13) is still correct if we use the exact form in (3.12), $F_x(r,0,0) = -4\epsilon r(r - \epsilon)^{-2}(r + \epsilon)^{-2}$, instead of the approximation $F_x(r,0,0) \approx -2\mu r^{-3}$.*

Exercise 3.4 *Consider the expression in (3.21). Prove that, for any $\rho < 1$, it has a minimum when $\theta = 0$ and a maximum when $\theta = \pi/2$.*

Exercise 3.5 *Prove that the single-file dipole configuration is stable and an energy minimum. (Hint: derive a formula for the general orientation of two dipoles in three dimensions, cf. Figure 2.2 in [204]. This can be done with one distance parameter and three angular parameters.)*

Exercise 3.6 *Describe the orientation of the dipoles that corresponds to $\theta = 2\pi$ in Figure 3.8.*

Exercise 3.7 *Show that the in-line dipole-dipole interaction energy is*

$$
-\frac{1}{r-1} + \frac{2}{r} - \frac{1}{r+1} \approx -\left(\frac{\partial^2}{\partial r^2}\right)\left(\frac{1}{r}\right) = -\frac{2}{r^3}
\tag{3.38}
$$

as $r \to \infty$ by considering the error in the second-difference operator represented by the left-hand side.

Exercise 3.8 *Compute the position of the maximum attractive force for the Lennard-Jones potential (3.26). Determine the magnitude of the force in terms of the parameters of the Lennard-Jones potential.*

Chapter 4

Protein basics

It is not our intention to provide a complete introduction to the structure of proteins. Instead, we suggest consulting texts [87, 332] for further information. Moreover, we suggest acquiring a molecular modeling set so that accurate three-dimensional models can be constructed. In addition, it will be useful to become familiar with a graphical viewer for PDB files (even the venerable 'rasmol' would be useful). We present some essential information and emphasize concepts needed later or ones that may be novel.

4.1 Chains of amino acid residues

Proteins are sequences of amino acids which are covalently bonded along a 'backbone.' The basic units of the backbone are depicted in Figure 4.1. In each unit, there is a residue denoted by R that is a molecule that can vary in size, and it is bonded to the central Carbon atom in the unit, called the C_α carbon. The twenty residues R of most interest in biology are represented in Figures 4.3—4.4. The peptide units bond together to form (arbitrarily long) sequences by forming a double bond between the N-terminus and the C-terminus, as shown in Figure 2.1.

Proteins of biological significance fold into a three-dimensional structure by adding hydrogen bonds between carbonyl and amide groups on the backbone of different amino acids. In addition,
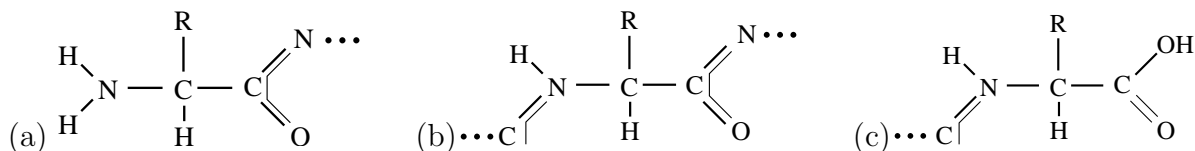
Figure 4.1: The basic units of a peptide sequence. (a) The initial, or N-terminal, unit. (a) The typical (internal) unit. (c) The final, or C-terminal, unit. The letter R stands for 'residue' and can be any of the twenty depicted in Figures 4.3—4.4. The atoms are not all co-planar. In particular, the four bonds around the central C_α Carbon are in a tetrahedral arrangement as shown in Figure 3.1. The dots before the Carbons and after the Nitrogens indicate the continuation of the peptide sequence.
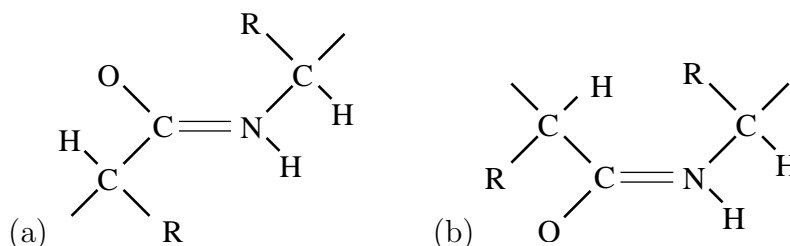
Figure 4.2: The rigid state of the peptide bond: (a) trans form, (b) cis form [39]. The letter R stands for 'residue' and can be any of the twenty depicted in Figures 4.3—4.4. The double bond between the central carbon and nitrogen keeps the peptide bond planar. Compare Figure 13.1.

other bonds, such as a salt bridge (Section 4.2.1) or a disulfide bond (Section 4.2.2), can form between particular amino acids (Cysteine has sulfur atoms in its sidechain). However, the hydrogen bond is the primary mode of structure formation in proteins.

The basic unit of the peptide group shown in Figure 4.2 comes in two forms that are related by a rotation around the C-N bond. The **trans** form (a) of the peptide bond is the most common state, but the **cis** form (b) has a small but significant occurrence [39, 180, 321]. The occurences of the cis form are overwhelmingly associated with prolines, but other residues are also involved [222, 205]. The covalent bond linking O-C-N is called a resonance, and it indicated by a polygonal line linking them in Figures 4.1 and 2.1. This particular bond will be considered in more detail in Chapter 13. In Figure 4.2 we represent the C-N bond as a double bond, the resonant state that confers rigidity to the peptide unit.

The peptide chain units are joined at the double bond indicated between the N and the O in Figure 4.2. Thus we refer to the coordinates of the nitrogen and hydrogen as N^{i+1} and H^{i+1} and to the coordinates of the oxygen and carbon as O^i and C^i.

At the ends of the chain, things are different, as depicted in Figure 4.1(a,c). The **N-terminus**, or **N-terminal end**, has an NH_2 group instead of just N, and nothing else attached, as shown in Figure 4.1(a). In the standard numbering scheme, this is the beginning of the chain. The **C-terminus**, or **C-terminal end**, has a COOH group instead of just CO, and nothing else attached, as shown in Figure 4.1(c). In the standard numbering scheme, this is the end of the chain.

4.1.1 Taxonomies of amino acids

There are many ways that one can categorize the amino acid sidechains of proteins. We are mainly interested in protein interactions, so we will focus initially on a scale that is based on interactivity. We postpone until Chapter 7 a full explanation of the rankings, but suffice it to say that we rank amino acid sidechains based on their likelihood to be found in a part of the protein surface that is involved in an interaction.

In the following, we will use the standard terminology for the common twenty amino acids.[1] In Table 4.1 we recall the naming conventions and the RNA codes for each residue. Complete

---

[1]There are more than twenty biologically related amino acids that have been identified, but we will limit our study to the twenty primary amino acids commonly found.

Full name of amino acid	three letter	single letter	The various RNA codes for this amino acid
alanine	Ala	A	GCU, GCC, GCA, GCG
arginine	Arg	R	CGU, CGC, CGA, CGG, AGA, AGG
asparagine	Asn	N	AAU, AAC
aspartate	Asp	D	GAU, GAC
cysteine	Cys	C	UGU, UGC
glutamine	Gln	Q	CAA, CAG
glutamate	Glu	E	GAA, GAG
glycine	Gly	G	GGU, GGC, GGA, GGG
histidine	His	H	CAU, CAC
isoleucine	Ile	I	AUU, AUC, AUA
leucine	Leu	L	UUA, UUG, CUU, CUC, CUA, CUG
lysine	Lys	K	AAA, AAG
methionine	Met	M	AUG
phenylalanine	Phe	F	UUU, UUC
proline	Pro	P	CCU, CCC, CCA, CCG
serine	Ser	S	UCU, UCC, UCA, UCG, AGU, AGC
threonine	Thr	T	ACU, ACC, ACA, ACG
tryptophan	Trp	W	UGG
tyrosine	Tyr	Y	UAU, UAC
valine	Val	V	GUU, GUC, GUA, GUG
stop codons			UAA, UAG, UGA

Table 4.1: Amino acids, their (three-letter and one-letter) abbreviations and the RNA codes for them. For completenes, the "stop" codons are listed on the last line.
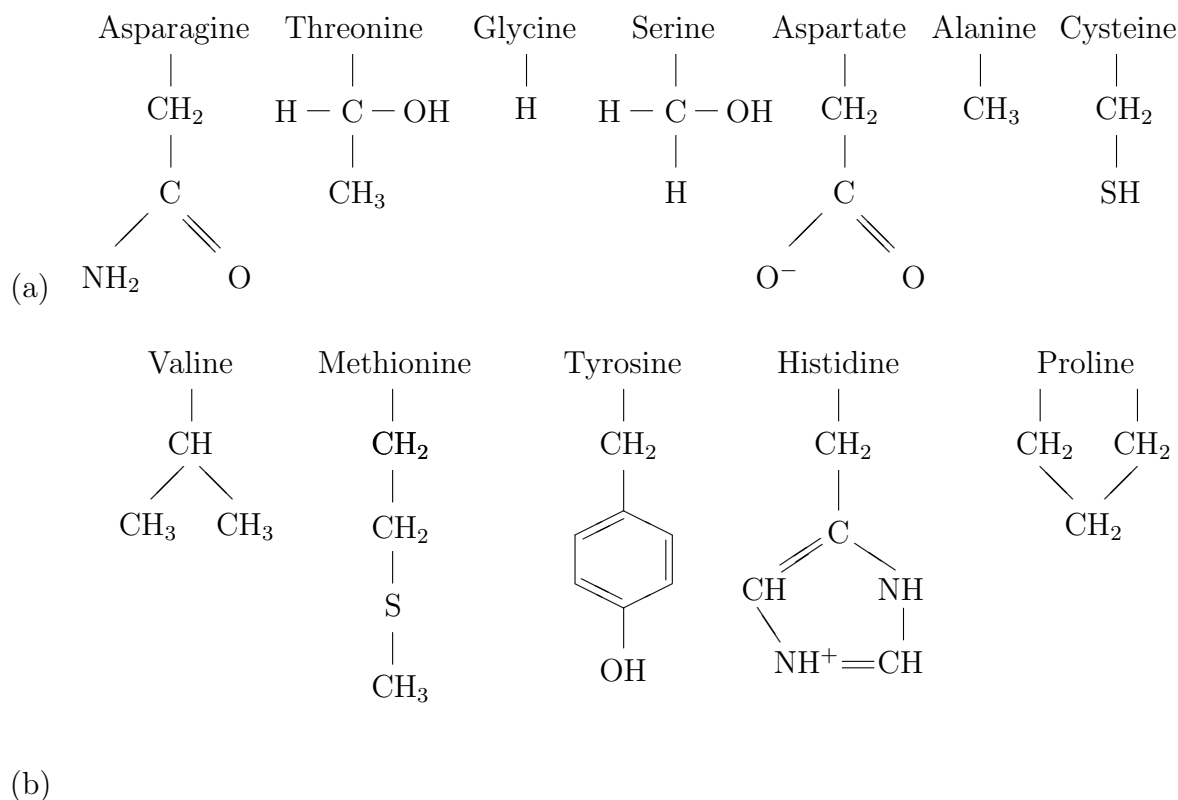
Figure 4.3: Periodic table of amino acid sidechains (residues). Not shown is the C_α carbon (see Figure 4.2), located at the top of the residue where the name appears. (a) The smallest, and most likely to be involved in protein-ligand interactions, ordered from the left (asparagine). (b) The middle ground in terms of interactivity.

descriptions of the sidechains for the amino acids can be found in Figures 4.3—4.4.

In Table 4.2, we present some elements of a taxonomy of sidechains. We give just two descriptors of sidechains, but even these are not completely independent. For example, all the hydrophilic residues are either charged or polar, and all of the neutral sidechains are hydrdophobic. However, some residues have both hydrophilic and hydrophobic regions, and are characterized as amphiphilic. For example, the aromatic residues are among the most hydrophobic even though two of them are polar, cf. Section 4.4.5. Even Phenylalanine has a small polarity that can interact with other molecules.

We focus here on the properties of individual sidechains, but these properties alone do not determine protein structure: the context is essential. Studying pairs of sidechains that are interacting in some way (e.g., ones that appear sequentially) gives a first approximation of context.

4.1.2   Wrapping of hydrogen bonds

A key element of protein structure is the protection of hydrogen bonds from water attack. A different taxonomy amino acids can be based on their role in the protection of hydrogen bonds. We will see in Chapter 7 that this correlates quite closely with the propensity to be at an interface.

Some hydrogen bonds are simply buried in the interior of a protein. Others are near the surface and potentially subject to water attack. These can only be protected by the sidechains of other nearby amino acids. Such protection is provided by the hydrophobic effect (cf. Section 2.5). The hydrophobic effect is complex [41, 392], but suffice it to say that a key element has to do with the fact that certain **non-polar groups**, such as the **carbonaceous groups** CH_n ($n = 1, 2, 3$), tend to repel polar molecules like water. They are non-polar and thus do not attract water strongly, and moreover, they are polarizable and thus damp nearby water fluctuations. Such carbonyl groups are common in amino acid sidechains; Val, Leu, Ile, Pro, and Phe have only such carbonaceous groups. We refer to the protection that such sidechains offer as the **wrapping of hydrogen bonds**. For reference, the number of nonpolar CH_n groups for each residue is listed in Table 4.2.

The standard thinking about sidechains has been to characterize them as being hydrophobic or hydrophilic or somewhere in between. Clearly a sidechain that is hydrophobic will repel water and thus protect anything around it from water attack. Conversely, a sidechain that is hydrophilic will attract water and thus might be complicit in compromising an exposed hydrogen bond. In some taxonomies [332], Arg, Lys, His, Gln, and Glu are listed as hydrophilic. However, we will see that they are indeed good wrappers. On the other hand, Ala is listed as hydrophobic and Gly, Ser, Thr, Cys and others are often listed as "in between" hydrophobic and hydrophilic. And we will see that they are among the most likely to be near underwrapped hydrogen bonds. This is not surprising since they are both polar (see Section 4.4.1) and have a small number of carbonaceous groups.

What is wrong with a simple philic/phobic dichotomy of amino acids is that the "call" of philic versus phobic is made primarily based on the final group in the sidechain (the bottom in Figures 4.3—4.4). For example, Lys is decreed to be hydrophilic when the bulk of its sidechain is a set of four carbonaceous groups. What is needed is a more complete picture of the role of all the groups in the entire sidechain. This requires a detailed understanding of this role, and in a sense that is a major object of this monograph. Thus it will require some in depth analysis and comparison with data to complete the story. However in the subsequent chapters this will be done, and it will appear that one can provide at least a broad classification, if not a linear ordering, of amino acid sidechains based on either their ability or propensity to wrap (or not) exposed hydrogen bonds or other electronic bonds.

The ordering of the most interactive proteins is based on a statistical analysis which is described in more detail in Chapter 7. We will also see there that these are likely to be associated with underwrapped hydrogen bonds. On the other hand, it is relatively easy to predict the order for good wrappers based on counting the number of carbonaceous groups. There is not a strict correlation between interactivity and bad wrapping, but a significant trend exists.

Arginine

Phenyl-
alanine

Tryptophan

CH_2

Leucine

CH_2

CH_2

CH_2

CH_2

CH_2

CH_2

CH_2

C

CH

CH_3 CH_3

CH

NH

NH

C

NH_2 NH_2^+

(c)

Lysine   Glutamate   Glutamine   Isoleucine

CH_2 CH_2 CH_2 $H - C - CH_3$

CH_2 CH_2 CH_2 CH_2

CH_2 C   C CH_3

CH_2 O^- O NH_2 O

NH_3^+

(d)
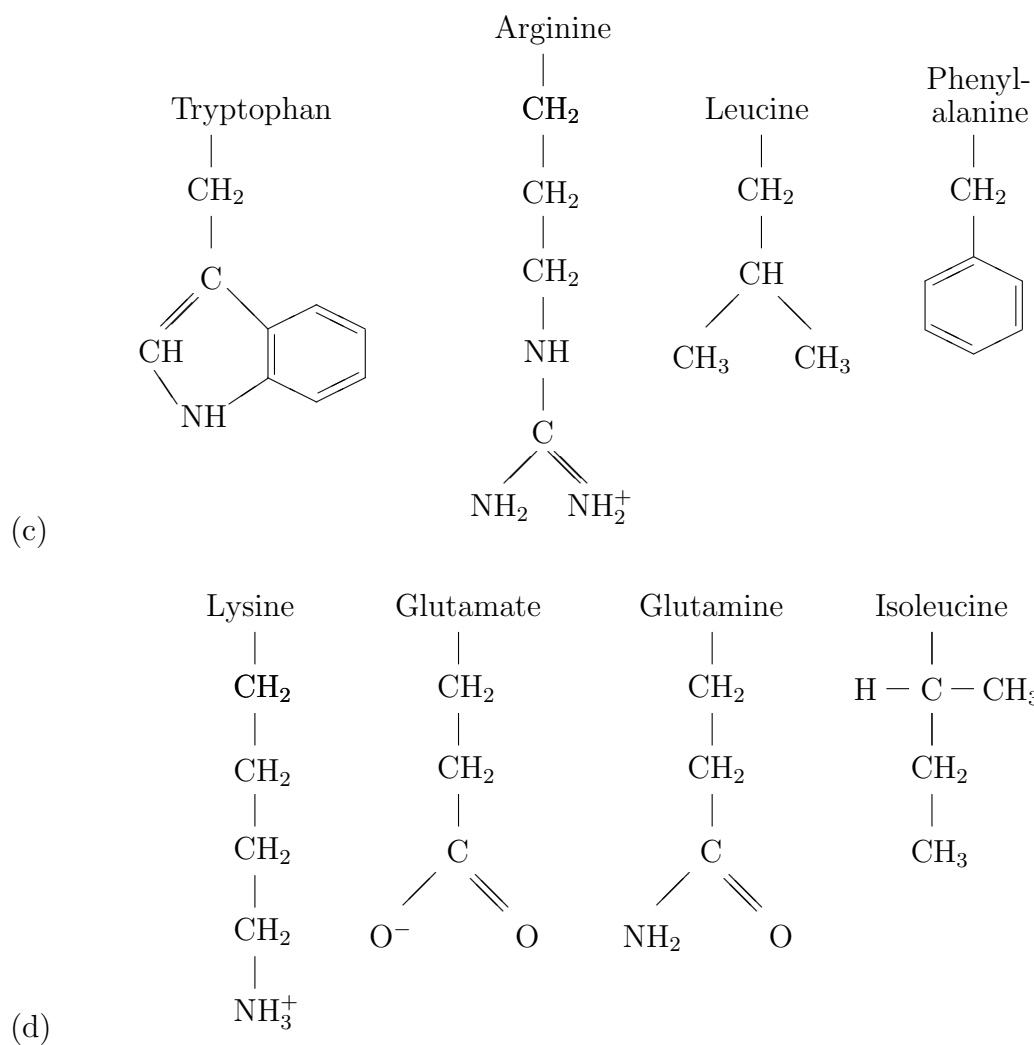
Figure 4.4: Periodic table of amino acid sidechains (residues). Not shown is the C_α carbon (see Figure 4.2), located at the top of the residue where the name appears. (c) Some amino acids less likely to be interactive. (d) The amino acids least likely to be involved in interactions.
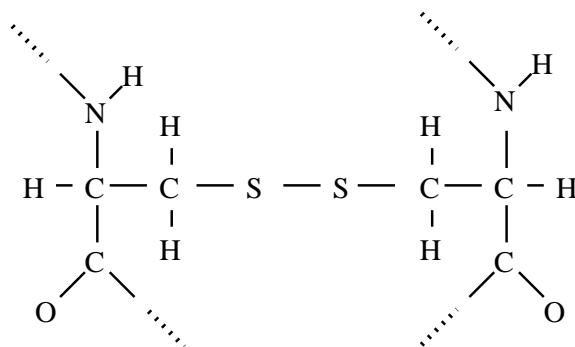
Figure 4.5: Schematic representation of a disulfide bond. The dotted lines indicate continuation of the peptide sequence. We leave as Exercise 4.10 to complete the drawing including the resonant bond character linking O-C-N as in Figure 4.1.

4.2 Special bonds

In addition to the covalent bonds of the backbone and the ubiquitous hydrogen bonds in proteins, there are two other bonds that are significant.

4.2.1 Salt Bridges

Certain sidechains are charged, as indicated in Table 4.2, and these can form an ionic bond, as depicted in Figure 3.2 in Section 3.1.2. Depending on the pH level, His may or may not be positively charge, but both Arg and Lys can be considered positively charged in most biological environments. Similarly, Asp and Glu are typically negatively charged. When sidechains of opposite charge form an ionic bond in a protein, it is called a **salt bridge** [84, 104]. Thus there are four (or six, depending on the charge on His) possible salt bridges.

Unmatched charged residues are often found on the surface of a protein where they can be solvated, but not inside a protein core where they would not make contact with water to neutralize the charge.

It is possible to define geometries to characterize salt bridges [104].

4.2.2 Disulfide bonds

Proteins are also held together by **disulfide bonds** or **disulfide bridges** which are bonds which form between two sulfurs on cysteines, as depicted in Figure 4.5. Specifically, the hydrogens attached to the sulfur atom on the two Cys sidechains are liberated, and a covalent bond forms between the two sulfur atoms. This is a much stronger bond than a hydrogen bond, but it is also much more specialized. It appears frequently in neurotoxins [171, 317]. These proteins would be highly disordered without the disulfide bridges. There appears to be a strong correlation between the number of disulfide bridges in proteins and their dehydron density (number of dehydrons as a

Full name of residue	three letter	single letter	water preference	charge or polarity	N	intrinsic pK_a	ΔV
Alanine	Ala	A	phobic	(backbone)	1	NA	-2.6
Arginine	Arg	R	amphi	positive	2	12	+7.9
Asparagine	Asn	N	philic	polar	1	NA	+7.0
Aspartate	Asp	D	philic	negative	1	3.7	+11.9
Cysteine	Cys	C	philic	polar	1	8.5	-1.0
Glutamine	Gln	Q	amphi	polar	2	NA	+1.3
Glutamate	Glu	E	amphi	negative	2	4.2	+8.5
Glycine	Gly	G	NA	(backbone)	0	NA	—
Histidine	His	H	philic	positive	1	6.5	+3.3
Isoleucine	Ile	I	phobic	neutral	4	NA	-2.6
Leucine	Leu	L	phobic	neutral	4	NA	-6.2
Lysine	Lys	K	amphi	positive	3	10.4	+7.6
Methionine	Met	M	amphi	polar	3	NA	+0.7
Phenylalanine	Phe	F	phobic	neutral	7	NA	-0.9
Proline	Pro	P	phobic	neutral	2	NA	-6.2
Serine	Ser	S	philic	polar	0	NA	+1.4
Threonine	Thr	T	amphi	polar	1	NA	+0.3
Tryptophan	Trp	W	amphi	polar	7	NA	+0.6
Tyrosine	Tyr	Y	amphi	polar	6	9.8	-0.3
Valine	Val	V	phobic	aliphatic	3	NA	-3.6

Table 4.2: A taxonomy of amino acids. The code for water interaction is: phobic, hydrophobic; philic, hydrophilic; amphi, amphiphilic. N is the number of CH_n groups in the sidechain. Values of pK_a for ionizable residues are taken from [396] (cf. Table 1.2 of [87]). The indication 'backbone' for the polarity of Alanine and Glycine means that the polarity of the backbone is significant due to the small size of the sidechain. The polar region of Tyrosine and Phenylalanine is limited to a small part of the side chain, the rest of which is neutral. When Histidine is not charged, then it is polar. ΔV is the change in volume in Å^3 [181] of sidechains between protein core and water (Section 5.4).
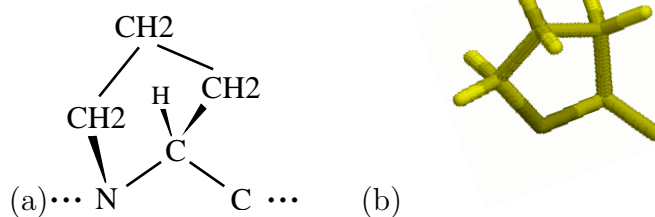
Figure 4.6: Proline sidechain configuration: (a) cartoon and (b) configuration of Pro165 in the PDB file 1QM0 (the hydrogen attached to the C_α carbon is not shown.

fraction of all hydrogen bonds): an increase in one disulfide bridge per hundred residues bonds correlates with an increase of five dehydrons per hundred backbone hydrogen bonds [131].

Disulfide bonds can also form between two separate proteins to form a larger system. This occurs in insulin and in antibodies.

4.3   Post-translational modifications

Proteins are not quite so simple as the protein sequence might imply. The term **post-translational modification** means changes that occur after the basic sequence has been set. Modifications (**glycosylation**, **phosphorylation**, etc.) add groups to sidechains and change the function of the resulting protein. A change in pH can cause the ends of some sidechains to be modified, as we discuss in Section 4.5.

Phosphorylation occurs by liberating the hydrogen atom in the OH group of Serine, Threonine and Tyrosine, and adding a complex of phosphate groups (see Section 12.2 for illustrations).

Phosphorylation can be inhibited by the presence of wrappers. Serine phosphorylates ten times more often than Tyrosine, even though the benzene ring presents the OH group further from the backbone.

Phosphorylation is expressed in PDB files by using a non-standard amino acid code, e.g., SEP for phosphoserine (phosphorylated serine), PTR for phosphotyrosine (phosphorylated tyrosine) and TPO for phosphothreonine (phosphorylated threonine).

4.4   Special side chains

There are many ways that sidechains can be classified, according to polarity, hydrophobicity and so on. When all such designations are taken into account, each sidechain becomes essentially unique. Indeed, it is advisable to study more complete descriptions of the unique properties of individual sidechains [87]. But there are some special properties of sidechains that deserve special mention here for emphasis.
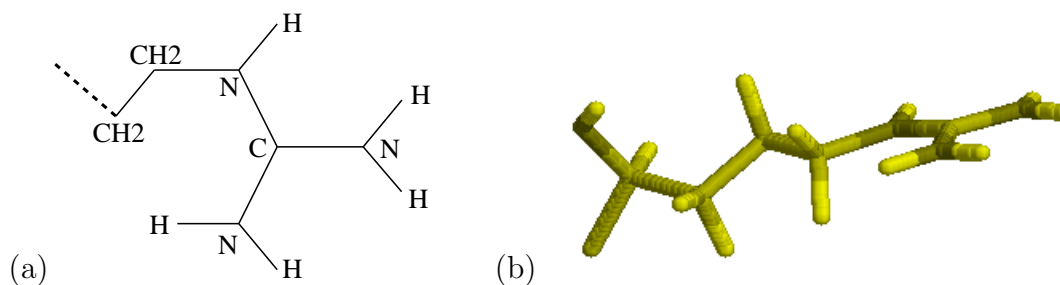
Figure 4.7: The planar structure of the terminal atoms of the Arginine sidechain: (a) cartoon in the plane of the terminal groups, (b) a rotated view of Arg220 in PDB file 1QM0 to show the planar structure from the side.

4.4.1 Glycine (and Alanine)

Glycine is special because it has essentially no sidechain. More precisely, it is the only aminoacid without a C_β carbon. As a result, it is appropriate to think of Gly as polar, since the polarity of the backbone itself has a significant impact on the environment near the sidechain. In this regard, alanine can also be viewed to be somewhat polar. Alanine has a C_β carbon, but no other heavy atoms in its sidechain, a feature unique to Ala.

4.4.2 Proline

Proline is unique because it connects twice to the backbone, as depicted in Figure 4.6. This causes a special rigidity not found with other residues. Proline has substantial implact on the stabilization of loops [14, 173, 175, 191, 281, 326]. Technically, it derives not from an amino acid but rather an imino acid [327]. The backbone nitrogen for proline has no attached hydrogen, due to the extra covalent bond from the sidechain. Thus the proline backbone cannot participate as a hydrogen bond donor. On the other hand, it can be a hydrogen bond acceptor.

    There is a special conformation of protein structures called PP2 (a.k.a. PPII or PII) which refers to the type of structure that a polyproline strand adopts [150, 370].

4.4.3 Arginine

The uniqueness of arginine is highlighted by the fact that the end of its residue closely resembles guanidine (CN_3H_5) and the guanidinium cation (CN_3H_6), cf. Figure 4.8. Compounds of guanidinium (e.g., guanidinium hydrochloride and guanidinium thiocyanate [284, 360]) have the ability to **denature** proteins, that is, to cause them to unfold [110, 434]. Urea (CON_2H_4, a.k.a. carbamide) is also a protein denaturant [361, 357, 62] and resembles guanadine except that the NH group is replaced by an oxygen, cf. Figure 4.8. How the denaturing process occurs is not fully understood [297, 360, 357], although the similarity of urea to the peptide backbone is assumed to play a significant role by intercolating between amide and carbonyl groups on the backbone that would otherwise be making hydrogen bonds [361, 62]. Urea can both denature proteins and dissociate protein complexes [357].
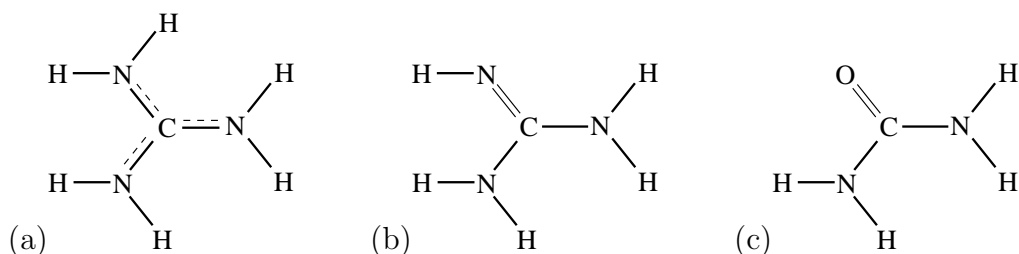
Figure 4.8: The molecules (a) guanidinium (ion), (b) guanidine and (c) urea. The net positive charge for the guanidinium ion (a) is distributed around the three nitrogen groups, as is indicated by the dashed lines.

One feature of the arginine residue is that the positive charge at the end of the residue is distributed quite broadly among the atoms at the end of the residue (see Table 8.5). How or why this might have a special effect is not completely explained.

It is very difficult to form natural water structures around the terminal (guanidinium) part of an arginine sidechain [284]. The atoms beyond the C_δ carbon are all in a plane, cf. Figure 4.7. Model building shows that it is impossible for waters attached to the terminal hydrogens (those attached to the three Nitrogens) to cohabitate, whereas the terminal N_3 group of Lysine is easily solvated, cf. Exercise 4.2. One can think of the planar structure of the terminal CN_3H_5 group as like a knife blade that cuts through water structures. Similarly, uric acid ($C_5H_4N_4O_3$, the cause of the disease gout) is a planar molecule that is not very soluble in water, despite its relation to urea, which is very soluble.

One property of arginine is that polyarginine is the polypeptide most able to cross a cell membrane without the help of a transporter molecule [294], and compounds rich in Arg have similar behavior [441].

4.4.4 Cysteine

What makes cysteine special is the ability of the sidechain to bond with another cysteine sidechain, making a disulfide bridge (Section 4.2.2). This is the only sidechain that forms a covalent bond with another sidechain.

4.4.5 Aromatic sidechains

Three sidechains (Tyr, Trp, Phe) have benzene rings as a significant part of their structure. At first, these appear simply hydrophobic, but the electron structure of aromatic rings is complex [105]. There is a doughnut of positive charge located in the plane of the carbon and hydrogen atoms, and the hole of the doughnut contains regions of negative charge extending to both sides of the main positive ring (see Figure 2A of [105]). This makes these side chains polar. Tyrosine is also polar in a more conventional way at the end of the sidechain due to the OH group there.

Tryptophan deserves special mention for various reasons, not just because of its pop-culture notoriety for sleep induction [307] and other behavioral impact [334]. It is the largest and most

complex sidechain, involving two types of rings, the indole ring in addition to the benzene ring.[2] Tryptophan is also the least common and most conserved (least likely to be mutated in homologous proteins, cf. Section 11.4) sidechain.

4.4.6 Remembering code names

Many of the single letter codes for sidechains are obvious (Alanine, Glycine, Histidine, Isoleucine, Leucine, Methionine, Proline, Serine, Threonine, Valine), but others require some method to remember. We propose here some non-serious mnemonic devices that may aid in retaining their assignments.

Asp and Glu are the negatively charged residues, and the alphabetic order also corresponds with the size order (Asp is smaller than Glu). The code names are also alphabetical (D and E); the choice of E corresponds to the charge *e* of the extra electron.

Two of the positive sidechains also have special codes. To remember the R for arginine is to think of it as the pirate's favorite sidechain. To "lyse" means to destroy or disorganize, so we can think of lysine as the Killer sidechain.

The biggest sidechains (the aromatic ones) also have letter codes which need special treatment. A way to remember the single letter code for Phe is to misspell it with the Ph changed to F. A way to remember the single letter code for Trp is that it is the Widest sidechain. A way to remember the single letter code for Tyr is to focus on the second letter Y in the name. The sidechain also looks like an inverted Y on top of another Y.

The two remaining proteins are comparable to Asp and Glu, but with nitrogen groups replacing one of the oxygens: asparagiNe and Qlutamine. The emphasis on Nitrogen is clear in Asn, since it is the third letter of the code. The letter G is one the most overloaded among first letters in the sidechain names, but Q is a close match for a poorly written G.

4.5 Sidechain ionization

We will not consider extensively *pH* effects, although these clearly involve a type of modulation of electrical forces. There is significant *pH* variation in different parts of cells, and thus it has a potential role in affecting protein-ligand interactions.

The effects of *pH* are both localized and dynamic in nature, since the number of ions that can be involved in protein-ligand interactions is not large. For example, a well solvated large biomolecule [421] can be modeled dynamically with just over 10^5 atoms, and significantly less than 10^5 water molecules. But at *pH* 7, there is just one hydronium molecule per 5.5508×10^8 water molecules (cf. Section 17.7.1). The number of water molecules in the simulation in [421] used fewer than 55,508 water molecules, and thus would not have included a hydronium ion until the *pH* was less than three. On the other hand, ions cluster around proteins since they have charged and polar residues, so a more complex model is required to account for their effects.

---

[2]In this regard tryptophan shares structure similar to the compound psilocybin which is known to fit into the same binding sites as the neurotransmitter serotonin.

The ends of some sidechains can vary depending on the ionic composition of the solvent [87]. The pH value (Section 17.7.1) relevant for ionization is out of the range of biological interest in most cases, with the exception of His.

We list the intrinsic pK_a values [87, 319, 396] in Table 4.2 for reference. This value is the pH at which half of the residues would be in each of the two protonation states. For example, for pH below 3.7, Asp would be more likely to be protonated, so that one of the terminal oxygens would instead be an OH group, as shown in Figure 4.9. In this case, it would be appropriate to refer to the residue as aspartic acid. Similarly, for pH below 4.2, Glu would more likely have an OH terminal group, as shown in Figure 4.9, and be called glutamic acid. By contrast, for pH above 8.5, a Cys residue would tend to lose the terminal hydrogen. Correspondingly, the other sidechains with $pK_a > 9$ in in Table 4.2 would also lose a hydrogen.

The relation between pH and pK_a is mathematically simple:

$$
\begin{aligned}
pK_a &= -\log_{10}\frac{[A^-][H^+]}{[AH]} = -\log_{10}\frac{[A^-]}{[AH]} - \log_{10}[H^+] \\
&= -\log_{10}\frac{[A^-]}{[AH]} + pH = \log_{10}\frac{[AH]}{[A^-]} + pH,
\end{aligned}
\tag{4.1}
$$

where $[X]$ denotes the equilibrium number of molecules X and AH denotes the protonated state of molecule A. Thus when $pH < pK_a$, the protonated state AH is more likely than the deprotonated state A^-. However, both states still coexist at all pK_a values; when $pK_a = pH - 1$, $[AH]$ is still one tenth of $[A^-]$. We should note that these statements are correct only for molecules A in pure water.

For simplicity, we refer to the residues in their form that is most common at physiological values of pH. The one sidechain that has a pK_a value near physiological pH (≈ 7) is His [42]. There are two ways in which His can become deprotonated, as indicated in Figure 4.9. The particular choice of site depends on factors similar to those that determine hydrogen location (cf. Section 6.2).

The concept of pK_a is not just a global concept. There is a local pK_a [7, 90, 156, 167, 169, 210, 224, 260, 291, 319, 389, 395] for each sidechain in its own particular local environment, and these values can be biologically relevant in many cases. For example, the pK_a of Glu35 in hen and turkey lysozyme is just over 6 instead of the nominal value of 4.2 [37]. Thus the protonation of the carboxyl sidechain of Glu can occur at biologically relevant values of pH in some cases. The particular oxygen that gets protonated will also be dependent on the local environment (cf. Section 6.2). The pK_a can also fluctuate in time, as local changes in chemical environment occur [287].

4.6   Salt ions

In addition to the pH variation, proteins are affected by small amounts of salt ions, such as Potassium and Sodium. They occur in greater volume than hydronium ions, but they have less mobility.
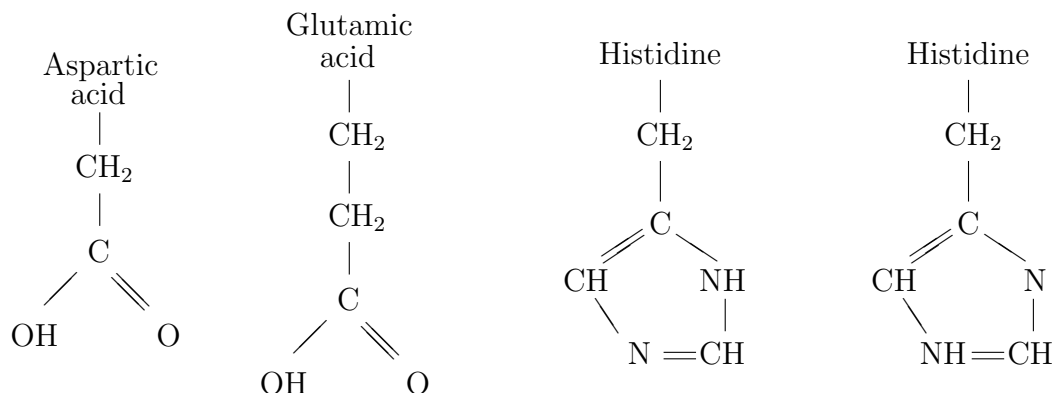
Figure 4.9: Neutralized sidechains. Either oxygen on Asp and Glu could be protonated. There are two forms of His as shown with either nitrogen being protonated [336].

4.7 Amino acid frequencies

We now consider a simple question as an introduction to datamining. We will see that there are two major components to datamining:

- the choice of lens (or filter or metric)

- the choice of dataset.

Here we take a simple lens: we ask about the presence or absence of residues in PDB files. In Table 4.3, we list the frequencies of each residue, as a percentage. To measure this, we count the number of time each one appears in the set of PDB files, we divide this by the total number of residues counted, and then multiply by 100 to get a percentage.

To be precise, if we have a dataset with N different types of characteristics (e.g., $N = 20$ and the characteristics are the different amino acid types), then the **frequency** f_i of the i-th characteristic is defined by

$$f_i = \frac{o_i}{\sum_{j=1}^{N} o_j},\tag{4.2}$$

where o_j is the number of occurrences of the j-th characteristic in the dataset. The f_i's form a probability distribution since they sum to 1. In some cases, frequencies are represented as percentages, in which case we simply multiply by 100 in (4.2) to get $p_i = 100 f_i$. The sum of the p_i's is then 100.

We will extend this later in Section 7.4 when we look at some more sophisticated metrics for examining the likelihood of finding things with a given property.

4.7.1   Choice of database

To make the exercise more interesting, we look at different datasets, and we see both similarities and differences that are significant. To start with, we have taken for simplicity the abundances published in [53], listed in column two in Table 4.3. We do not claim that the small set of proteins used in [53] provides the optimal reference to measure relative abundance in this setting, but it certainly is a plausible data set to use. Moreover, the evolution of amino acid abundances studied in [53] is a significant issue in its own right.

The datasets used to generate the third through sixth columns in Table 4.3 were based on the 124,729 protein chains analyzed in the non-redundant PDB dated 3 April 2009 (nrpdb.040309.txt). PDB chains listed as cluster center (group rank 1), using the BLAST pvalue 10^{-7} to determine similarity, were included (there are 14,588 such PDB chains). Of these, 9,141 chains with flaws in the sidechains (a nonzero value in columns G, H, I, or J in nrpdb.040309.txt) were eliminated, and PDB files having a resolution worse than 2 Ångstroms were also eliminated. Also eliminated were files having no useful temperature factors. This elimination process left a total of 1,989 representative chains. There were 409,730 total residues analyzed, and the corresponding frequencies are listed in column three in Table 4.3.

We will see in Section 5.1 that there are distinctly different parts of proteins based on their secondary structure, which can be described in three groups: helices, sheets and loops. The latter is defined to be the complement of the union of the former two classes. Of the total 409,730 residues in the nonredundant dataset, 162,565 are in helices, 91,665 are in sheets, and 155,481 are in loops (19 residues had conflicting structure information). The fourth, fifth and sixth columns in Table 4.3 differentiate between where the residues occur in secondary structure, helices and sheets or loops. Loops are by definition less structured regions of proteins. Indeed, the latter regions may not have any structure at all, and the corresponding protein sequences are likely underrepresented in the PDB because they frequently cannot form any stable structure, and thus are not imaged via the experimental techniques used to determine the data represented in the PDB [201].

We have a total of five datasets represented in Table 4.3, one for each column. However, the dataset corresponding to the third column is the union of the three datasets representing the final three columns. We would therefore expect that the values in the third column would lie between the min and the max of the three values in the fourth through sixth columns, and indeed they do in all cases. However, the values in the second column come from an unrelated dataset, and it is not surprising that the correlation is less simple. Nevertheless, the values in the second column lie between the values in the fourth through sixth columns 65% of the time, and four of the outliers are Cys, His, Met, and Trp, the four least frequently found sidechains. More importantly, the orderings for the two datasets are similar: both datasets indicate that Leucine is the most prevalent residue overall, and that Cys, His, Met, and Trp are the least common. However, we also see that the prevalence of Leu is based in helices and sheets; its prevalence in loops is about what you would expect by chance (5 percent).

Two other outliers of interest are Gly and Pro; comparing the last four columns, we see that these become dominant in loops, representing more than one-fifth of the residues in loops in the PDB. This suggests that four-fifths of all residues in loops are neighbors of Gly and Pro, if not one of them. Pro makes the structure very rigid nearby, and Gly enhances flexibility. Pro and Gly

Res.	[53]	All	Helix	Sheet	Loop
Ala	7.77	8.46	11.4	6.38	6.66
Arg	6.27	4.68	5.41	4.33	4.14
Asn	3.36	4.62	4.17	2.86	6.13
Asp	5.42	5.89	5.80	3.30	7.52
Cys	0.78	1.48	1.22	1.85	1.55
Gln	3.15	3.70	4.55	2.95	3.25
Glu	8.59	6.23	8.06	4.45	5.37
Gly	7.30	8.09	5.21	5.38	12.7
His	1.92	2.26	2.11	2.27	2.42
Ile	6.66	5.50	5.14	9.63	3.43
Leu	8.91	8.48	10.5	9.66	5.70
Lys	7.76	5.79	6.41	4.71	5.78
Met	2.41	1.97	2.27	1.90	1.70
Phe	3.61	4.01	3.91	5.70	3.11
Pro	4.35	4.64	2.66	1.91	8.32
Ser	4.66	6.15	5.80	5.10	7.14
Thr	4.87	5.79	4.79	6.94	6.15
Trp	1.02	1.56	1.58	2.01	1.26
Tyr	3.00	3.71	3.41	5.44	2.99
Val	8.17	6.99	5.66	13.2	4.70

Table 4.3: Amino acids frequencies (as percentages) in different data sets. The first column gives the three-letter code, and the second column is the frequency reported in [53]. The third through sixth columns are based on the non-redundant dataset described in the text. The third column is the frequency in that dataset for all residues, the fourth column is the frequency in helices, the fifth column is the frequency in sheets, and the sixth column is the frequency in loops.

play the role of bookends for the rigidity-flexibilty axis, and their prevalence suggests a concept of structure in loops.

We also see that some residues are more likely to be found in a helix or sheet: Ala and Glu are examples of the former, Ile and Val are examples of the latter.

4.7.2   PDB details

Although the datamining task described here seems simple, it already introduces a serious challenge in dealing with PDB files. The exact location of many atoms is not known, and **alternate locations** for them are sometimes listed in a PDB file. Their presence is indicated by the **altloc** indicator in the 17th column of a PDB file. Typical indicators are captital letters at the beginning of the alphabet. So it would be common to see ALYS and BLYS with the same residue sequence number. If such occurrences were not flagged, some residues might be overcounted.

It might seem that we could just check the residue sequence number: columns 23 through 26 in a PDB file. However, this is not sufficient, because there are other reasons that the residue sequence number might be the same for different entries. The **icode** or **insertion code** (column 27) allows for residues to be inserted in a sequence without altering some previously established numbering. Thus we need to check both the altloc and icode columns to get the right answer.

PDB files based on NMR data can contain many "models" of a protein complex. Each model is a complete representation of the complex, and there can be dozens of them. We can think of the different models as snapshots at different times. Since the models all represent the same protein(s), we have to consider only one of them in counting frequencies of residues and other features. Each model is preceded by a line in the PDB file of the form

```
MODEL        nm
```

where `nm` represents the model number (it could be just a single digit). In parsing a PDB file, you have to look for these indicators to go by to know which model you are looking at.

A PDB file often represents a complex of several proteins. The individual proteins are in unique **chains**, indicated by a single character in column 22. Often one wants to limit the analysis of frequency data to a specific chain in a PDB file. The **nonredundant PDB** (**NRPDB**) is a database of individual PDB chains that represent to varying degrees unique folds in the PDB. This database can be crucial in terms of avoiding bias in sampling properties of PDB files.

Finally, there is a small challenge in terms of reporting the frequencies of residues in secondary structure. A very small number of residues are ambiguously reported as being in both helix and sheet. This number is small enough that it can typically be ignored, but it can lead to confusing results if one expects the number of residues in helix, sheet and loop regions to add up exactly.

4.7.3 Hetatoms

In addition to the standard amino-acid residues discussed so far, PDB files include additional molecules referred to as **hetatoms**. The most common hetatom is water, with three-letter code HOH. Some hetatoms appear in the same way as regular amino-acid residues, such as PTR. This is a version of Tyrosine modified by phosphorylation, the covalent bonding of a phosphate group (PO_4) to the end of the sidechain of Tyrosine; some of these novel residues are listed in Table 4.4. Other hetatoms, like HOH, stand apart, not covalently bonded to other parts of the PDB strucure elements; some of these are listed in Table 4.5.

4.8 Exercises

Exercise 4.1 *Draw all the atoms in the tri-peptide GAG, including the C-terminal and N-terminal ends.*

Exercise 4.2 *Using a model set, build the terminal atoms for Lys and Arg together with some water molecules bound to them. Use this to explain why it is easy to solvate Lys and hard to solvate Arg.*

PDB code	sidechain name	analog/comment/formula
ACE	acetyl group	truncated residue
AIB	alpha-aminoisobutyric acid	truncated Val (C_β removed)
DAL	D-Alanine	D conformation of Ala
DBU	(2e)-2-aminobut-2-enoic acid	$C_4H_7NO_2$
HYP	hydroxyproline	OH group attached to C_γ of Pro
MSE	Selenomethionine	Methionine (S replaced by Se)
PCA	pyroglutamic acid	$C_5H_7NO_3$
PHL	L-Phenylalaninol	Phenylalanine [371]
PTR	phosphotyrosine	phosphorylated tyrosine
SEP	phosphoserine	phosphorylated serine
TPO	phosphothreonine	phosphorylated threonine

Table 4.4: Some common hetero-atoms in PDB files which appear in peptide sequences.

PDB code	chemical formula	common name
CA	Ca^+	calcium ion
CL	Cl^-	chloride ion
CCN	C_2H_3N	acetonitrile
GAL	$C_6H_{12}O_6$	beta-d-galactose
HOH	H_2O	water
MOH	CH_4O	methanol
NAG	$C_8H_{15}NO_6$	n-acetyl-d-glucosamine
NGA	$C_8H_{15}NO_6$	N-acetyl-D-galactosamine

Table 4.5: Some common hetero-atoms in PDB files which appear as independent molecules.

Exercise 4.3 *What is the typical distance between the two oxygens at the end of the sidechains of Asp (OD1 and OD2) and Glu (OE1 and OE2)? (Hint: take a small set of PDB files and compute the distribution of distances for all occurrences in those files.)*

Exercise 4.4 *What is the typical distance between the two heavy atoms at the end of the sidechains of Asn (OD1 and ND2) and Gln (OE1 and NE2)? (Hint: take a small set of PDB files and compute the distribution of distances for all occurrences in those files.)*

Exercise 4.5 *Three points determine a plane, but there are four atoms in some sidechains that are supposed to lie in a plane (e.g., Asp and Glu). Determine the plane that best fits four points in three dimensions by least squares. Test this on Asp and Glu in several PDB files. How big is the deviation of the four points from the plane (compute the squared deviation)?*

Exercise 4.6 *Another way to determine how planar four points are in three-dimensional space is to compute the volume of the tetrahedra spanned by these four atoms (as vertices). Test this on Asp and Glu in several PDB files, and contrast this with the corresponding data for Leu (CB, CG, CD1, CD2).*

Exercise 4.7 *Determine how planar the terminal four atoms (Cx, Cy, Oz1, Oz2) are in Asp (x,y,z=B,C,D) and Glu (x,y,z=C,D,E) in several PDB files. To determine planarity, use the techniques of both Exercise 4.5 and Exercise 4.6. How do the distributions of planarity measure differ? Contrast this with the corresponding data for Leu (CB, CG, CD1, CD2).*

Exercise 4.8 *Scan a large set of PDB files to see which ones have the largest number of Cys residues. Determine which ones are in a disulfide bond based on proximity of the terminal sulfurs. Plot the distribution of PDB files as a function of numbers of disulfide bonds and the distribution of PDB files as a function of numbers of isolated Cys residues. Redo the calculations plotting instead using numbers of disulfide bonds (respectively, numbers of isolated Cys residues) per hundred residues in each PDB file. What are the proteins (and their function) in the PDB files with the top ten numbers of disulfide bonds (respectively, numbers of isolated Cys residues) per hundred residues in each PDB file? Compare [131].*

Exercise 4.9 *Determine how long each sidechain is by scanning the PDB. That is, determine the distribution of distances from the C_α carbon to the terminal (heavy) atoms for each residue (amino acid) type (thus ignore Gly). Is the furthest atom always the same for each sidechain? If not, give the distribution for the different atoms. Is the atom furthest from C_α always in the sidechain (hint: check the small residues carefully).*

Exercise 4.10 *In Figure 4.5, the resonant bond linking O-C-N has been omitted. Re-draw this figure including the rest of the bonds in the peptide unit as done in Figure 4.1.*

Exercise 4.11 *Repeat the experiment reported in Table 4.3 with a different set of proteins. Pick a set of proteins according to some rule (e.g., all tyrosine kinases) and count the frequencies of the different amino acid residues. Compare with Table 4.3.*

Exercise 4.12 *Write a code to count the number of residues in a PDB file both by ignoring the altloc indicator and by taking it into account. Test the code on PDB file 3EO6. Verify by another technique that your count is correct.*

Exercise 4.13 *Write a code to count the number of residues in a PDB file both by ignoring the icode indicator and by taking it into account. Test the code on PDB file 1BIO. Verify by another technique that your count is correct.*

Exercise 4.14 *Compare the frequency odering of residues given by columns 2 and 3 in Table 4.3. Determine the extent to which you can cluster groups of residues that are ordered in the same way by both columns. For example, Leu is the most frequent by both columns, and Cys, His, Met, and Trp are the least common according to both. What group of residues is second-most common according to both columns (and so forth)?*

Chapter 5

Protein Structure

We now review the basic ideas about protein structure. An example of a five-residue peptide sequence is given in Figure 2.1. Such an all-atom representation is too busy in most cases, so it is useful to look for a simpler representation. A cartoon of a peptide sequence is depicted in Figure 5.1. Here we retain only certain features, such as C_α carbon and attached residue, and both the dipole (cf. Figure 3.4) and the double bond of the peptide backbone. This representation is useful in talking about the types of hydrogen bonds that are formed in proteins, as we depict in Figure 5.3. However, in many cases, only the sequence of residues is significant. The representation of proteins as a linear sequence of amino acid residues is called the **primary structure**. More precisely, we can represent it as a string of characters drawn from the twenty-character set A, C, D, ..., Y, W in column three in Table 4.1. For example, the structure of the six-residue DRYYRE [146] is discussed in Section 10.5.3.

The primary structure representation of proteins is fundamental, but it does not directly explain the function of proteins. Essentially all proteins only function in a three-dimensional structure. That structure can be described in a hierarchical fashion based on structural subunits, as we now describe.

5.1   Hydrogen bonds and secondary structure

Proteins can be described using a hierarchy of structure. The next type after the primary, linear structure is called **secondary structure**. Many components of secondary structure have been identified, but the main ones may be catagorized by two primary types: alpha-helices and beta-sheets (a.k.a., α-helices and β-sheets). These form the basic units of secondary structure, and they can be identified in part by the pattern of hydrogen bonds they make, as depicted in Figure 5.3. These units combine to form 'domains' or 'folds' that are characteristic structural patterns that can be viewed as widgets used to build protein structure, and presumably govern its function. Structural relationships among these widgets form interesting networks, as will be described in Section 5.5.
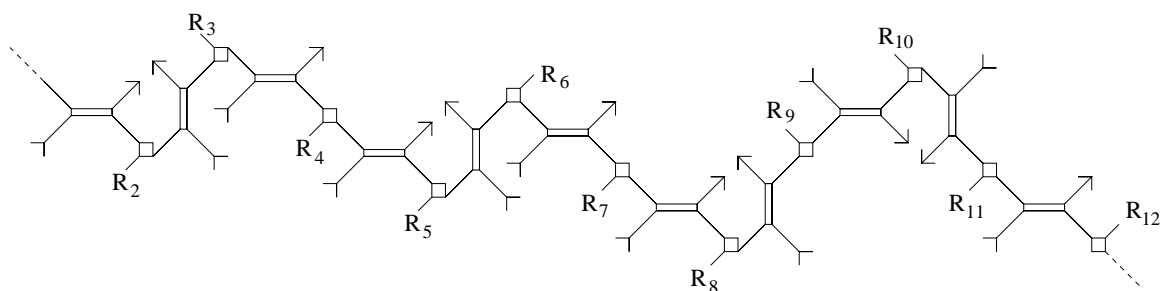
Figure 5.1: Cartoon of a peptide sequence where all of the peptides are in the trans form (cf. Figure 4.2). The small boxes represent the C-alpha carbons, the arrow heads represent the amide groups NH, the arrow tails represent the carbonyl groups CO, and the thin rectangular boxes are the double bond between the backbone C and N. The different residues are indicated by R's. The numbering scheme is increasing from left to right, so that the arrow formed by the carbonyl-amide pair points in the direction of increasing residue number. The three-dimensional nature of the protein is left to the imagination, but in particular where the arrow heads appear to be close in the plane of the figure they would be separated in the direction perpendicular to the page.

5.1.1 Secondary structure units

Alpha helices are helical arrangements of the subsequent peptide complexes with a distinctive hydrogen bond arrangement between the amide (NH) and carbonyl (OC) groups in peptides separated by k steps in the sequence, where primarily $k = 4$ but with $k = 3$ and $k = 5$ also occurring less frequently. An example of a protein fragment that forms a rather long helix is given in Figure 5.2. What is shown is just the 'backbone' representation: a piecewise linear curve in three dimensions with the vertices at the C_α carbons. The hydrogen bond arrangement in a helix is depicted in Figure 5.3(a) between two such peptide groups. The helices can be either left-handed or right-handed, but protein structures are dominated by right-handed helices [218, 313, 379].

Beta sheets represent different hydrogen bond arrangements, as depicted in Figure 5.3: (b) is the anti-parallel arrangement and (c) is the parallel. Both structures are essentially flat, in contrast to the helical structure in (a).

Both alpha-helices and beta-sheets can be distinguished based on the angles formed between the protein backbone units, as described in Section 5.2. It is also possible to find distinctive patterns in the distribution of residues found in helices and sheets [419]. For example, the residues most likely found in helices were found to be E, A, L, M, Q, K, R, H (in order of likelihood), in sheets V, I, Y, C, W, F, T, and in turns G, N, P, S, D. The distributions of pairs of residues in various structures is even more distinctive [266].

Other structural units include turns and loops (or coils). The former involve short (three to four) peptide units, whereas the latter can be arbitarily long. The PDB classification eliminated 'turn' as a special classification in 2009. The term **loop** is now taken to mean the complement of α-helix and β-sheet regions in protein sequence.

There is a characteristic alternation of hydrophobic and hydrophilic sidechains in helices and sheets [112]. Not suprisingly, the frequency of alternation is approximately two in beta-sheets,
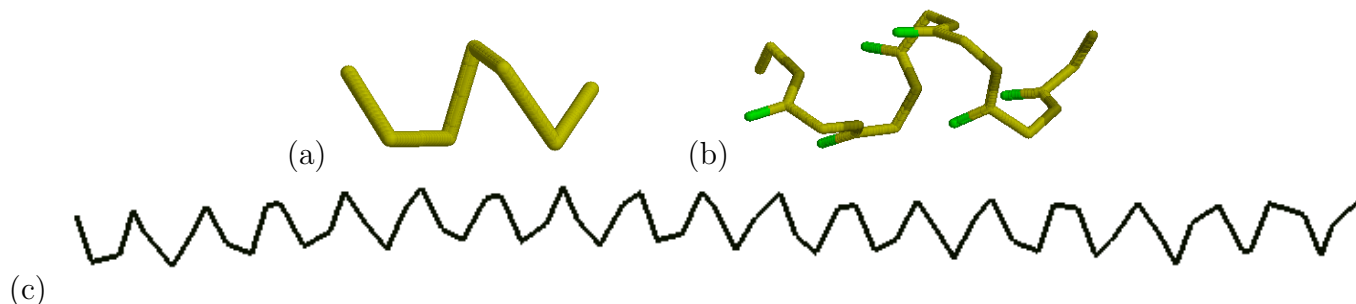
Figure 5.2: Two helices (one short, one long) in chain B in PDB file 1HTM, which depicts the Influenza Hæmagglutinin HA2 chain. (a) and (b) represent the helix formed by the seven residues B-Asn-146 through B-Ile-152: (a) shows only the backbone and (b) indicates all of the backbone heavy atoms, showing in particular the direction of the carbonil (C-O) orientation, pointing toward the four nitrogen residues ahead in the sequence. (c) is the backbone representation of the helix formed by the sixty-six residues B-Ser-40 through B-Gln-105.
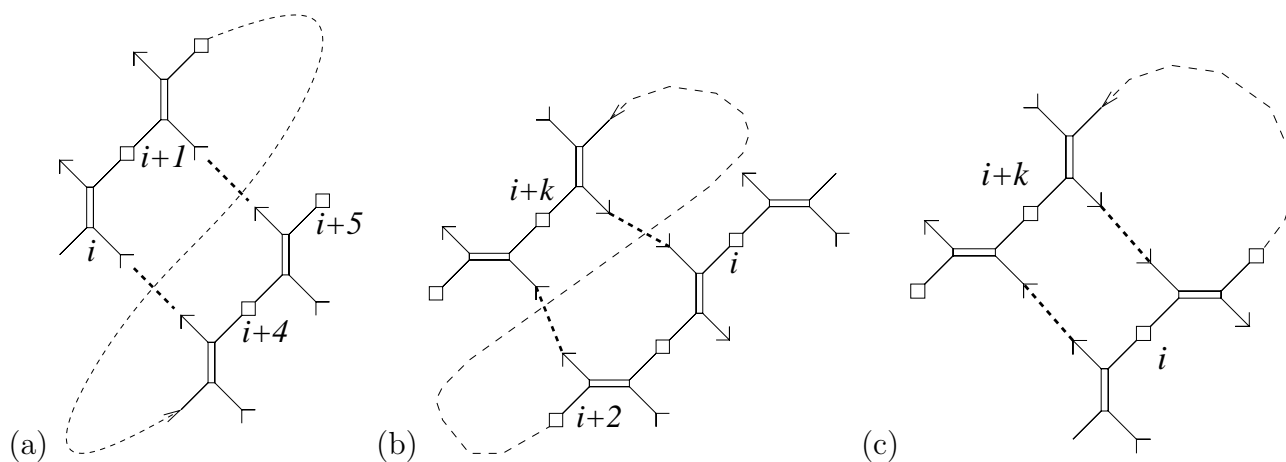
Figure 5.3: The hydrogen bond (dashed line) configuration in (a) α-helix, (b) parallel β-sheet, and (c) antiparallel β-sheet. The dotted lines indicate how the backbone is connected. The amide (N-H) groups are depicted by arrow heads and the carbonyl (O-C) groups are depicted by arrow tails, thus indicating the dipole of the backbone.
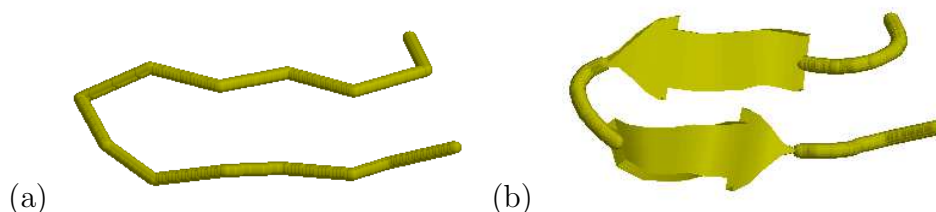
Figure 5.4: Two views of the peptide in PDB file 1K43 which forms an antiparallel β-sheet: (a) backbone, (b) ribbon cartoon.
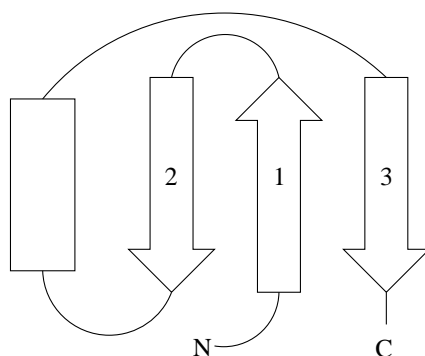
Figure 5.5: Example of a simple fold with three beta sheets and one alpha helix (in the primary sequence between sheets 2 and 3). The letters N and C indicate the N-terminal and C-terminal ends. Proximity of the edges of the beta sheets indicates hydrogen bonding.

so that one side of the sheet tends to be hydrophobic and the other side hydrophilic. In alpha-helices, the period closely matches the number k of residues between the donor and acceptor of the mainchain hydrogen bonds.

5.1.2   Folds/domains

A **fold** or **domain** is a collection of basic structural units, as defined in Section 5.1.1, together with topological information on relations among the basic units [424]. The Structural Classification of Proteins (or SCOP) database [8] provides a classification of these folds [424]. A hypothetical example is depicted in Figure 5.5.

The topological representation of folds can be viewed as a type of language based on the basic units (α and β) as characters. This set of characters can be extended to include other units (e.g., turns). This linguistic approach to structure has been used as the basis of approaches to secondary structure prediction [15, 111, 63]. The I-sites library of a dozen or so structural units has been used [63] to facilitate the prediction. Thus we might think of folds (domains) as representing words in a linguistic representation of structure. With this view, proteins represent phrases in the language with particular significance. A single protein can consist of a single fold, or it may be made up of several different folds.

SCOP release 1.73 (November 2007) divides proteins into seven classes of protein structure

Name of fold classes	number of folds in class
All alpha helix proteins	258
All beta sheet proteins	165
Alpha and beta proteins a/b	141
Alpha and beta proteins a+b	334
Multi-domain proteins alpha and beta	53
Membrane and cell surface proteins and peptides	50
Small proteins	85
Coiled coil proteins	7
Low resolution protein structures	26
Peptides	120
Designed proteins	44

Table 5.1: The major classes in SCOP release 1.73. The a/b class consists of mainly parallel beta sheets, whereas the a+b class consists of mainly antiparallel beta sheets. The last four lines of the table are not considered 'true classes' of protein folds.

groupings, together with four additional groupings of special cases. These are listed in Table 5.1, which shows the number of folds for each class. Not counting the last four classes, there are 1086 folds represented (cf. [425]). Thus we can think of the set of known folds as a small dictionary of the words formed in the language based on the characters of secondary structural elements.

Combinations of folds interact [264] to form a variety of structures; the shape that they adopt is called the **tertiary structure**. The combination of several proteins in a unified (functional) structure is called a **protein complex**. The shape that a protein complex adopts is called its **quartenary structure**.

5.2 Mechanical properties of proteins

The Protein Data Bank (PDB) supports a simple mechanical view of proteins. The positions of the backbone and sidechain atoms are specified, together with the positions of some observed water molecules and other atoms. This basic information allows the derivation of extensive additional information, as we will explain subsequently. But for the moment, we simply recall some information on the static description of proteins.

5.2.1 Conformational geometry of proteins

We recall the basic ingredients of the peptide group from Figure 4.2. If x is a given residue, then $N(x)$, $H(x)$, $C(x)$ and $O(x)$ denote the position vectors of the corresponding atoms in the peptide group. For the remaining atoms, the standard notation from the PDB is as follows:

$$C_\alpha(x), C_\beta(x), C_\gamma(x), C_\delta(x), C_\epsilon(x), C_\zeta(x), C_\eta(x)$$
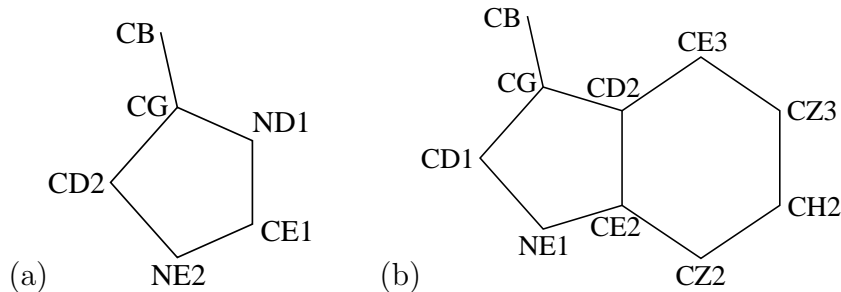
Figure 5.6: PDB notation for (a) histidine and (b) tryptophan.

are the $\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \eta$ carbons (denoted in plain text in the PDB by CA, CB, CG, CD, CE, CZ, CH) in the sidechain structure of residue x. Most of these can also appear with subscripts, e.g., $C_{\gamma i}$ for $i = 1, 2$ in Ile and Val. Correspondingly, $N_{\delta i}(x), N_{\epsilon i}(x), N_{\eta i}(x)$ are the i-th δ, ϵ, η nitrogens, denoted in plain text in the PDB by NDi, NEi, NHi for $i = 1, 2$. Notation for oxygens is similar. Unfortunately, the plain text descriptor OH for O_η in Tyr is a bit confusing, since this oxygen has an attached hydrogen. The PDB descriptors for Histidine and Tryptophan are depicted in Figure 5.6.

We can view $C_\alpha(x), N_{\delta i}(x)$, etc., as three-dimensional vectors, using the corresponding coordinates from the PDB. For amino acids x_i, x_{i+1} which are adjacent in the protein sequence, the *backbone vector* is defined as

$$\mathcal{B} = C_\alpha(x_{i+1}) - C_\alpha(x_i). \tag{5.1}$$

The *sidechain vector* $\mathcal{S}(x)$ for a given amino acid x, defined by

$$\mathcal{S}(x) = C_\beta(x) - C_\alpha(x), \tag{5.2}$$

will be used to measure sidechain orientation. $\mathcal{S}$ involves the direction of only the initial segment in the sidechain, but we will see that it is a significant indicator of sidechain conformation. For $x = Gly$, we can substitute the location of the sole hydrogen atom in the residue in place of C_β. For each neighboring residue pair x_i, x_{i+1}, the sidechain angle $\theta(x_i, x_{i+1})$ is defined by

$$\cos \theta(x_i, x_{i+1}) = \frac{\mathcal{S}(x_i) \cdot \mathcal{S}(x_{i+1})}{|\mathcal{S}(x_i)||\mathcal{S}(x_{i+1})|}, \tag{5.3}$$

where $\mathcal{B}$ is defined in (5.1), and $A \cdot B$ denotes the vector dot-product.

It is not common to characterize the secondary structures (helix and sheet) by θ, but θ is strongly correlated with secondary structure, as shown in Figure 5.7, and it gives a simple interpretation. Values $70 \leq \theta \leq 120$ are typical of α-helices, since each subsequent residue turns about 90 degrees in order to achieve a complete (360 degree) turn in four steps (or 72 degrees for five steps, or 120 degrees for three steps). Similarly $140 \leq \theta \leq 180$ is typical of β-sheets, so that the sidechains are parallel but alternate in direction, with one exception. Some β-sheets have occasional 'spacers' in which θ is small, in keeping with the planar nature of sheets.

The distribution of the θ angle peaks roughly at 44, 82 and 167 degrees as shown in Figure 5.7. The peptide bond makes it difficult for θ to be much less than 50 degrees, thus the smaller peak
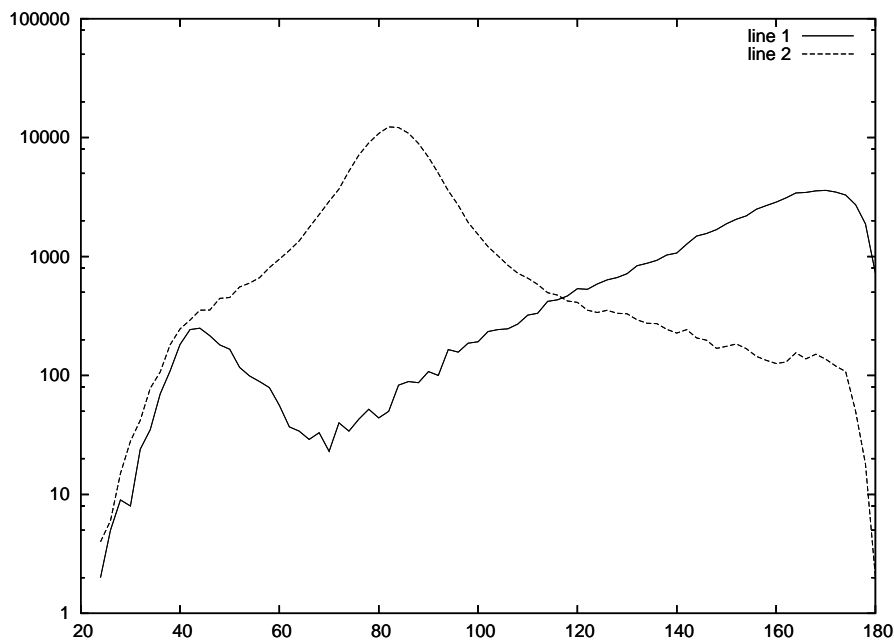
Figure 5.7: The distribution of θ angles in sheets (solid line) and helices (dashed line).

corresponds to a motif where the side chains align as closely as possible. A small number of these occur in beta sheets, but the majority of them constitute an independent motif whose properties deserve further study.

The different structural motifs have characteristic sidechain compositions [16, 266]. For the larger values of θ, hydrophobic residues are found in most pairs; β-sheets have alternating hydrophobic and hydrophilic pairs [266]. By contrast [266], the most common pairs involve predominantly polar or charged residues for $\theta \leq 50$. The ends (or caps) of α-helices necessarily must be different from the middle to terminate the structure [16].

5.2.2 Dihedral angles

We also recall the standard mainchain **dihedral** or **torsion** angles. Given a sequence of four points A, B, C, D in $\mathbb{R}^3$, the **dihedral angle** represents the angle between the planes spanned by subsequent triples of points, namely between the planes spanned by A, B, C and by B, C, D, as depicted in Figure 5.8. For two intersecting planes, there are in general two angles between them. Let μ denote the smaller (positive) angle, in radians; the other angle is $\pi - \mu$ radians. (Of course, both angles could be $\pi/2$.) Choosing which angle to call the dihedral angle is a convention, although it is possible to define it in a consistent way so that it depends continuously on the points A, B, C, D. The dihedral angle can be determined by the orientation of the normal vectors, with the normal determined in a consistent way. In Figure 5.8, the normal vectors are determined using the left-hand rule.

The angle can be defined in terms of the angle between the normal vectors for the two planes.
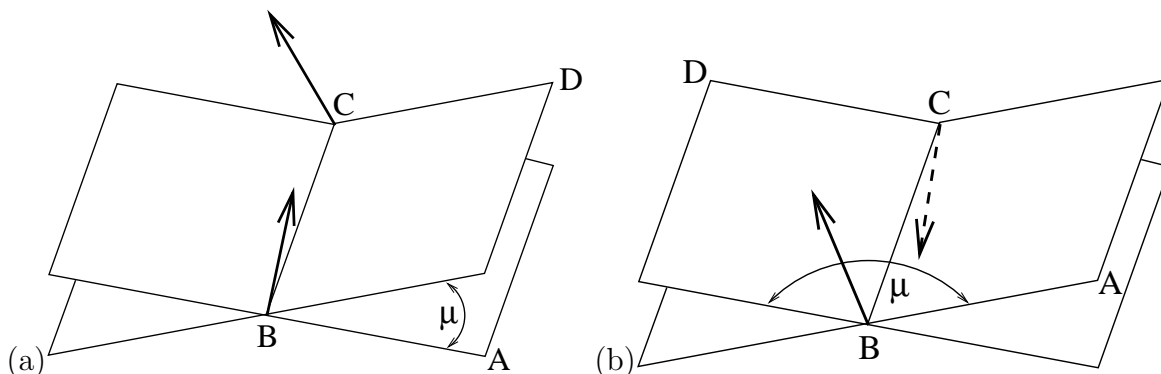
Figure 5.8: The geometry of the dihedral angle $[A, B, C, D]$ in two different configurations: (a) cis form [39], (b) trans form. The normals are defined using the left-hand rule.
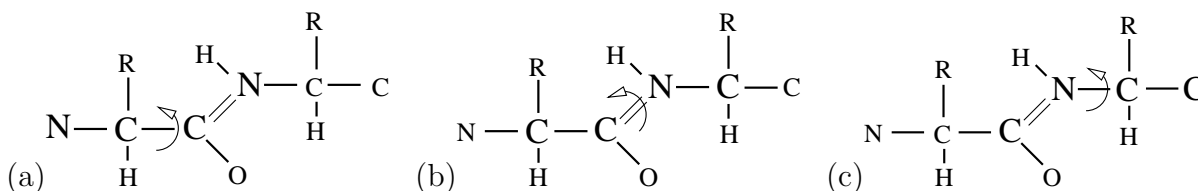
Figure 5.9: The primary dihedral angles: (a) ψ, (b) ω, (c) ϕ. The four points used in each definition are in a larger font. The arrows indicate the axes of rotation.

The normal to the plane spanned by A, B, C is proportional to $n_1 = (A - B) \times (C - B)$, where $\times$ denotes the vector 'cross' product in $\mathbb{R}^3$. Similarly, the normal to the plane spanned by B, C, D is proportional to $n_2 = (B - C) \times (D - C)$. The angle μ between these two normals is given by

$$\cos \mu = \frac{n_1 \cdot n_2}{|n_1| \, |n_2|}, \tag{5.4}$$

where $|n_i|$ denotes the Euclidean length of n_i. Let $[A, B, C, D] = \mu$ denote the dihedral (or torsion) angle between the planes defined by the points A, B, C and B, C, D. The dihedral angle can be viewed as a rotation about the line $\overline{BC}$ at the intersection of the two planes.

Then the ψ, ω and ϕ angles are defined by

$$\begin{aligned}
\psi(x_i) &= [N(x_i), C_\alpha(x_i), C(x_i), N(x_{i+1})] \\
\omega(x_{i+1}) &= [C_\alpha(x_i), C(x_i), N(x_{i+1}), C_\alpha(x_{i+1})] \\
\phi(x_{i+1}) &= [C(x_i), N(x_{i+1}), C_\alpha(x_{i+1}), C(x_{i+1})].
\end{aligned} \tag{5.5}$$

It might appear that there is a possible degeneracy, if three of the consecutive points are collinear. Since this cannot happen for the three angles in (5.5), we do not worry about this case. The reason that the positions of N-C-C are not collinear is the tetrahedral structure depicted in Figure 3.1.

In Chapter 13 we study the effect of a polar environment on the flexibility of peptide bond rotation in the ω angle. The typical trans form of the peptide bond as depicted in Figure 4.2(a) is associated with $\omega = \pi$ radians. The cis form of the peptide bond as depicted in Figure 4.2(b) is associated with $\omega = 0$ radians.
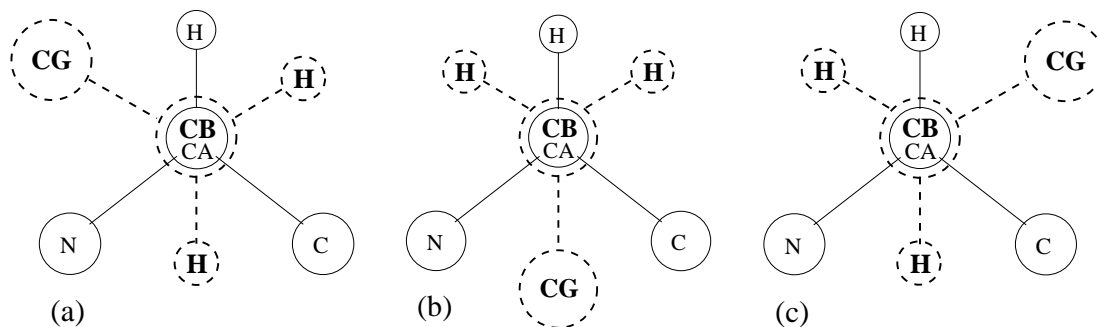
Figure 5.10: The primary sidechain rotamer conformations (a) gauche+, (b) gauche-, and (c) trans, corresponding to χ_1 values of (a) -60 degrees, (b) +60 degrees and (c) 180 degrees. The view is oriented so that the C_α and C_β atoms are aligned perpendicular to the plane of the page. The closer (and therefore larger) atoms are indicated with dashed lines and bold letters. The hydrogens for C_β are indicated. The atom marked 'XG' corresponds to either a C_γ or an O_γ atom.

5.2.3 ϕ, ψ versus ψ, ϕ: the role of θ

The pair of angles ϕ_i, ψ_i captures the rotation of the peptide chain around the i-th C_α carbon atom. The θ angle measures the rotation that corresponds with comparing angles ψ_i, ϕ_{i+1} in successive peptides (cf. Exercise 5.8). This correlation has recently been observed to have significant predictive power [152].

The conformations of ϕ_i, ψ_i are typical of different secondary structures, such as α-helix or β-sheet. The Ramachandran plot [199] depicts the distributions of angles that are commonly adopted (cf. Exercise 5.11).

5.2.4   Sidechain rotamers

The sidechains are not rigid, so the geometric description of a sidechain requires more information than ϕ, ψ and so forth. As we use the ϕ, ψ angles to define the positions of the backbone relative to the position of the N-terminal end, we can also use dihedral angles to define the positions of the sidechains. For example, we can define

$$\chi_1 = [N, C_\alpha, C_\beta, X_\gamma] \tag{5.6}$$

for all side chains that have a C_β atom attached to an additional 'heavy' atom (not Hydrogen). This excludes Glycine and Alanine, but includes all the other sidechains. There may be two candidates for the X_γ atom, but the angle is (or should be, according to standard models) the same for either choice due to the planar structure of terminal groups.

Libraries of angular orientations of the different segments have been developed [272, 338]. The possible orientations are not uniformly distributed in many cases, but rather show a strong bias for a few discrete orientations. For example, carbon chains typically orient so that the hydrogen atoms are in complementary positions. In Figure 5.10, the three primary conformations are shown for side

chains with C_β and C_γ constituents. These conformations are known as **gauche+**, **gauche-**, and **trans**, corresponding to mean χ_1 values of -60 degrees, +60 degrees and 180 degrees, respectively. However, the distributions can change depending on local neighbor context [267].

5.3 Pair frequencies in proteins

We expand upon Section 4.7 which looked at amino acid residue frequencies in different secondary structures. Instead we ask what *pairs* of residues are found frequently in different structural motifs. This requires us to develop some technology to make this question precise. It also introduces some complexities in terms of analyzing PDB files.

5.3.1 Pair frequency definitions

In addition to looking at the frequencies of individual residues, one can also look at the frequencies of pairings. A standard tool for doing this is the **odds ratio**. Suppose that f_i is the frequency of the i-th amino acid in some dataset, and suppose that C_{ij} is the frequency of the pairing of the i-th amino acid with the j-th amino acid. Then the odds ratio O_{ij} is defined as

$$O_{ij} = \frac{C_{ij}}{f_i f_j} \tag{5.7}$$

and has the following simple interpretation. If the pairing of the i-th amino acid with the j-th amino acid were random and uncorrelated, then we would have $C_{ij} = f_i f_j$, and thus $O_{ij} = 1$. Therefore an odds ratio bigger than 1 implies that the pairing is more common than would be expected for a random pairing, and conversely if it is less than 1.

Let us break down the process in more detail. Suppose that the pairs i, j are observed in an amount n_{ij}. Then by the standard definition of frequency (4.2), the pair frequency is

$$C_{ij} = \frac{n_{ij}}{N}, \qquad \text{where} \quad N = \sum_{i,j} n_{ij}. \tag{5.8}$$

If the pairs are not ordered, then the frequency of occurrence of residue i in a pair would be given by

$$f_i = \sum_j C_{ij}, \tag{5.9}$$

by the following reasoning. The quantify

$$o_i = \sum_j n_{ij}, \tag{5.10}$$

is the number of occurences of the i-th residue in any of the pairs. Thus by (4.2),

$$f_i = \frac{o_i}{\sum_j o_j} = \frac{\sum_j n_{ij}}{N} = \sum_j C_{ij}, \tag{5.11}$$
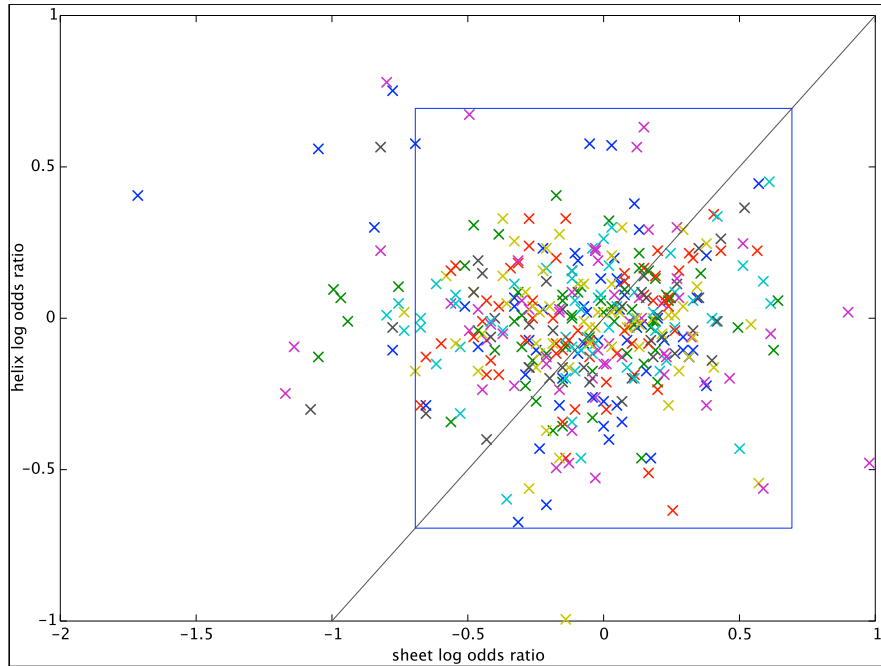
Figure 5.11: Pair frequencies for helices and sheets.

proving (5.9). Therefore

$$O_{ij} = \frac{n_{ij}}{N f_i f_j} = N \frac{n_{ij}}{\sum_k n_{ik} \sum_k n_{jk}}, \tag{5.12}$$

providing a way to compute the odds ration directly from the occurrences n_{ij}.

The **log odds ratio** is often defined by simply taking the logarithm of the odds ratio. This has the benefit of making the more likely pairings positive and the less likely pairings negative. Moreover, since the odds ratio is a multiplicative quantity, the logarithm linearizes it so that the quantities less than 1 are not compressed compared with the ones greater than 1. In [170], a quantity G_{ij} is defined by multiplying the log odds ratio by a numerical factor of 10.

In the case of counting residue pairs in secondary structure in proteins, a further complication occurs: the order of the pair matters. That is, the frequency of the pair where A occurs just before C can be different from the the frequency of the pair where C comes before A. This also means that the simple question of how often A appears becomes more complex. Do we mean how often in appears in a pair AX or in a pair XA? Thus we can define C_{ij} to be the frequency of the pair where the amino acid i is followed by j. Then

$$f_i^b = \sum_j C_{ij} \tag{5.13}$$

is the frequency of occurrence of residue i before another residue in a sequence, and

$$f_j^a = \sum_i C_{ij} \tag{5.14}$$

is the frequency of occurrence of residue i after another residue in a sequence. Thus the correct way to write the odds ratio for the occurrence of i before j is

$$O_{ij} = \frac{C_{ij}}{f_i^b f_j^a} = N \frac{n_{ij}}{\sum_k n_{ik} \sum_k n_{kj}}. \tag{5.15}$$

If the matrix of occurrences n_{ij} were symmetric, then this distinction would not be necessary.

5.3.2 Application to protein sequences

We now apply these ideas to residue pairs in different secondary structure. We saw in Section 4.7 that there are different frequencies for residues in different secondary structure, and we might expect the pair frequencies to have an even stronger differentiating behavior. In Table 4.3, we see that individual frequencies can vary by almost an order of magnitude. However, we will see that pair frequencies are more complicated to interpret.

There are several residue pairs that have extreme probabilities of occurrence in helices versus sheets, as indicated in Table 5.2 and Figure 5.11. The range of odds ratios in Table 5.2 is an order of magnitude, as we found for residue frequencies in Table 4.3. But the extreme odds ratios in Table 5.2 are dominated by pairs with Cys, which is itself very rare. Thus high values of the odds ratio occur in part because of the small denominator.

The box in the middle of the figure segregates pairs for which the odds ratio is either greater than 2 or less than $\frac{1}{2}$. We can take the criterion to define what we mean by highly likely or highly unlikely. The pairs inside the box are not more (or less) likely by a factor of 2 than random, and almost all 400 pairs fall within this box. On the other hand, over 5% of the pairs are unlikely (by a factor greater than 2 over random) to be found in sheets. By contrast, no pair is that unlikely to be found in helices.

One potential application of the residue pair analysis might be to help identify secondary structure by sequence analysis [388]. However, as Figure 5.11 shows, the odds ratios are similar for most of the residue pairs for helices and sheets. There are only two pairs that are both highly likely to be in helices and highly unlikely to be in sheets, according to our factor of 2 standard.

5.3.3 Like-charged pairs

Conventional wisdom would suggest that like-charged residues would not likely be found adjacent in proteins, since they repel each other [271]. Thus it is interesting to look at the occurences of like-charged pairs in sequence. In Table 5.3, the frequencies of such pairs are given for residues that appear in helices. Surprisingly, some pairs actually occur more frequently than chance would predict.

For like-charged pairs, we might have expected odds ratios much less than one, whereas all of the like-charged pairs involving Asp, Glu, Arg, Lys are no less than 0.89, and many are greater than 1. It should be remembered that His is somewhat of a wild card in that its protonation state is strongly dependent on pH in the normal physiological range. Moreover, the highest odds ratio is

helix	sheet	res. pair
2.18	0.45	His Cys
2.12	0.46	Cys Pro
1.50	0.18	Pro Pro
0.37	0.87	Cys Met
0.62	2.66	Lys Cys
1.02	2.46	Cys Lys
1.88	1.16	Cys Cys

Table 5.2: Amino acid pairs with extreme values for odds ratios. Listed are the two highest odds ratio pairs for helices and sheets as well as the lowest odds ratio pairs for both. In addition, the duplex pair of Cys residues is included for reference.

res	Asp	Glu	Arg	His	Lys
Asp	0.97	1.03	0.88	1.06	0.88
Glu	0.96	1.13	1.04	1.07	1.06
Arg	1.12	1.16	1.36	1.06	0.98
His	0.65	1.00	0.63	1.08	0.78
Lys	1.07	1.21	0.89	0.96	1.03

Table 5.3: Odds ratios for charged amino acid sequence-adjacent pairs in helices.

not for a putative salt bridge (residues of opposite sign) but rather for the repulsive, like-charged pair Arg-Arg.

The low odds ratios for oppositely charged sequence-nieghbor pairs is somewhat misleading in that most salt bridges are formed by pairs that differ in sequence by 3 or 4 [104].

5.4   Volume of protein constituents

The number of atoms in amino acid side chains varies significantly. Although atoms do not fill space in the way we would imagine from a terrestrial point of view, it is possible to associate a volume [44] with them that is useful in comprehending them. In 1975, Chothia initiated a study of the size of sidechains and the change in size in the core of proteins, an early use of datamining in the PDB. That study was later revisited [181], and subsequent studies have further refined estimates of sidechain volume, including sizes of individual atom groups [400]. For reference, we have summarized in Tables 5.4–5.6 some estimates [400] of the sizes of the constituent groups inside the core of proteins. The numbers presented represent the dispersion in mean values over the different data sets used [400]. For reference, a water molecule occupies about $30Å^3$ (see Section 17.7.3).

The definition of 'volume' of sidechain or atom-group is important. It might be best to think of this as an 'excluded volume' in the following sense. The approach of [181, 400] uses a Voronoi

decomposition of space based on vertices at the location of the heavy atoms in proteins (excluding hydrogens). We will give the essentials of this approach but describe certain important details only briefly.

The basic Voronoi decomposition involves polytopes defined for each point in the input set S, defined as follows. The polytope P_s for each $s \in S$ is defined as the closure of the set of points in space closer to s than to any other point $t \in S$. Thus for $s \neq t$ in S, the interiors of P_s and P_t will not intersect. What is somewhat magic about the set of polytopes is that they form a decomposition of space (there are no voids). But this is just because, for every point in space, there is a closest point in S. Some of the polytopes are infinite, and these are not useful in the application to proteins. In [400], a modified procedure was used that involves moving faces in the interior, and presumably adding faces at the boundary, based on van der Waahls radii determined separately for each atom. The volume of a residue is defined as the sum of volumes of its atoms. In Figure 5.12, we have depicted only the interior of a Voronoi diagram, excluding the boundary where such infinite domains would occur.

The faces of each polytope are perpendicular to lines between nearby vertices, and are equidistant from the two vertices. Every vertex (heavy atom) is associated with a unique polytope (that contains no other heavy atoms). The heavy atom need not be at the center of the polytope. In Figure 5.12 we show an extreme case in which the points in S can get arbitrarily close to the boundary of the enclosing P_s. In [181, 400], a slight modification is made to the basic Voronoi definition by moving the faces along the line joining the two atoms based on the relative van der Waals radii of the two atoms.

Figure 5.12: Voronoi diagram for a set of points in two dimensions.

At first blush, it might seem that the volumes occupied e.g. by the carbonaceous groups in Table 5.4 are too small, because we know that oil floats on water. But a water molecule is twenty to fifty percent heavier than a CH_n group (for $n = 0, \ldots, 3$), so the size estimates are quite in line with what we would expect. It also might seem that there is a very big variation in volume of different sidechain constituents. However, if we think in terms of the size of a box (or ball) of comparable volume, then the side of the box (or diameter of the ball) does not vary quite so much. For reference, in Table 5.7 we give the relevant volumes for boxes of various sizes, ranging from 2Å to 3.4Å on a side, and we see that this range of linear dimensions accounts fully for the range in volumes observed in Tables 5.4–5.6.

We have omitted sizes for some atom groups from the tables; for example, the volume of most sidechain oxygens is 14.9-19.2Å^3, with the exception of $O(\gamma^1)$ in threonine, whole volume is about 10Å^3. There are some obvious trends, e.g., in Table 5.4 it is clear that the volume of the carbonaeous groups CH_n is an increasing function of n, but the increase in size is greater than linear. On the other hand, although there is a similar trend for the nitrogen-based groups NH_n, the increase is less pronounced; in particular, the volume of the NH_3 group in lysine is only 20.6-21.9Å^3, smaller than the NH_2 groups of argenine.

But there is also a more subtle size issue that is solvent dependent. One of the significant

volume (Å^3)	n	t	carbon groups
9.2—10.1	0	12	D(γ), E(δ), F(γ), H(γ), N(γ), Q(δ), R(ζ), W($\gamma, \delta^2, \epsilon^2$), Y($\gamma, \zeta$)
14.1—14.8	1	3	I(β), L(γ), V(β)
20.0—21.7	2	16	H(δ^2, ϵ^1), F($\delta^{1,2}, \epsilon^{1,2}, \zeta$), W($\delta^1, \epsilon^3, \zeta^{2,3}, \eta^2$), Y($\delta^{1,2}, \epsilon^{1,2}$)
22.5—24.2	2	25	E(γ), I(γ^1), K(γ, δ, ϵ), M(γ), P(δ), Q(γ), R(γ, δ) plus 15 β's
25.2—25.8	2	2	P(β, γ)
35.7—38.5	3	9	A(β), I(γ^2, δ^1), L($\delta^{1,2}$), M(ϵ), T(γ^2), V($\gamma^{1,2}$)

Table 5.4: Volume of carbonaceous atom groups CH_n in protein sidechains [400]. The group sized 22.5—24.2 also includes all C_β carbons with the exception of Ala, Ile, Pro, and Val which are listed in other size groups. t is the number of carbonaceous groups in the category and n is the number of hydrogens in these groups.

volume (Å^3)	n	t	nitrogen groups
14.8—17.0	1	4	H(δ^1, ϵ^2), R(ϵ), W(ϵ^1)
20.6—23.9	2-3	5	R($\eta^{1,2}$), K(ζ), N(δ^2), Q(ϵ^2)

Table 5.5: Volume of nitrogen atom groups NH_n in protein sidechains [400]. t is the number of carbonaceous groups in the category and n is the number of hydrogens in these groups.

conclusions [181] is that hydrophobic residues occupy less volume inside the core of a protein than they do in bulk water. Similarly, hydrophilic residues occupy more volume inside the core of a protein than they do in bulk water. We have listed the change in volume of the various sidechains in Table 4.2. Given our general understanding of the hydrophobic effect, this is not surprising. However, it gives a clear understanding of an important packing effect.

In typical proteins, the increase in volume due to burying hydrophilic residues is compensated by the decrease in volume due to burying hydrophobic residues. That is, the net volume change upon folding is typically quite small. However, for other systems, such a balance does not seem to be so close. For example, cell membranes are made of lipid layers composed substantially of hydrophobic chains. Thus simple pressure tends to keep such cell membranes intact. To break apart, the cell membranes constituents would have to undergo a substantial increase in volume and thus induce a significant increase in pressure.

The volumetric cost of burying hydrophilic residues makes one wonder why they appear inside proteins at all. However, without them the electrostatic landscape of the protein would be far less complex. Moreover, if proteins had only hydrophobic cores, they would be harder to unfold. Both of these effects contribute to an understanding for why charged and polar residues are found in protein cores.

The volumes for the sidechain atoms plus the backbone atoms is about the same in the two papers [181, 400], although the apportionment between backbone and sidechain differs systematically. This would lead to different values for the sidechain volume in Table 4.2, but they would just be shifted by a fixed amount. Thus the relative size change between hydrophobic residues and hydrophilic

atom	group of 17	alanine	glycine	proline
C	8.4—8.9	8.8—8.9	9.5—9.8	8.7—8.8
O	15.7—16.3	16.0—16.3	16.1—16.5	15.8—16.3
N	13.3—14.1	13.8—14.0	14.5—14.9	8.5—8.8
CA	12.9—13.5	14.0—14.1	23.3—23.8	13.8—14.0

Table 5.6: Volume of protein backbone atom groups (Å^3) [400]; 'group of 17' refers to the residues other than Ala, Gly and Pro.

r	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0	3.1	3.2	3.3	3.4
r^3	9.26	10.6	12.2	13.8	15.6	17.6	19.7	22.0	24.4	27.0	29.8	32.8	35.9	39.3

Table 5.7: Relation between volume and length in the range of volumes relevant for proteins.

residues would remain the same.

5.5 Fold networks

We have followed a natural progression in the hierarchy of structure of proteins. A natural next step is to look at interactions between different structural units. To study interactivity from a more global viewpoint, we need some new mathematical technology. A natural representation for interactions is to use graph theory. We will see that appropriate concepts from this theory provide useful ways to compare interactions quantitatively.

Relations among proteins can be determined by various means, and here we look at relations between the basic tertiary structural units of proteins. This will provide both a baseline of what to expect in terms of the graph theory of protein interactions as well as an application of some basic concepts of tertiary structure of proteins. The notion of 'fold' or 'domain' characterizes the basic unit of tertiary structure of proteins, as described in Section 5.1.2. A fold consists of basic units of secondary structures together with relations among them. Secondary structure consists of different types of helices and beta sheets and other motifs, as described in Section 5.1. These stuctural subunits form different groupings (folds) that have characteristic shapes that are seen repeatedly in different proteins.

Although most proteins consist of a single domain, a significant number contain multiple domains [408]. The distribution of multiple domains is known [424] to be exponential. More precisely, the probability $p(k)$ of having k domains was found [424] to be closely approximated by

$$p(k) \approx 0.85 e^{0.41(k-1)}. \tag{5.16}$$

Such a distribution implies [424] "the evolution of multidomain proteins by random combination of domains."

One way to form a relation among proteins is to define the vertices of a graph to be the protein domains and the edges of the graph to be pairs of domains found in a single protein [429]. The resulting graph of protein domain connectivity has many nodes with low degree with just a few highly connected nodes, in agreement with a power-law distribution [342, 429]. More precisely, if we let $p(k)$ be the probability of having k neighbors in this graph, we find

$$p(k) \approx ck^{-\gamma}. \tag{5.17}$$

A network with a distribution (5.17) is often called a **scale free network**i [36]. This provides a useful baseline regarding what to expect in terms of protein interactions. We will see just such a distribution in Figure 11.8.

The definition of domain interaction requires some explanation. If all three-dimensional structures were available in the PDB, this would be simplified significantly. We discuss the use of the PDB in this way further in Section 11.4.2. However, the current fraction of proteins available in the PDB is tiny, and it is expected to remain so for a long time due to the cost and technical obstacles of structure determination. In [429], interaction was determined from different databases devoted to the characterization of protein domains, including both structural means and other techniques such as sequence alignment which do not require structural information. One of the databases used in [429] was Pfam [151]. The objective of "Pfam is a comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile hidden Markov models" [151]. For a description of profile hidden Markov models, see [111].

Knowing the sequence description of a domain allows the determination of proteins with common domains via sequence alignment. If one of the sequences is represented in the PDB, then a fold can be assigned to all the similar sequences. It may seem odd that it would be possible to predict what is essentially structural information (a fold) for a large number of proteins, but one view suggests that relatively few PDB structures are needed [406]. On the other hand, sequence similarity does not always imply fold similarity [174].

The distribution of folds in different species is not uniform. Quite the contrary, the distribution appears to provide a distinct signature for the fourteen different species studied in [424] (twenty species were studied in [342]). Thus it is natuaral to consider the connectivity of folds separately for each species, and the distribution of connectivity of folds for each species was found [342, 429] to follow a power-law with a distinctive decay rate for each species, with "those of smaller genomes displaying a steeper decay" [342], that is, a larger γ in (5.17). We will see a similar type of behavior in Figure 11.8 when we consider the relationship between wrapping and structural connectivity.

We have described two characteristic probability distributions here, the exponential (5.16) and the power-law (5.17). These are both quite different from the more familiar Gaussian (normal) distribution. Since all of these are being fit to a finite number of data points, it is a reasonable question to ask if these distributions are really different enough to distinguish them with just a few points. To address this question, Figure 5.13 presents a comparison of the Gaussian distribution with the power-law distribution (5.17) and the exponential distribution (5.16). On thing revealed by Figure 5.13 is that data that appears to fall along a linear curve in a log-log plot much more closely fits a power-law distribution than it does either a Gaussian or an exponential distribution. Such distributions of data will be seen in Figure 11.8.

5.6   Exercises

Exercise 5.1 *Suppose that we create a string of letters from an alphabet A as follows. We start with an empty string σ. With probability p, we pick a letter $x \in A$ and define a new string $\sigma \leftarrow \sigma + x$ where $+$ means 'append to the string.' With probability $q = 1 - p$, we stop. Then the probability of having a string of length zero is q, and the probability of having a string of length one is pq. Continuing in this way, show that the probability of having a string of length k is $p^k q$. Prove that*

$$\sum_{k=0}^{\infty} p^k q = 1, \tag{5.18}$$

and that $f(k) = p^k q = ce^{-\gamma k}$ for some c and γ. Explain why this supports the statement that the evolution of multidomain proteins can be modeled by the "random combination of domains" [424].

Exercise 5.2 *Use protein constituent volume data [181, 400] (cf. Tables 5.4–5.6) to estimate the density of typical proteins. What is the mass per residue of typical proteins? (Hint: use Table 6.2 to estimate the relative abundance of each residue.)*

Exercise 5.3 *Plot a Gaussian distribution and an exponential distribution as in Figure 5.13 but on a log-linear plot. That is, use a logarithmic scale on the vertical scale but an ordinary (linear) scale on the horizontal axis. Explain how to distinguish these distributions by this device.*

Exercise 5.4 *Determine whether the choice of atom X_γ in (5.6) matters for Asp and Glu by examining high-resolution structures. For each Asp (and Glu), compute the difference of dihedral angles for the two choices of terminal atoms, and plot the distribution of angle distances, in degrees. Compare with Exercise 4.7.*

Exercise 5.5 *Draw figures analogous to Figure 5.10 for the residues Thr and Ser, assuming that they adopt the same basic gauche± and trans configurations. How many possible orientations of the atoms are there? Check to see if the rotomeric states you propose are probable [272].*

Exercise 5.6 *Draw figures analogous to Figure 5.10 for the residues Asp and Asn, assuming that they adopt the same basic gauche± and trans configurations. Include the possible orientations of the terminal pairs (OD1 and XD2) of atoms (where X=O for Asp and X=N for Asn). How many possible orientations of the atoms are there? Check to see if the rotomeric states you propose are probable [272]. (Hint: recall the planarity of the terminal four atoms, Exercise 4.7.)*

Exercise 5.7 *Explain how Figure 5.3(a) represents either a right-handed or left-handed helix depending on whether the dashed line goes above or below the indicated hydrogen bonds, that is, toward or away from the viewer. Recall that a helix is right-handed if it turns clockwise as it moves away from you. (Hint: cut a narrow strip of paper and mark donors and acceptors for the hydrogen bonds at regular intervals. Experiment by twisting the paper into both right-handed and left-handed helices.)*

Exercise 5.8 *In typical peptide bonds, the ω angle is constrained to so that the peptide bond is planar (cf. Figure 13.1). In this case, there is a relationship imposed between θ, ϕ and ψ. Determine what this relationship is.*

Exercise 5.9 *Proteins are* oriented*: there is a C-terminal end and an N-terminal end. Determine whether there is a bias in α-helices in proteins with regard to their* macrodipole μ *which is defined as follows. Suppose that a helix consists of the sequence $p_i, p_{i+1}, \ldots, p_{i+\ell}$ where each p_j denotes an amino-acid sidechain. Let $\mathcal{C}(p)$ denote the charge of the sidechain p, that is, $\mathcal{C}(D) = \mathcal{C}(E) = -1$ and $\mathcal{C}(K) = \mathcal{C}(R) = \mathcal{C}(H) = +1$, with $\mathcal{C}(p) = 0$ for all other p. Define*

$$\mu(p_i, p_{i+1}, \ldots, p_{i_\ell}) = \sum_{j=0}^{\ell} \mathcal{C}(p_{i+j}) \left(j - \tfrac{1}{2}\ell\right) \tag{5.19}$$

Plot the distribution of μ over a set of proteins. Compare with the peptide dipole, which can be modeled as a charge of $+0.5$ at the N-terminus of the helix and a charge of -0.5 at the C-terminus of the helix. How does this differ for left-handed helices versus right-handed helices? What happens if you set $\mathcal{C}(H) = 0$? (Hint: the PDB identifies helical regions of protein sequences. The peptide dipole in our simplification is just ℓ, so μ/ℓ provides a direct comparison.)

Exercise 5.10 *Consider the definition of macrodipole introduced in Exercise 5.9. Explain why the α-helical polypeptide $Glu_{20}Ala_{20}$ would be more stable than $Ala_{20}Glu_{20}$.*

Exercise 5.11 *Determine the Ramachandran plot [199] for a set of proteins. That is, plot the ϕ_i and ψ_i angles for all peptides in the set. Use a different symbol or color for the cases where the i-th peptide is said to be a helix or sheet in the PDB file.*

Exercise 5.12 *From Figure 4.2, we would conclude that the distances between two C-alpha carbons is much smaller when the peptide bond is in the cis form. Compare the C-alpha distances with the ω angle.*

Exercise 5.13 *Correlate the distance between C_α carbons separated in sequence by k with the secondary structure call in a given PDB file. That is, give the (three, separate) distributions of distances as a function of k for helices and sheets and loops (that is, for residues not in either structure).*

Exercise 5.14 *Consider the distribution of distances between C_α carbons separated in sequence by $k = 2$ in a given PDB file. Explain the bimodal character in PDB file 2P9R. Correlate the distance with ϕ and ψ angles.*

Exercise 5.15 *Correlate the distances between CA carbons separated in sequence by k with the secondary structure call in a given PDB file. That is, give the (three, separate) distributions of distances as a function of k for helices and sheets and loops (that is, for residues not in either structure).*
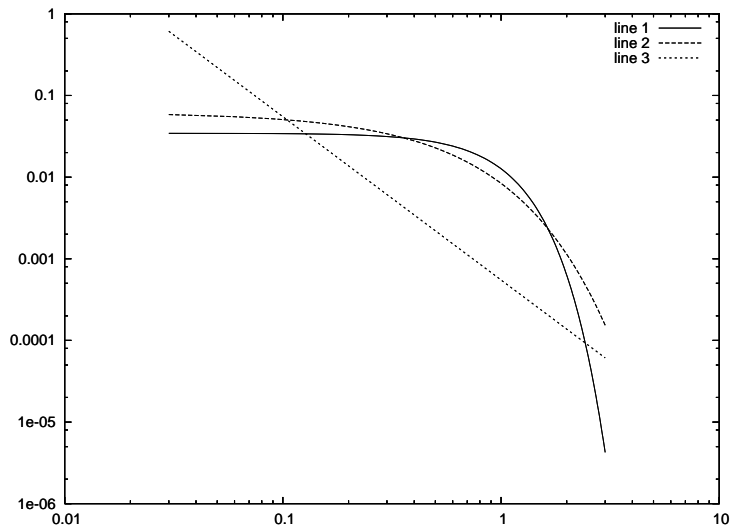
Figure 5.13: Comparisons of the probability distributions $c_1 e^{-x^2}$ (solid line), $c_2 e^{-2x}$ (dashed line), and $c_3 x^{-2}$ (dotted straight line) on the interval $[0.03, 3]$. The constants c_i are chosen to make each distribution integrate to one on $[0.03, 3]$.

Chapter 6

Hydrogen bonds

The concept of the hydrogen bond was established by 1920 [249], and possibly earlier [228, 83]. Hydrogen bonds are the most important bond in biochemistry, so we need to understand them in some depth. Unfortunately, there are several challenges. First of all, although hydrogen bonds in proteins have been considered extensively [30, 394, 242], they are not yet fully understood and are still actively studied [208, 209]. Secondly, in most PDB files, hydrogens are not listed at all, due to the difficulty of locating them by typical imaging techniques. Thus an initial step is to place hydrogens in the appropriate places. We will explain how this is done in simple cases and what the problems are in the difficult cases. We describe how their locations can be inferred starting in Section 6.2. Many different techniques are in use, and comparisons of the different techniques have been made [155].

Hydrogen bonds can be viewed as our simplest example of a switch that is crucial to protein biophysics. Although detailed models do not yet exist to quantify the behavior of a hydrogen bond as a switch, we will see qualitatively that their geometry determines their quality, and that a change in geometry effecitvely can reduce their viability. Their dependence on distance is not a simple one like what we have seen so far for point charges and dipoles. Although hydrogen bonds are often modeled as dipole-dipole interactions, it is known that this model is insufficient. They have a stronger angular dependence and they also involve induction, somewhat like van der Waals interactions.

The general hydrogen bond is of the form XH - - Y where X and Y are 'heavy atoms' such as F, N, O, S or even C in some cases. The X atom is called the **donor** of the bond, and the Y atom is called the **acceptor** of the bond. An example based on the interaction of water molecules is given in Figure 6.1.
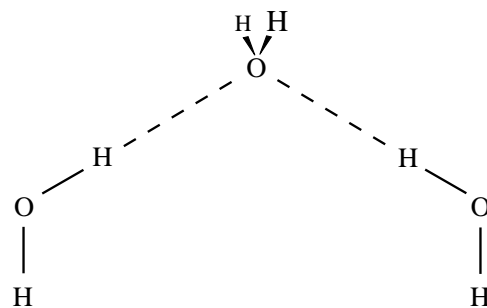
Figure 6.1: Cartoon of the configuration of three water molecules and the hydrogen bonds formed, depicted by the dashed lines. The oxygen atoms accept two hydrogen bonds, whereas the hydrogen atoms are involved in donating only a single bond.

Acceptor	Donor	System	$R(\text{Å})$	$\Delta E(\text{kcal})$
HF	NH_3	$HF–HNH_2$	3.45	1.3
H_2O	NH_3	$H_2O–HNH_2$	3.41	2.3
H_3N	NH_3	$H_3N–HNH_2$	3.49	2.7
HF	H_2O	$HF–HOH$	3.08	3.0
H_2O	H_2O	$H_2O–HOH$	3.00	5.3
H_3N	H_2O	$H_3N–HOH$	3.12	5.8
HF	HF	$HF–HF$	2.72	9.4
H_2O	HF	$H_2O–HF$	2.75	11.7
H_3N	HF	$H_3N–HF$	2.88	4.6

Table 6.1: Distances and energies of various hydrogen bonds. R is the distance (in Ångstroms) between the donor and acceptor (heavy) atoms. The energy ΔE of the hydrogen bond is given in kcal/mole. Note that R "is primarily a function of the degree of positive charge on the hydrogen in the H bond" [233].

6.1 Types of hydrogen bonds

Hydrogen bonds differ based on the heavy atoms that are involved. The variation in bond distance and strength is illustrated in Table 6.1 which has been extracted from [233]. What is clear from this data is that the donor type (the side of the bond that includes the hydrogen) is the primary determinant of the hydrogen bond strength (and length) in these cases. This is interpreted to mean that the charge dipole of the donor is the determining factor [233]. In some sources (including Wikipedia), the electronegativity of the constituents is given as the key factor. But according to [233], "the ability of proton donors and acceptors to form hydrogen bonds (X-H ... Y) is more closely related to their respective acidity or basicity than to the electronegativities of X and Y."

One basic question about hydrogen bonds is whether the hydrogen is in a symmetric position between the donor and acceptor, or whether it favors one side (donor) over the other. The answer is: yes and no [335]. Both situations arise in nature, and there is an intriguing bifurcation between the two configurations, as depicted by the caricature in Figure 6.2. Depicted is a curve that was fit [335] to extensive data on bond lengths of OH - - O hydrogen bonds. The horizontal axis is the distance between the oxygen centers, and the vertical coordinate is the (larger) distance between oxygen and hydrogen. The upper-left segment, where the O-H distance is exactly half of the O-O distance, is the symmetric arrangement. In this configuration, you cannot distinguish one of the oxygen atoms as the donor; both are donor and acceptor.

The dashed parts of the curves indicate where data has been found in both configurations. But what is striking is the void in the O-H distance region between 1.1Å and nearly 1.2Å. Thinking in bifurcation terms, one can stretch the O-O distance in the symmetric configuration, but at a certain point it loses stability and has to jump to the asymmetric one in which the hydrogen has a preferred partner. Moreover, as the O-O distance continues to increase, the (smaller) O-H distance *decreases*, as the influence of the other oxygen decreases with increasing distance. Note that the O-O distance

(for waters) reported in Table 6.1 is 3.0Å, thus clearly in the asymmetric regime (almost off the chart in Figure 6.2).

6.1.1   Hydrogen bonds in proteins

As indicated in Table 6.1, hydrogen bonds vary in character depending on the donor and acceptor. In proteins, there are two classes of donors and acceptors, mainchain (or backbone) and sidechain. All backbone nitrogens (with the exception of proline, unless it is N-terminal) can act as donors of hydrogen bonds, and all backbone oxygens can be acceptors of hydrogen bonds. These donors and acceptors were represented as outgoing and incoming arrows in Figure 5.1. In addition, many of the standard sidechains can act as donors or acceptors, as listed in Table 6.2. Note that certain atoms can be both donors and acceptors, as are the oxygens in water.

Given two classes of contributors, mainchain (M) and sidechain (S), there are four classes of bond pairs: M-M, M-S, S-M, and S-S. We have differentiated between S-M and M-S depending on whether the donor or acceptor is M or S, but in some cases these two classes are lumped into one class.

Given the rigidity of the backbone and the flexibility of the sidechains, it would be reasonable to assume that S-S bonds were the most common and M-M the least. Curiously, it is just the opposite [382]. In Chapter 15, we will see that mainchain-mainchain are much more common. By simple counts in a database of 1547 nonredundant structures, the number of M-M bonds is nearly four times the number of mainchain-sidechain (M-S and S-M) bonds combined, and it is seven times the number of sidechain-sidechain bonds. On the other hand, one finds a significant number of potential sidechain-water hydrogen bonds in many PDB files. These include apparent water bridges [333, 430]. It is not clear how fully waters in PDB files are reported, but their importance to protein structure is significant.

Typical hydrogen donors would make only one hydrogen bond, whereas typical oxygen acceptors can make two hydrogen bonds. However, more complex patterns are possible; see the figures on page 139 of [209]. In some cases, the network of hydrogen bonds can be complicated, as shown in Figure 6.3.

Figure 6.2: Cartoon of the bifurcation of O-H..O hydrogen bonds from a symmetric arrangement to an asymmetric arrangement, based on Figure 4 of [335]. The horizontal axis is the O-O distance and the vertical coordinate is the O-H distance (both in Ångstroms). The upper-left segment is the symmetric arrangement. Note that the water-water hydrogen bond is well away from the symmetric region. The dashed curves indicate where data has been found in both states.

6.1.2   Hydrogen bond strength

Assessing the strength of hydrogen bonds remains a significant challenge [250, 384]. There is a strong angular dependence for the energy of the hydrogen bond [309, 394]. Moreover, the nature

Full name of amino acid	three letter	single letter	Donors (PDB name)	Acceptors (PDB name)
Arginine	Arg	R	NE, NH1, NH2	—
Asparagine	Asn	N	ND2	OD1
Aspartate	Asp	D	—	OD1, OD2
Cysteine	Cys	C	SG*	SG
Glutamine	Gln	Q	NE2	OE1
Glutamate	Glu	E	—	OE1, OE2
Histidine	His	H	ND1, NE2	ND1, NE2
Lysine	Lys	K	NZ	—
Methionine	Met	M	—	SD
Serine	Ser	S	OG	OG
Threonine	Thr	T	OG1	OG1
Tryptophan	Trp	W	NE1	—
Tyrosine	Tyr	Y	OH	OH

Table 6.2: Donors and acceptors for sidechain hydrogen bonds. *If a Cys is involved in a disulfide bridge, it cannot be a hydrogen bond donor.
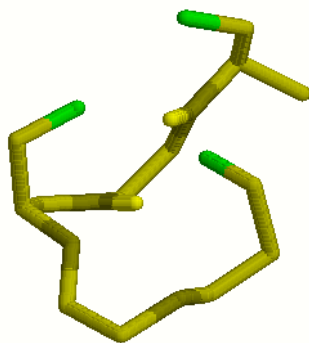
Figure 6.3: Three serines (52, 54 and 56) from chain H the PDB file 1DQM come together to form a complex of hydrogen bonds. The amide groups on the peptide base of H-Gly55 and H-Ser56 provide donors to the oxygens on the end of the sidechain of H-Ser52. The sidechain-sidechain hydrogen bonds among the terminal OH groups on the serines depend on the hydrogen placements.

Full name of non-standard residue or molecule	PDB three letters	Donors	Acceptors
Acetyl group	ACE		O
Glycerol	GOL	O1, O2, O3	O1, O2, O3
Nitrate Ion	NO3		O1, O2, O3
Phosphotyrosine	PTR	N, O2P‡, O3P‡	O, OH, O1P, O2P, O3P
Pyroglutamic acid	PCA	N†	O, OE
Phosphono group	PHS		O1P, O2P, O3P
Phosphate Ion	PO4		O1, O2, O3, O4
Sulphate Ion	SO4		O1, O2, O3, O4

Table 6.3: PDB codes for donor and acceptor atoms in some nonstandard residues and molecules. Key: † Only N-terminus. ‡ In case that the hydrogens PHO2, PHO3 exist in the PDB files.

of the hydrogen bond can depend on the context: even backbone-backbone hydrogen bonds can be different in alpha helices and beta sheets [231, 248, 394].

One might hope that modeling the hydrogen bond as a simple dipole-dipole interaction (Section 10.2.1) would be sufficient to capture the angular dependence. But a purely partial-charge model of hydrogen bonds is not sufficient to capture the angular dependence of the energy: "At the distances where H bonding occurs, the dipole moment approximation is a poor one and higher multipoles must be considered" [233], as we confirm in Section 10.2.1.

Attempts have been made to model hydrogen bonds via more sophisticated interactions. In additon to partial charges, dipole, quadrapole and higher representations of the donor and acceptor groups have been used [58, 59]. A model of this type for the hydrogen bonds in water has been proposed that includes terms for the polarization of water [40, 293]. The difficulty with models of this type is that the multipole expansion converges rapidly only for large separation of the donor and acceptor. Thus these models provide very accurate representations of the asymptotic behavior of the interaction for large separation distances R, but for values of R of close to native separation distances it is only slowly converging. As a remedy to this, distributed multipole expansions, in which the representation involves partial charges (and dipoles, etc.) at many positions, have been proposed [383, 385].

A model to represent the angular and distance dependence of the energy of the hydrogen bond, based only on the atomic distances among the primary constituents, has been proposed [309], in which the dominant term appears to be a strong repulsion term between the like-charged atoms. Such a model is simple to implement because it uses exactly the same data as a dipole model, but with a more complex form and with additional data derived from ab initio quantum chemistry calculations.

More recently, a model based on distance and angular data has been proposed [75] that has been fitted to approximation quantum (DFT) computations and empirical data on hydrogen bonds at interfaces. This model uses a reduced representation based on only one distance and two angles,
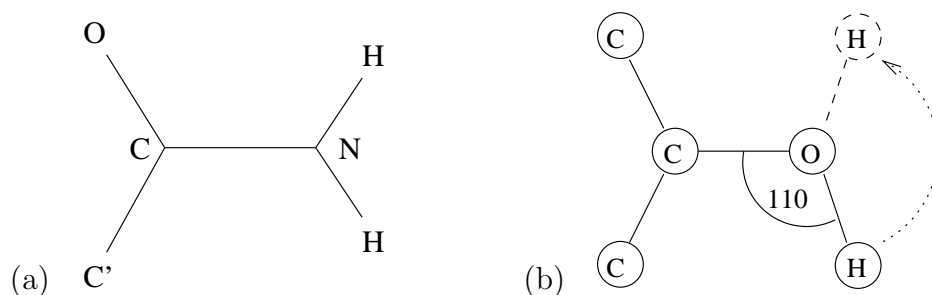
Figure 6.4: Hydrogen placement for three sidechains. (a) Asn and Gln. Shown is the terminal group of atoms for the sidechains. The atom marked C' denotes the preceding carbon in the sidechain, viz., CB for Asn and CG for Gln. (b) Tyr. The hydrogen is in the plane of the aromatic ring, with the angle between C-O and O-H being 110 degrees. Both positions are possible for the terminal hydrogen.

and it is similar to a model used earlier [89]. The earlier model was based on one distance and three angles, with the angle not used in [75] playing only a minor role.

The accurate simulation of one of the simplest hydrogen bonds, in the water dimer, has been of recent interest [227], even though this computation has been carried out for several decades [233]. The fact that this simple interaction is still studied is an indicator of the difficulty of determining information about general hydrogen bonds. Models of the water trimer, tetramer, and hexamer have also recieved recent attention [172, 223, 268].

6.2 Identification of hydrogen positions

Most PDB files do not include locations of hydrogens. Only the heavier atoms are seen accurately in the typical imaging technologies. In general, hydrogen placement is a difficult problem [155, 394]. However, in many cases, the positions of the missing hydrogens can be inferred according to simple rules. We begin with the simplest situations to predict and then consider the more complex ones.

Most hydrogens can be located uniquely. In particular, the Appendix in [299] depicts the locations of such hydrogens, as well as providing precise numerical coordinates for their locations. However, other hydrogen positions are not uniquely determined.

The position of most hydrogens can be modeled by the bond lengths and angles given in [299]. A program called HBPLUS [286] was developed based on this information to provide hydrogen positions in a PDB format. More recently, sophisticated software has emerged to provide estimates of hydrogen positions, including decisions about sidechain ionization (cf. Section 4.5) [155, 167]. The program Reduce is commonly used today [251, 427].

6.2.1 Fixed hydrogen positions

The position of the hydrogen that is attached to the mainchain nitrogen (see Figure 3.1a) can be estimated by a simple formula. The C-O vector and the H-N vector are very nearly parallel, so one can simply take

$$H = N + |C - O|^{-1}(C - O) \tag{6.1}$$

since the N-H distance is approximately one Ångstrom. We leave as an exercise (Exercise 6.1) to make the small correction suggested by the figure on page 282 in [324]; also see the more recent corrections to estimates of bond lengths and angles in [207].

As another simple example, the position of the hydrogens that are attached to the terminal nitrogen in Asn and Gln can also be estimated by a simple formula. The terminal O-C-NH$_2$ group of atoms are all coplanar, and the angles formed by the hydrogens around the nitrogen are all 120 degrees, as depicted in Figure 6.2.1(a). The angle between the C-N and the C-O vectors is very close to 120 degrees [299], so the C-O vector and one of the N-H vectors are very nearly parallel. So one can again take

$$H^1 = N + |C - O|^{-1}(C - O) \tag{6.2}$$

as the location for one of the hydrogens attached to N, since again the N-H distance is approximately one Ångstrom. For the other hydrogen bond, the direction we want is the bisector of the C-O and C-N directions. Thus the second hydrogen position can be defined as

$$H^2 = N + \tfrac{1}{2}\left(|O - C|^{-1}(O - C) + |N - C|^{-1}(N - C)\right) \tag{6.3}$$

We leave as an exercise (Exercise 6.2) to make the small corrections suggested by Figure 13 in [299].

The hydrogen attached to the terminal oxygen in the tyrosine sidechain has two potential positions. The hydrogen must be in the plane of the aromatic ring, but there are two positions that it can take. This is depicted in Figure 6.2.1(b). The one which makes the stronger H-bond with an acceptor is presumably the one that is adopted.

6.2.2 Variable hydrogen positions

The terminal OH groups in serine and threonine are even less determined, in that the hydrogen can be in any position in a circle indicated in Figure 6.5(a). A Cys sidechain that is not engaged in a disulfide bond would be similar, as shown in Figure 6.5(b).

6.2.3 Ambiguous hydrogen positions

An interesting example of the ambiguity of the assignment of the hydrogen location for serines and threonines occurs in the PDB file 1C08. In chain B, Thr30 and Ser28 form a sidechain-sidechain hydrogen bond involving the terminal OH groups, as indicated in Figure 6.6. But which is the donor and which is the acceptor cannot be differentiated by the data in the PDB file in a simple way. Model building shows that both are possible, and indeed there could be a resonance (Section 13.1) between the two states. One state may be forced by the local environment, but without further determining
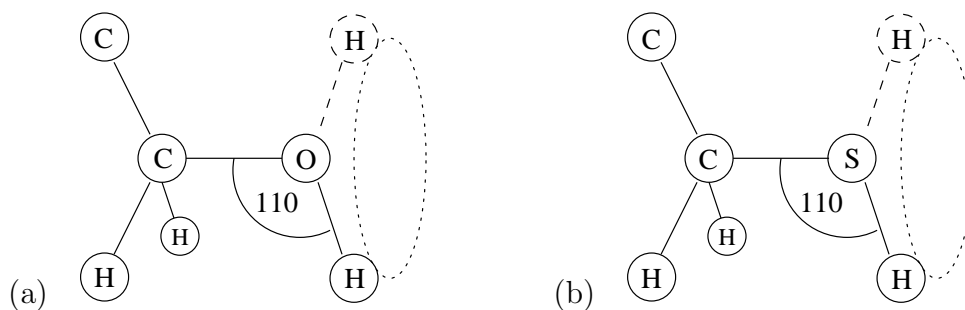
Figure 6.5: Hydrogen placement for (a) Ser and Thr and (b) Cys: anywhere on the dotted circle. The angle between C-O and O-H being 110 degrees, but the hydrogen is otherwise unconstrained.
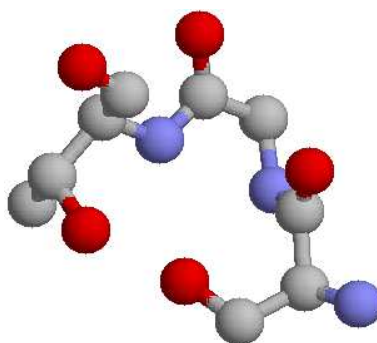
Figure 6.6: Ambiguous hydrogen placement for serine-28 (lower right)—threonine-30 (upper left) sidechain-sidechain hydrogen bond involving the terminal O-H groups; from the B chain in the PDB file 1C08. The sidechain of isoleucine-29 has been omitted but the backbone atoms are shown connecting the two residues. Only the oxygen atoms in the terminal O-H groups are shown.
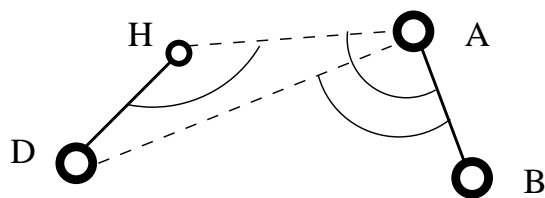
Figure 6.7: Geometric model for hydrogen bonds: D is the donor atom, H the hydrogen, A the acceptor, B acceptor antecedent (i.e. an atom one covalent bond away from the acceptor).

factors both states are possible. It is possible to critique the detailed geometry by considering the quality of the corresponding dipole-dipole interaction (see Section 10.2.1). According to this metric, Thr30 is the preferred donor.

6.3   Geometric criteria for hydrogen bonds

The simplest approach to approximating the angular dependence of the hydrogen bond is to use angular limits, as well as distance limits, in the definition. Each hydrogen bond can be defined by the geometric criteria (Figure 6.7) based on those used in [30], as we now enumerate:

1. Distance between donor and acceptor $|D - A| < 3.5$Å

2. Distance between hydrogen and acceptor $|H - A| < 2.5$Å

3. Angle of donor-hydrogen-acceptor $\angle DHA > 90°$

4. Angle of donor-acceptor-acceptor antecedent $\angle DAB > 90°$

5. Angle of hydrogen-acceptor-acceptor antecedent $\angle HAB > 90°$

To be declared a hydrogen bond, all five criteria must be satisfied.

It is surprising that there is no widely accepted model for hydrogen bonds other than these geometric constraints. There is no reliable quantitative estimate of the strength of hydrogen bonds based on the angle and distance parameters used in the geometric definition even when the context of the bonds is specified (e.g., the hydrogen bonds in particular secondary structures). Note that the five parameters of this model are not independent; only four independent parameters are required.

6.4   Special classes of hydrogen bonds

Hydrogen bonds can occur in a wide variety of situations. Here we gather some cases that might otherwise be obscure.

6.4.1   Salt bridge hydrogen bonds

Electrostatic interactions between postively charged residues, such as Lys and Arg, and negatively charged residues, Asp and Glu, can be characterized ambiguously as both salt bridges and hydrogen bonds in many cases [243, 244]. Also, His can be positively charged in some cases.

The constiutents of a salt bridge have significant energy cost for desolvation. Thus interactants for which none of the heavy atoms were less that 4Å apart were considered destabilizing [243]. A 4Å has been used [244] as a cut-off distance for distinguishing between different salt-bridge interactions. There is some statistical evidence to support this [154].

It is possible to define geometries relevant for salt bridges [104] that can potentially be used to differentiate between true salt bridges and ones that only form hydrogen bonds.

6.4.2   Carboxyl-carboxylate hydrogen bonds

Under suitable conditions, the terminal groups of Asp and Glu can become protonated. The resulting OH group can then form hydrogen bonds with oxygens, including the ones in the terminal groups of other Asp and Glu residues [423]. These are referred to as carboxyl-carboxylate hydrogen bonds. Although these bonds would be expected in low pH environments [359], they have been found to be critical elements of ion channels [301]. In typical PDB structures, the hydrogen in a carboxyl-carboxylate hydrogen bond would not be visible. Thus it could be associated with either oxygen unless further information is available to reveal the association.

6.4.3   Aromatic hydrogen bonds

In general, any negatively charged entity might provide an acceptor for a hydrogen bond. We will see that the faces of the aromatic rings in Phe, Tyr and Trp provide appropriate negative charge regions that act as acceptors of hydrogen bonds. We will postpone the discussion of them until a more detailed analysis of the aromatic sidechains (Chapter 12).

6.4.4   Carbonaceous hydrogen bonds

In some cases, carbonaceous groups (CH) can act as donors for hydrogen bonds [213, 282].

6.5   Exercises

Exercise 6.1 *Refine the formula (6.1) to give a more precise location for the hydrogen attached to the nitrogen in the peptide bond, e.g., following the figure on page 282 in [324] or the more recent corrections to estimates of bond lengths and angles in [207, 397].*

Exercise 6.2 *Refine the formulas (6.2) and (6.3) to give a more precise location for the hydrogens attached to the terminal nitrogen in the residues Asn and Gln, using the data in Figure 13 in [299].*

Exercise 6.3 *Use the model for the energy of a hydrogen bond in [309] to estimate the strength of hydrogen bonds. Apply this to antibody-antigen interfaces to investigate the evolution of the intermolecular hydrogen bonds at the interfaces.*

Exercise 6.4 *Hydrogen positions can be inferred using neutron diffraction data, because hydrogen is a strong neutron scatterer. There are over a hundred PDB files including neutron diffraction data. Use this data to critique the models for hydrogen locations presented in this chapter.*

Exercise 6.5 *Helical secondary structure is formed by amide-carbonyl hydrogen bonding between peptides i and j where $3 \leq |i - j| \leq 5$. Determine how frequent it is to have $i - j = k$ for the different possible values of $k = -5, -4, -3, 3, 4, 5$. Are there instances where amide-carbonyl hydrogen bonding between peptides i and j where $|i - j| = 2$ or $|i - j| = 6$?*

Exercise 6.6 *The C-O (carbonyl) groups in the peptide backbone can make two hydrogen bonds (typically), whereas the N-H (amide) group usually forms only one hydrogen bond. How common is it for carbonyl groups to make two bonds in helical secondary structures? In β-sheet structures? How often are the bonds mainchain-mainchain bonds, versus sidechain-mainchain bonds? As a first step, you can define helical carbonyls to be ones where there is bonding between peptides i and j where $3 \leq |i - j| \leq 5$, but determine how many double bonds there are for each value of $k = i - j$.*

Exercise 6.7 *Determine the angular dependence of the mainchain-mainchain hydrogen bond. What is the distribution of O–H distances and C-O, N-H angles? Consider the different classes of bonds separately: those in (1) parallel and (2) anti-parallel sheets, and those in helices of separation $k = \pm 3, \pm 4, \pm 5$ (cases 3-8). How does the bond distance and angle correlate? What is the mean distance and angle in each case?*

Exercise 6.8 *The pairing of Gln (i) and Asp (i+4) in alpha helices can make a sidechain-sidechain hydrogen bond in addition to the mainchain-mainchain hydrogen bond[202]. What is the odds ratio of this pairing? How favorable is the geometry for this sidechain-sidechain hydrogen bond (determine the distribution of angles and distances for these bonds).*

Exercise 6.9 *The pairing of Gln (i) and Asp (i+4) in alpha helices can make a sidechain-sidechain hydrogen bond in addition to the mainchain-mainchain hydrogen bond[202]. What other sidechain pairs can lead to sidechain-sidechain hydrogen bonds in alpha helices or beta sheets?*

Chapter 7

Determinants of protein-protein interfaces

We now turn to a key question: what factors are most influential in protein-ligand binding? We review attempts to answer this question both to give a sense of the historical development and also to emphasize key aspects of the datamining techniques used. Later in the book we will clarify the role of dehydrons in this process, but for now we proceed naïvely to get a sense of how the ideas developed. Although we ultimately want to consider general ligands, we only work with ones that are themselves proteins in this chapter.

Protein associations are at the core of biological processes, and their physical basis, often attributed to favorable pairwise interactions, has been an active topic of research [47, 78, 160, 215, 270, 298, 348, 412]. A common belief has been that hydrophobic interactions play a central role in protein-ligand binding. According to [399], the "hydrophobic effect [...] is believed to play a dominant role in protein-protein interactions." The guiding principle of [170] was that in the past "interfaces were found to have more hydrophobic residues than the rest of the protein surface. Recent large-scale studies of protein complexes have confirmed the importance of hydrophobicity in proteinprotein interactions." However, more precise comments could also be found. According to [430], "The prevailing view holds that the hydrophobic effect has a dominant role in stabilizing protein structures." Whether or not hydrophobic effects are important for protein-ligand binding, one might also expect that the sort of bonds that help proteins form their basic structure would also be involved in joining two different proteins together.

7.1 Direct bonds

Both hydrogen bonds and salt bridges play a significant role at protein interfaces. The mean density of hydrogen bonds between two different proteins at an interface is reported to be $0.474/\nu m^2$ [430], about one per 2.11 square nanometers, corresponding to a grid of size 1.45Å on a side. For reference, the interfacial area per atom reported in [71] has a mean of 9.25Å^2; this corresponds to 10.81 atoms per νm^2, or about one atom per grid block with a grid size of about 3.04Å. Grids of these respective sizes are indicated in Figure 7.1; a typical patch would consist of a square consisting of about 13

grid steps on each side. (The mean patch size is 1558Å^2; data are based on the 46 single-patch interfaces listed in Table I of [71].)

The average number of hydrogen bonds per interface studied in [430] is about ten. Thus there is a small discrepancy between the data in [430] and [71], as is typical for different studies of interfaces. On the other hand, the average number of salt bridges per interface is only two [430]. Disulfide bonds play a more limited and specialized role.

It might be that the story of protein-protein interactions ends here, with the intermolecular hydrogen bonds and salt bridges being the whole story. However, three of the 54 high-resolution structures studied in [430] have no hydrogen bonds or salt bridges, and another dozen have no salt bridges and five or fewer hydrogen bonds. Not surprisingly, we will begin to see indications of the role of intramolecular hydrogen bonds that become enhanced upon binding, as we depicted in Figure 2.8.

One factor that complicates the picture of protein-protein interactions is the appearance of water molecules which play a structural role, as opposed to simply mediating interactions via dielectric effects. In the protein interfaces studied in [430], polar atom pairs bridged by water across the interface with hydrogen bonds were potentially more numerous than direct hydrogen bond pairs, with each water molecule connecting 3.8 cross-chain atom pairs on average. This unusually high number is not fully explained, but the prospect was suggested that "water molecules are most likely to allocate their hydrogen bonds to the protein atoms dynamically, in the so-called flip-flop mechanism" [430]. If there are 3 potential donors and 2 potential acceptors, there could be as many as 6 different donor-acceptor pairs, if the geometry permits. An additional piece of information would be provided by the number of distinct donors and acceptors potentially interacting with waters.
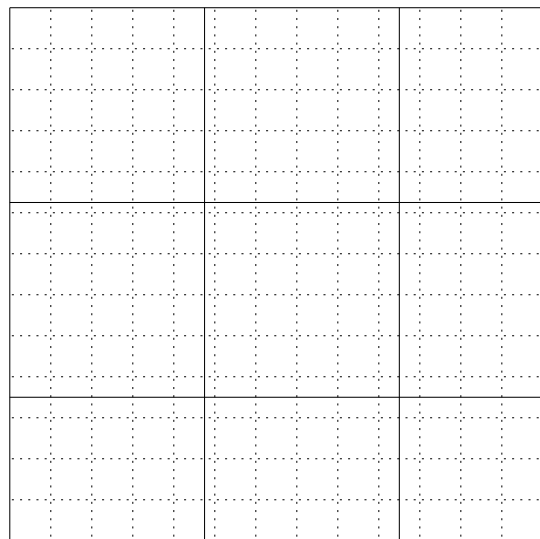
Figure 7.1: Meshes indicating the average density of atoms and hydrogen bonds at typical protein-protein interfaces, based on the data from [71, 430]. The finer, dotted mesh indicates the atom density (one per square) and the coarser, solid mesh indicates the hydrogen bond density (one per two squares).

7.2 Hydrophobic interactions

There have been several attempts to define a hydrophobicity scale for protein sidechains as a guide to protein-ligand binding [44, 112, 246, 291, 296, 314, 339]. A similar, but distinct, concept is that of **lipophilicity** [283] which measures the extent to which substances dissolve in a non-polar solvent. Although the scales are designated as hydrophobicity measures, they are really intended to be proxies for the local dielectric environment [291]. One characteristic feature of most hydrophobicity measures is that the scale attempts to balance hydrophobicity with hydrophilicity, in such a way
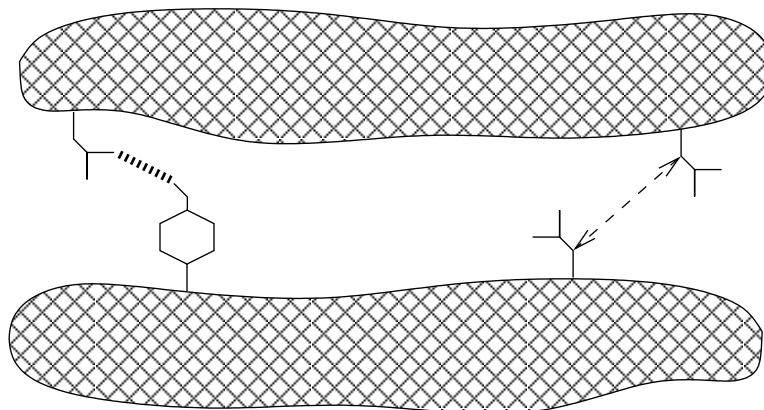
Figure 7.2: Cartoon showing a protein-protein interface and some possible ways to characterize interactions between residues. On the right, two residues (Asp and Asn fit the description, but Gln, Glu or Val are similar) have their C_β carbons close. On the left, a hydrogen bond between Tyr and Gln is depicted, but with the C_β carbons of each at a greater distance.

that amphiphilic residues tend to be in the middle of the scale. However, a hydrophilic residue does not cancel the hydrophobic effect in a simple way, at least regarding its impact on the local dielectric. A hydrophilic residue surrounded by hydrophobic groups will not have a strong effect on the dielectric environment. As we will see in Chapter 16, it is both the abundance and the mobility of water molecules that contributes to the dielectric effect. A small number of (confined) water molecules hydrogen bonded to a singular polar group on a protein sidechain will not cause a significant increase in the local dielectric. We will present here a scale for sidechains based on datamining protein interfaces that turns out to correlate closely with the amount of wrapping.

We review one attempt [170] to discover hydrophobic interactions by examining protein-protein interfaces. The hypothesized form of the interactions in such studies determined the basic choices that guided the data mining. In particular, a definition of 'sidechain interaction' using only the proximity of C_β carbons was used [170] to quantify interactivity; it does not take into account individual sidechain features. This is depicted in Figure 7.2 in which two sidechains on the right have close C_β carbons, but two on the left are not so close even though they form a hydrogen bond. Such a definition is appropriate for postulated hydrophobic interactions which are generally nonspecific, but it is not designed to detect more subtle relationships. By a re-examination of the data, the studies actually provide confirmation that hydrophobic-hydrophobic interactions are not prominent in the interfaces studied.

Indeed, hydrophobic-polar pairings at protein-protein interfaces are frequent and challenge the commonly held view regarding hydrophobic interactions. The prediction and explanation of binding sites for soluble proteins require that we quantify pairwise energy contributions. But also we must concurrently explain the extent to which surrounding water is immobilized or excluded from the interactive residue pairs. As proteins associate, their local solvent environments become modified in ways that can dramatically affect the intramolecular energy [21, 120, 128, 132, 148, 306, 412].

Water removal from hydrophobic patches on protein surfaces has a high thermodynamic benefit
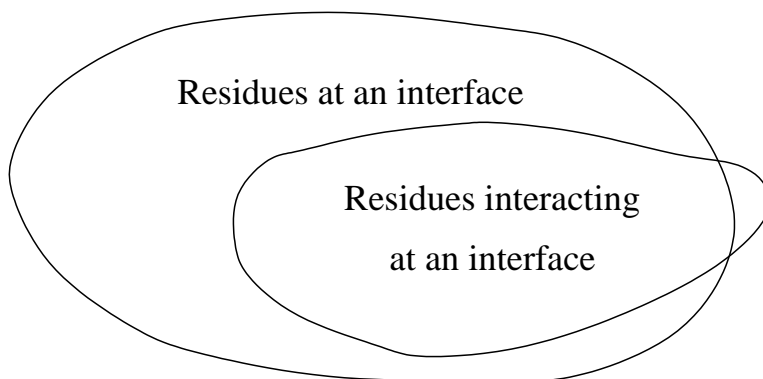
Figure 7.3: Cartoon showing possible relationship between two datasets.

[47, 78, 160, 215, 270, 298, 348, 412], due to an entropic gain by the solvent. The water next to hydrophobic patches lacks interaction partners (hydrogen bond partners), and in moving to a bulk environment it gains hydrogen bonds. Thus, hydrophobic patches are possible binding regions provided there is a geometric match on the binding partner. However, most protein surfaces have ratios of hydrophilic to hydrophobic residues reportedly ranging from 7:1 to 10:1 [142]. Moreover, hydrophobic patches involved in associations at an interface are often paired with polar groups [378]. We will ultimately explain how this can be energetically favorable, but we begin with some small steps to understand better the nature of protein interfaces.

7.3   Amino acids at protein-protein interfaces

We begin with a simple use of datamining applied to the understanding of amino acid tendencies at interfaces. There are different questions that one can ask, and of course it is natural that amino acids get ranked in different orders accordingly. For simplicity, we contrast just two that have appeared in the litterature, but we also review others in Section 7.5.3. The data here is drawn primarily from [53, 147, 170]. The two basic questions we address are the following.

- What residues are most likely to be *found* at an interface?

- What residues are most likely to be *interacting* at an interface?

It is important to realize how these questions differ, and how they drive the resulting data mining experiments.

Unfortunately, there is no universal way to define a protein-protein interface. The basic idea is that it is the region of the proteins in which there are two sidechains, one from each protein, having a specified relationship. For example, two such sidechains are said to be 'interacting' in [170] if the distance between their C_β carbons (C_α for Gly) was less than 6 Ångstroms. A shortcoming of this definition is indicated in Figure 7.2, in which two sidechains making a hydrogen bond across the interface will not be counted as interacting, whereas two others may be counted even though there is no obvious type of interaction.

3-letter code	1-letter code	Nonpolar Carbons	Interface Rel. Prop.	Dehydron Rel. Prop.	Hydro-pathy	Levitt s value
Asn	N	1	+1.28	+1.63	-3.5	0.2
Thr	T	1	+1.10	+1.41	-0.7	-0.4
Gly	G	0	+0.99	+1.42	-0.4	0.0
Ser	S	0	+0.60	+0.80	-0.8	0.3
Asp	D	1	+0.34	+0.76	-3.5	2.5
Ala	A	1	+0.29	+0.60	1.8	-0.5
Cys	C	0	+0.25	+0.24	2.5	-1.0
Val	V	3	+0.20	-0.31	4.2	-1.5
Met	M	1	+0.10	+0.10	1.9	-1.3
Tyr	Y	6	+0.10	+0.10	-1.3	-2.3
His	H	1	-0.25	-0.25	-3.2	-0.5
Pro	P	2	-0.25	-0.25	-1.6	-1.4
Trp	W	7	-0.33	-0.40	-0.9	-3.4
Arg	R	2	-0.35	-0.40	-4.5	3.0
Leu	L	4	-0.35	-1.10	3.8	-1.8
Phe	F	7	-0.40	-0.40	2.8	-2.5
Lys	K	3	-0.42	-0.38	-3.9	3.0
Glu	E	2	-0.50	-0.11	-3.5	2.5
Gln	Q	2	-0.62	-0.60	-3.5	0.2
Ile	I	4	-0.70	-0.92	4.5	-1.8

Table 7.1: Amino acids ranked according to their likelihood of being found at protein-protein interfaces. The second column indicates the number of nonpolar carbon groups in the side chain (see Table 8.2). Interface and dehydron relative propensity (Rel. Prop.) is given as $R_i - 5$ as in (7.5). Dehydron Propensity is also presented as frequency $f - 5$; 5% is the average propensity to be at an interface or engaged in a dehydron. The hydropathy scale of Kyte et al. [246] and the hydrophobicity values s of Levitt [257] are included for reference.

On the other hand, the JAIL database [177, 409] defines an interface by requiring two atoms from sidechains from separate parts of protein structures to be within 4.5Å, in which case the two entire residues are considered to be interacting. This definition further requires at least 5 C_α atoms to be within the interaction zone [409]. Furthermore, distinctions are made between **obligate interactions**, between two fold domains within a single protein, and transient (nonobligate) interactions, between different proteins [280, 409].

Another approach that has been used to define an interface uses the concept of **solvent-accessible surface**. This has a mathematically rigorous definition, although it does not describe accurately the solvation of a protein. It should really be called a **sphere-accessible surface**. To define this surface, we first define a domain that could be called the van der Waals domain of a protein:

$$W = \cup_j \left\{ x \mid \text{distance}(x, a_j) \le r_j \right\}, \tag{7.1}$$

where r_j is the van der Waals radius of atom j in the protein, which is located at the point a_j. We then define a set of points which are centers of (open) spheres (of radius $\lambda > 0$) contained in the exterior of W:

$$X_\lambda = \left\{ x \mid \text{distance}(x, W) > \lambda \right\}. \tag{7.2}$$

The solvent-accessible surface S_λ is the boundary of X_λ:

$$S_\lambda = \partial X_\lambda. \tag{7.3}$$

The physical justification of this definition is not that water is roughly spherical, but rather that the locations of the points in a water molecule, for arbitrary possible orientations of a water molecule, should be represented by a spherical volume. Thus we can say that (with the appropriate λ), X_λ is a set of points where we can be sure that a single water molecule could be accommodated (without steric conflict). The set of points where water could actually go is potentially larger, since it is not a spherical object. On the other hand, if we considered the hydrophobicity of a patch of protein, it might be that water would not be very likely to get as close as S_λ might suggest. So S_λ gives only an approximation to an actual solvation surface, and the errors in prediction could be on either side of the surface.

The contact area for a protein-protein interface may be estimated by comparing the solvent-accessible (or other estimated solvation) surface area of the two proteins separately with that of the joined unit. In [71], the sidechains contributing to this discrepancy in area are counted as being in the interface. Another approach to identify interfaces uses threading techniques originally used to predict protein folding [273]. Since there is little agreement in how interfaces are defined, we will not attempt to give the details in each case. Differences in conclusion could be largely affected by differences in definitions.

The site specificity of protein-protein interactions has been widely studied due to its central biological significance [95, 170, 200, 215, 216, 217]. Hydrophobic residues such as Leu and Val are more abundant at protein-ligand interfaces. As a result, the role of hydrophobic residues in the removal of water surrounding the protein surface has been assumed to be a dominant factor for association [140, 376]. But it is also true that such residues are more abundant in proteins in general (see Table 7.2).

The first question [147] we consider is about the amino acid composition of protein–protein interfaces. This can be done by simply counting, once an identification has been made regarding which amino acids are at an interface. However, simple frequencies are misleading: Leu is the most common residue at interfaces, but it is also overwhelmingly the most common residue in most proteins. Thus one has to normalize by the natural frequencies of amino acids in proteins [53].

The second question [170] is about the amino acid composition for pairs of amino acids at interfaces *that are interacting.* There are many ways to define interaction, but proximity [170] is a natural metric. Thus, two residues are defined [170] as interacting if their C_β coordinates differ by at most 6Å (with C_α used for Gly). This notion is simplistic in that the C_β atom is only the first in the sequence. The key feature of this definition of interactivity is that it does not discriminate *how* sidechains might be interacting. Moreover, it might be that two sidechains are forming a bond and yet their C_β atoms are further apart than 6Å.

In view of Section 7.1, it would be reasonable to pre-filter potential interactions by known interactions, such as hydrogen bonds, salt bridges, and so forth (e.g., cation-π interactions and dehydrons) [79]. But this approach was not taken in [170].

Let us compare and contrast the two questions. The first question seeks to determine clues for protein-protein association by investigating all residues, suitably normalized. The second question assumes that proximity of sidechain pairs is a significant factor in protein-protein association, and thus looks for consequences of restricting to such pairs. Not surprisingly, each question returns different answers regarding the relative significance of different residues. In Figure 7.3, we depict the difference between the two data sets. We allow for the fact that being 'at the interface' may be differently defined in each case, leading to the possibility that neither set contains the other.

The distribution of amino acid composition in proteins displays evolutionary trends [53], and this can require extra care to reveal subtle relationships. Here we limit our investigations to fairly strong trends for simplicity. However, the precise numerical data presented would differ if different databases were chosen for the primary data being used.

7.4 Interface propensity

A common belief is that hydrophobic residues on the surface of proteins are likely candidates to support interfaces in protein-protein association. In Section 7.5, we present evidence that supports this case with suitable clarifications. However, [147] presents data with a distinctively different conclusion, by focusing on all residues found at an interface and normalizing the relative abundance of residues at the interface by their over-all abundances. The residues with the highest relative propensity [147] to be at interfaces are, in decreasing order of frequency, Asn, Thr, Gly, Ser, Asp, Ala, and Cys, the group depicted in Figure 4.3(a). None of these residues is distinctively hydrophobic. This is quite a surprising result, and it demands an explanation.

If we have two datasets with the same characteristics, with frequencies f_i and g_i, respectively, as defined in (4.2), then one can define a **relative frequency**

$$r_i = f_i/g_i \tag{7.4}$$

of the characteristics between the two datasets. There are some problems with this measure of occurrence. First of all, it might happen that $g_k = 0$ for some k, making the interpretation difficult. Related to this is the need for normalization in order to be able to relate two different comparisons. In [147], the following approach was taken.

Define a normalized **relative propensity** via

$$R_i = \frac{r_i}{\sum_{j=1}^{N} r_j}. \tag{7.5}$$

These relative propensities sum to one, so we can think of them like ordinary frequencies. Similarly, we multiply by 100 in (4.2) to convert to percentages as the unit of "frequency."

If we apply this approach to datasets of proteins, and the characteristics are the different amino acid constituents, then we obtain the scheme used in [147]. In this case, the sum of the relative propensities (in percentage units) is one hundred, so the mean is five. In Table 7.1, data from [147] is presented in terms of the deviation of these relative propensities from the mean of five. That is, the data represent $100R_i - 5$.

The unusual ranking of residues in Table 7.1 was explained in [147] by noting that it correlates closely with the propensity to be engaged in under-wrapped backbone hydrogen bonds, among amino acids acting as either proton donors or acceptors for mainchain hydrogen bonds. These data are presented in the fifth column in Table 7.1, and the correlation is striking. Such bonds, in turn, are determinants of protein-protein associations, as discussed subsequently.

Since we expect a significant number of intermolecular hydrogen bonds (and some salt bridges) at interfaces, we might expect residues capable of making them (cf. Table 6.2) to be more likely at interfaces. But these residues are uniformly distributed in Table 7.1, not clustered near the top. If anything, the charged residues are clustered near the bottom. This implies that another factor determines the propensity to be at an interface, as suggested in [147], namely, the amount of wrapping a residue can provide.

As noted in [147], the seven residues in Figure 4.3(a), with the highest propensity for being engaged in under-desolvated hydrogen bonds, also have at most one torsional degree of freedom in their side chain. Thus, the entropic loss resulting from the conformational hindrance of the sidechains upon protein association is minimal with these sidechains, so that the energetic benefit of intermolecular protection of pre-formed hydrogen bonds is most beneficial. The only purely hydrophobic residue that has an appreciable propensity to be in an interface is Val (cf. Figure 4.3(b)), with only one sidechain rotameric degree of freedom. Therefore, its conformational hindrance upon binding also entails minimal loss in conformational entropy.

Consideration of the residues ranked at the bottom of Table 7.1 demonstrates that hydrophobic residues on the protein surface are infrequent relative to their over-all abundance. This implies that they are negatively selected to be part of binding regions, and thus they must play a secondary role in terms of binding.

Note that the polar residues (Asn, Asp, Ser, Cys and Thr) with a minimal distance from their polar groups to the backbone are likely to be engaged in dehydrons, according to Table 7.1. It is presumed [147] that this arises not only because they have minimal nonpolar carbonaceous groups, but also because the relative proximity of their polar groups to a backbone hydrogen bond may limit

Res. Code	Pairing Rel. Prop.	Pairing Rel. Freq.	Pairing Freq. [170]	Total Abundance [53]	Interface Rel. Prop.	Rim/Core freq. [71]
Cys	5.4	2.40	1.87	0.78	+0.25	0.45
Trp	1.9	1.60	1.63	1.02	-0.33	0.32
Pro	1.7	1.55	6.74	4.35	-0.25	1.24
Ser	1.5	1.50	7.01	4.66	+0.60	1.04
Asn	1.3	1.46	4.90	3.36	+1.28	1.19
Thr	1.1	1.41	6.87	4.87	+1.10	1.19
His	0.76	1.33	2.56	1.92	-0.25	0.52
Tyr	0.32	1.23	3.70	3.00	+0.10	0.67
Gly	0.11	1.18	8.59	7.30	+0.99	1.16
Ala	0.11	1.18	9.18	7.77	+0.29	0.95
Phe	-0.15	1.12	4.02	3.61	-0.40	0.33
Gln	-0.33	1.08	3.41	3.15	-0.62	1.03
Met	-0.72	0.99	2.38	2.41	+0.10	0.54
Asp	-0.98	0.93	5.06	5.42	+0.34	1.48
Val	-1.2	0.87	7.12	8.17	+0.20	1.09
Leu	-1.6	0.79	7.05	8.91	-0.35	0.82
Ile	-1.8	0.75	5.00	6.66	-0.70	0.76
Arg	-1.9	0.71	4.46	6.27	-0.35	1.19
Glu	-2.6	0.55	4.71	8.59	-0.50	1.87
Lys	-2.9	0.48	3.73	7.76	-3.9	2.16

Table 7.2: Amino acids which occur in pairs at interfaces and their relative abundances as determined in [170]. Primary data is taken from the indicated references. Relative Propensity is defined in (7.5) and Relative Frequency is defined in (7.4). Interface Relative Propensity from Table 7.1 is included for comparison. The rim versus core frequencies ratio indicates the ratio of frequencies of residues found in the rim versus core of a protein interface [71], as depicted in Figure 7.4.

further clustering of hydrophobic groups around the bond. Gly is itself the greatest under-wrapper and can even be thought of as polar due to the fact that the polar environment of the peptide bond is exposed; Ala is the penultimate under-wrapper and may also exhibit some of the polar qualities of Gly (cf. Section 4.4.1).

7.5   Amino acid pairs at interfaces

We now return to the second question raised at the beginning of the chapter regarding the amino acid composition for interacting pairs of amino acids at interfaces. We recall the warm-up exercise regarding the pair frequencies of amino acid residues in secondary structure presented in Section 5.3. This gives a baseline for what to expect regarding pair frequencies in general. We review the results in [170] which attempts to explain the common belief that hydrophobic interactions should dominate. We show in fact that a re-examination of the the results in [170] indicates something contrary.

7.5.1   Pair interactions at interfaces

We review the results in [170] which use proximity as an interaction metric in which two residues are defined as interacting if their C_β coordinates differ by at most 6Å. We utilize the basic technology regarding pair frequencies developed in Section 5.3. In [170], a quantity G_{ij} is defined by multiplying the log odds ratio by a numerical factor of ten. We recall that the hypothesis of [170] was that hydrophobic interactions would be the most prevalent. In this setting, some dominant residue pairs are indeed hydrophobic, although it is pointed out in [170] that they "occurred more often in large contact surfaces, while polar residues prevailed in small surfaces," anticipating the subsequent discussion regarding "core" versus "rim" residues [71]. We present in Table 7.2 the residues and their relative propensities, as defined in (7.5), in decreasing order, as determined in [170].

Two of the residues in Table 7.2 with greatest relative propensity, namely Trp and Pro, are distinctively hydrophobic, as expected in the basic hypothesis of [170]. However, these are also two of the most unique residues, as discussed in Section 4.4. Trp-Pro can be involved in a what is called a 'sandwich' [343], so this is indeed a result in line with the basic hypothesis. But there are other sidechains which are purely hydrophobic, and have few other interesting features, such as Val, Leu, Ile and Phe. If the hypothesis of [170] were correct, these sidechains would be dominant among the leading pairs. Instead, other high-ranking residue pairs in Table 7.2 involve residues ranked at the top in Table 7.1. Indeed, one might wonder why there would be any differences in the two tables. The differences between Table 7.2 and Table 7.1 reflect the fact that we are now asking about residues which are in proximity of a specific residue.

Since Table 7.2 does not provide relative abundances directly, we need to say how these have been derived. The fundamental data in Table 7.2 is Table II on page 93 in [170], which lists the "contact" matrix C_{ij}. This is a matrix that counts the number of times that residue i contacts (is within the proximity radius of) residue j. Summing a column (or row) of C_{ij} and normalizing appropriately gives the total frequency F_i of the i-th amino acid involved in such pairings. More

Res. Pair	log odds ratio	odds ratio
Cys-Cys	0.626	1.87
Trp-Pro	0.351	1.42
Asp-His	0.220	1.25
Arg-Trp	0.205	1.23
Asp-Ser	0.202	1.22
Asp-Thr	0.191	1.21
Cys-Ser	0.181	1.20
Asp-Gln	0.174	1.19
Met-Met	0.145	1.16
Cys-His	0.136	1.15

Table 7.3: Highest log-odds ratios reported in Table III in [170]. Note that the numbers listed in that table are the log-odds ratios inflated by a factor of $A = 10$.

precisely, to report frequencies as a percentage, define

$$F_i = 100 \frac{\sum_{j=1}^{20} C_{ij}}{\sum_{i,j=1}^{20} C_{ij}} \tag{7.6}$$

to be the amino acid pairing frequency, shown in the column entitled 'Pairing Freq. [170]' in Table 7.2.

The abundance of each amino acid in such pairings needs to be normalized by an appropriate measure. Here we have taken for simplicity the abundances published in [53] which are reproduced in the column entitled 'Total Abundance [53]' in Table 7.2. We do not claim that this provides the optimal reference to measure relative abundance in this setting, but it certainly is a plausible data set to use. The data shown in the column entitled 'Pairing Rel. Freq.' in Table 7.2 represents the ratio of F_i, defined in (7.6), to the abundances reported in [53].

The fact that Cys appears to have the highest relative abundance in pairs at interfaces reflects the simple fact that when Cys appears paired with another residue, it is unusually frequently paired with another Cys to form a disulfide bond (Section 4.2.2), as confirmed in [170].

It is noteworthy that the odds ratios indicated in Table III of [170], the largest of which we have reproduced in Table 7.3, are all between one half and two. That is, there are no pairs which occur even as much as twice as frequently as would be expected randomly (or half as frequently). The pair in [170] with the highest odds ratio (1.87) is Cys-Cys, a disulfide bridge. Although Cys is uncommon, when it does appear we can expect it to be involved in a disulfide bridge. The next highest odds ratio pair is Trp-Pro (1.42), which pairs two of the most unique sidechains (Sections 4.4.2 and 4.4.5). The lack of rotational freedom in proline may be significant since there is no entropic loss in the pairing, but the story is likely more complex, e.g., Trp-Pro can be involved in a sandwich [343].

The subsequent four pairs in [170] with the next highest odds ratios involve charged residues: Asp-His (1.25), Arg-Trp (1.23), Asp-Ser (1.22) and Asp-Thr (1.21). The first of these pairs is a salt-bridge, and the second is a charge-polar interaction known as a cation-π interaction [88, 164, 431]

(see Section 3.1.4 and Section 12.2) based on the special polarity of aromatic residues (Section 4.4.5). The latter two pairs are charged and polar residues as well, and their interactions could well be based on hydrogen bonds. The next four pairs in ranking of odds ratio are Cys-Ser (1.20), Asp-Gln (1.19), Met-Met (1.16) and Cys-His (1.15).

These pairings show a similar mix of polar interactions, not the expected hydrophobic-hydrophobic interactions. In particular, why are Phe-Phe, or Phe-Leu, or the other purely hydrophobic residue pairs not found as commonly at interfaces? In [170], the typical configuration of Arg-Trp is pictured, a likely cation-π interaction, and similar polar pairings are highlighted, such as Lys-Lys, with an odds ratio 0.81, which is a fairly high odds ratio for pairs which repel each other.

There is no absolute scale on which to measure odds ratios, and the significance of any deviation from one is context dependent. But it is notable that the pair frequencies reported in [170] are much smaller than found for alpha helices or beta sheets as indicated in Figure 5.11. We interpret that to mean that the hydrophobic pairs involved in interfaces are more nearly random, none of which occur with very high odds ratios. This does not mean that the approach of [170] was flawed, rather that the hypothesis was proved to be incorrect. If the dominant interactions had been hydrophobic, these techniques would have discovered them. On the other hand, it could be possible to pre-filter the data to eliminate identifiable interactions, and this might provide interesting new data.

7.5.2   Core versus rim

When we add the further analysis in [71] which differentiated the prevalence of core versus rim residues in protein interfaces, the picture is clarified. In [71], interface topology was characterized in detail, and it was found that interfaces could typically be described in terms of discrete patches of about 1600 Å^2 in area. For each patch, the boundary (rim) residues were identified versus the interior (core) residues, each of which constituted about half of the total.

Figure 7.4: Cartoon showing a protein-protein interface subdivided into rim versus core regions.

The statistics for amino acid preferences for the rim versus the core are reproduced in Table 7.2. There is a strong correlation between being charged or polar and preferring the rim, as indicated in Table 7.4.

Similarly, it is noteworthy that the variance in relative propensities is much greater for pairs of interacting residues at interfaces (Table 7.2) than it is for all (unrestricted) residues at interfaces (Table 7.1). This is not surprising because we have selected for a particular subset of pairs (instead of including all pairs). Combining the previous two observations, we can say that interacting pairs at the core of interfaces are more likely to involve a hydrophobic residue, but the pair compositions involving hydrophobes are nearly random.

To illustrate the sensitivity of results depending on the database chosen, we review the results in [26] which is very similar in spirit to [71], the difference being the use of homodimers for the study of interfaces. In Table 7.4, we present this data, with the residues reordered to give the rim/core

Res. Code	Rel. Prop.	Rel. Freq.	Pair Freq.[170]	Total Abundance [53]	Rim/Core freq. [71]	Homodimer Rim/Core [26]
Lys	-2.9	0.48	3.73	7.76	2.16	2.19
Glu	-2.6	0.55	4.71	8.59	1.87	1.48
Asp	-0.98	0.93	5.06	5.42	1.48	1.61
Pro	1.7	1.55	6.74	4.35	1.24	1.51
Asn	1.3	1.46	4.90	3.36	1.19	1.49
Thr	1.1	1.41	6.87	4.87	1.19	1.16
Gly	0.11	1.18	8.59	7.30	1.16	1.38
Arg	-1.9	0.71	4.46	6.27	1.19	0.85
Val	-1.2	0.87	7.12	8.17	1.09	0.83
Ser	1.5	1.50	7.01	4.66	1.04	1.15
Gln	-0.33	1.08	3.41	3.15	1.03	1.22
Ala	0.11	1.18	9.18	7.77	0.95	0.93
Leu	-1.6	0.79	7.05	8.91	0.82	0.61
Ile	-1.8	0.75	5.00	6.66	0.76	0.55
Tyr	0.32	1.23	3.70	3.00	0.67	0.58
Met	-0.72	0.99	2.38	2.41	0.54	0.68
His	0.76	1.33	2.56	1.92	0.52	0.85
Cys	5.4	2.40	1.87	0.78	0.45	0.81
Phe	-0.15	1.12	4.02	3.61	0.33	0.40
Trp	1.9	1.60	1.63	1.02	0.32	0.60

Table 7.4: Amino acids which occur in pairs at interfaces and their relative abundances. Primary data is taken from the indicated references.

preferences in order for the data in [71] to facilitate comparison with the data in [26]. What we see is the same general trend, namely that charged and polar residues prefer the rim, but with changes in the particular rankings among the different groups. However, there is a significant reversal in the roles of arginine and valine [26].

The dissection trilogy is completed in [27] in which an attempt is made to determine aminoacid distributions for 'nonspecific' interactions. This is intended to be a proxy for any surfaces which might bind however briefly to other protein surfaces. The dataset is determined by looking at crystal contact surfaces in the PDB. We leave as an exercise to compare the data for these surfaces with the other data presented here. See [27] for a comparison with the data in [71] and [26].

7.5.3   Further studies

We have made several observations based on analyzing existing data sets. These conclusions should be viewed as preliminary since these data sets must be viewed as incomplete. Our primary intent was to introduce a methodology for exploring such data sets and to indicate the type of results that can be obtained.

Our basic analysis of pairwise interaction data was taken from [170]. However, the methodology is quite similar to that of the earlier paper [399], although there are differences in the way the interior (and non-interior) sidechains in the interaction zone are defined. That is, the classification of rim and core residues in the interface [170] is different in definition from exposed and interior residues in the interface in [399], although similar in spirit. Figure 3B of [399] shows how the residues that are interacting (proximate) in an interface are very similar in composition to ones in the interior of proteins.

Protein-ligand interfaces differ in function, and interfaces with different function can have different composition. In [215], basic differences between protein-antibody and enzyme-inhibitor pairs, as well as others, are explored. Using more extensive datasets available more recently, this approach has been refined to allow classification of interface type based on aminoacid composition [315].

The aromatic sidechains do play a special role in protein interfaces through what is called a cation-π interaction [164] (see Section 12.2). The special polar nature of the aromatic residues (Section 4.4.5) provides the opportunity for interaction with positively charged (cation) residues (Lys, Arg, His), however other types of bonds can be formed as well, including a type of hydrogen bond [258]. The cation-π motifs play a special role in protein interfaces [88, 431]. The cation-π interaction also has a significant role in α-helix stabilization [369].

A study of the role of evolution on protein interface composition can be found in [65]. In [179, 275], interacting amino acids across interfaces are studied and compared with regard to conservation and hot spots.

Protein-protein interactions can be classified in different ways, e.g., by how transient they are, and studies have been done to examine differences in size of interaction zones and sidechain propensities [311, 312].

Identification of individual sidechains that may play the role of 'anchors' in protein-ligand recognition is studied in [344] via molecular dynamics simulations. Individual residues are identified that appear to fit into geometric features on paired protein surfaces both in crystal structures and in the

dynamic simulations.

It is possible to refine the concept of sidechain interactions to one involving the interactions of individual atoms in structures. This approach has been suggested [82] as a way to discriminate between correct structures and incorrect ones. In [82], this concept was proposed as a way to critique structures being determined based on experimental imaging techniques, but the same concept could be applied to discriminate between native and **decoy structures** that are proposed via computational techniques.

7.6 Hot spots

In [47], an attempt is made to identify so-called **hot spots** on protein surfaces. They report on the results of an experimental technique called **alanine scanning** in which residues are replaced by alanine and compared with the original protein by some activity assay. A hot spot is defined to be a residue X for which the change in energy due to the mutation X→Ala is large.

What they discover is that the most common sidechains at hot spots are the ones that are bulkiest: Trp, Tyr and Arg. This is not surprising since the replacement by Ala has the greatest geometric, electrostatic, and hydrophobic effects for these residues. However, such substitutions might be extremely rare in nature.

What might be a better test of importance would be other mutations, e.g., ones which do not change the volume or geometry of the side chain. Systematic replacement of all amino acids by all other amino acids is clearly an order of magnitude more work than just replacing by a fixed side chain. Having a better model of what governs protein-protein interactions could lead to a more directed study of sidechain mutation effects. A correlation has been found between experiments and the level of residue conservation among related interfaces [275].

7.7 Conclusions about interfaces

Two main conclusions were obtained. The first is that residue hydrophobicity is not the primary attribute that determines proximity of a residue to interaction sites. Instead, there is a different 'interactivity' order that governs the likelihood of an amino acid residue being in an active zone. This interactivity scale is related strongly to the number of nonpolar constituents of sidechains, which governs the local dielectric environment. Thus the likelihood of a residue being at an interface is to some extent *inversely* proportional to its hydrophobicity.

On the other hand, pairwise interactions with hydrophobic residues may play a secondary role in protein-protein interactions, especially in the interior, or core, regions of interaction domains. Moreover, their interactions tend to be less specific than might be the case in other pairings, such as in alpha helices and beta sheets. The role of hydrophobic sidechains in such interactions is not revealed by such an analysis. In particular, the definition of 'interaction' has been taken to be simple proximity, so it is misleading to infer that there is any identified form of interaction.

Interfaces can be studied using the PRISM system [316].

7.8   Exercises

Exercise 7.1 *Compare the data for the surfaces in [26, 27, 71] by constructing a table analogous to Table 7.4.*

Exercise 7.2 *The aminoacid frequencies for different datasets constitute probability distributions on the set of aminoacids. Different datasets have different distributions. In [27], the distributions for nonspecific interaction surfaces are compared with the distributions for other surfaces [26, 71]. The comparison metric is the L^2 norm. Consider the effect of using the KL-divergence, Jensen-Shannon metric, and the earth-moving metric.*

Exercise 7.3 *The frequency of location at interfaces provides a linear ranking (Table 7.1) of residues that can be useful in making predictions based on techniques from learning theory. As an example, consider using this to identify under-wrapped hydrogen bonds in α-helices directly from sequence data. For an α-helix, there will be hydrogen bonds formed between residues at a distance of 3, 4, or 5 residues. Generate data from a protein sequence by computing the product of the product of interface ranks of two neighbors. That is, for a sequence abcd define $x = rank(a)rank(b)$ and $y = rank(c)rank(d)$. Thus for every four letter sequence, we assign a pair of numbers (x, y) in the unit square. If there is a dehydron associated with abcd then we expect (x, y) near zero. Using data from the PDB, construct a support-vector machine to separate dehydrons from wrapped hydrogen bonds. Then use this machine to predict dehydrons in sequences for which the sequence is not known.*

Exercise 7.4 *Compute the distribution of distances from water molecules (HOH) in PDB files to the nearest (a) CH_n, (b) NH, (c) OH, and (d) CO sidechain groups. Give the distances from the HOH coordinates to the nearest heavy atom (oxygen in the case of CO). Omit distances greater than 6 Å.*

Exercise 7.5 *Compute the distribution of distances from water molecules (HOH) in PDB files to the nearest (a) NH and (b) CO mainchain groups. Give the distances from the HOH coordinates to the nearest heavy atom (oxygen in the case of CO). Omit distances greater than 6 Å.*

Exercise 7.6 *Re-do the analysis in the chapter using the abundance data in Table 4.3 instead of [53].*

Exercise 7.7 *How frequently does Cys appear in the PDB bonded to another Cys versus nonbonded to a Cys? Use a simple criterion of proximity of the terminal heavy atoms as a proxy for the formation of a disulfide bridge. Scan a large data set of PDB files and determine the relative frequencies of the two cases. Plot the distribution of distances between the terminal heavy atoms of the nearest Cys residues. Also plot the joint distribution of distances and the CSSC dihedral angle.*

Exercise 7.8 *Define a distance function for sidechains in some way, e.g., by computing the distance between the centroids [154]. Let $\Delta r > 0$ be fixed and define frequencies as follows for a given protein*

database. Let $f_{i,j,k}$ denote the frequency of occurences of sidechain pairs i and j whose distance r is in the interval $(k-1)\Delta r \leq r < k\Delta r$. Let

$$f_k = \sum_{i,j} f_{i,j,k}$$

be the frequency of occurence of any sidechain pair with this distance, and let

$$f_{i,j} = \sum_k f_{i,j,k}$$

be the frequency of occurence of the pair i,j at any distance. Then the quotient

$$r_{i,j,k} := \frac{f_{i,j,k}}{f_k f_{i,j}}$$

measures the likelihood of finding a pair i,j at distance $k\Delta r$. Plotting $r_{i,j,k}$ as a function of $r = k\Delta r$ suggests the distances at which the two residues i and j are most likely to be located. This can be given an energetic interpreation [154].

Chapter 8

Wrapping electrostatic bonds

We begin by reviewing the history of research regarding the hydrophobic effect and its effect on protein structure and stability. A key concept is the identification of a minimal unit of hydrophobicity: a single nonpolar carbonaceous group CH_n, $n = 0, 1, 2, 3$, a common constituent of protein sidechains. We then review the concepts leading to characterization of nonpolar versus polar groups in molecules, the key idea being electronegativity of atoms. Following these preliminaries, we explain ways that the hydrophobic effect has been defined in terms of wrapping by nonpolar carbonaceous groups. We then see that the genetic code can be understood as minimizing change in polarity and charge of residues upon single-letter genetic mutations while faciltating exploration of wrapping space.

8.1 Role of nonpolar groups

By 1959, the role of hydrophobicity in protein chemistry was firmly established [221]. Soon afterward [159, 229], the role of hydrophobicity in enhancing the stability and strength of hydrogen bonds in proteins was demonstrated. However, the story developed slowly, and a careful interpretation is required. The paper [229] studied a model molecule, N-methylacetamiden, that is similar to the peptide backbone in structure and forms the same kind of amide-carbonyl (NH–OC) hydrogen bond formed by the backbone of proteins. Infrared absorption measurements were performed to assess the strength and stability of the hydrogen bonds formed by N-methylacetamiden in various solvents (including water) with different degrees of polarity. The paper's main conclusion might be misinterpreted as saying that hydrogen bonds are not significant for proteins in water [229]: "It seems unlikely, therefore, that interpeptide hydrogen bonds contribute significantly to the stabilization of macromolecular configuration in aqueous solution." However, the authors did confirm the opposite view in less polar solvents, so we would now say that their study indicated the value of hydrophobic protection of hydrogen bonds in proteins.

The subsequent paper [159] also studied model molecules, including N-methylacetamiden, in solvents based on varying ratios of trans-dichloroethylene and cis-dichloroethylene, via infrared spectroscopy. They established that "the free energy and enthalpy of association of the amides can be expressed as a function of the reciprocal of the dielectric constant." Although the variation in

dielectric constants achieved with these solvents only reached a level of one-tenth that of water, this paper quantified the effect of dielectric modulation on the strength and stability of hydrogen bonds in systems similar to proteins. Thus it remained only to connect the variation in the dielectric constant to quantifiable variations in protein composition.

Although the energetic role of peptide hydrogen bonds remains a subject of significant interest [31, 32], it now seems clear that the variation in hydrophobicity in proteins has a significant and quantifiable effect on the behavior of proteins [98]. According to [430], "The prevailing view holds that the hydrophobic effect has a dominant role in stabilizing protein structures." The quantitative use of hydrophobicity as a marker for hot spots (Section 7.6) in proteins is reported by diverse groups [57, 139].

Attempts to quantify the hydrophobicity of different sidechains has a long history [257]. The role of hydrophobic residues in strengthening hydrogen bonds has been studied by many techniques. The concept we call wrapping here is very similar to what has been termed **blocking** [28] and **shielding** [166, 274]. We prefer the term wrapping since it evokes the image of providing a protective layer around a charged environment. The term 'shielding' has a related meaning in electronics, but it is also easy to confuse with 'screening' which for us is what the water dielectric performs. The material used for shielding in a coaxial cable is a type of cylindrical screen, and it is a conductor, not an insulator.

In an experimental study [28] of hydrogen exchange [29], the authors stated that (hydrophobic) "amino acid side chains can enhance peptide group hydrogen bond strength in protein structures by obstructing the competing hydrogen bond to solvent in the unfolded state. Available data indicate that the steric blocking effect contributes an average of 0.5 kJ per residue to protein hydrogen bond strength and accounts for the intrinsic beta-sheet propensities of the amino acids." Although this result is clearly quantitative, it should be understood that the experimental technique is indirect. Hydrogen exchange [29] refers to the exchange of hydrogen for deuterium in a highly deuterated environment, and it most directly measures the lack of hydrogen bonds.

Numerical simulations of peptides also contributed to the growth in understanding of the quantitative effect of hydrophobic groups on hydrogen bonds. Based on computational simulations [404], the authors stated that their results provided "a sound basis with which to discuss the nature of the interactions, such as hydrophobicity, charge-charge interaction, and solvent polarization effects, that stabilize right-handed alpha-helical conformations."

One might ask what minimal quantum of wrapping might be identifiable as affecting the strength or stability of a hydrogen bond. The work on hydrogen exchange [28, 29] shows differences in the effect on hydrogen bonds for various hydrophobic sidechains (Ala, Val, Leu, Ile) which differ only in the number of carbonaceous groups. More recent experiments [274] have looked directly at the propensity to form alpha-helical structures of polypeptides (13 residues) which consisted of X=Gly, Ala, Val, Leu, or Ile flanked on either side by four alanine residues with additional terminal residues (Ac-KAAAAXAAAAKGY-NH2). These experiments directly measured the strength and stability of hydrogen bonds in these small proteins. The experimental evidence [274] again shows differences between the different sidechains X in terms of their ability to increase helix propensity, and hence their effect on the hydrogen bonds supporting helix formation. This observation was further developed in a series of papers [18, 19, 20, 21]. More recent, and more complex, experiments

[165] confirm that hydrogen bond strength is enhanced by a nonpolar environment.

8.1.1 Unit of hydrophobicity

Based on the accumulated evidence described previously, we take a *single carbonaceous group to be an identifiable unit of hydrophobicity.* There is perhaps a smaller, or another, unit of interest, but at least this gives us a basis for quantification of the modulation of the dielectric effect. It is perhaps surprising that such a small unit could have a measurable effect on hydrophocity, but we already remarked in Chapter 1 on comparable effects of a single carbonaceous group regarding toxicity of alchohols and antifreezes.

It is possible that removal of water can be promoted by components of sidechains other than purely carbonaceous ones. For example, we noted that the arginine residue does not solvate well [284], in addition to the fact that it contains significant carbonaceous groups. A computational study [166] of a 21-residue peptide including a triple (tandem) repeat of the sidechains AAARA concluded that "the Arg side chain partially shields the carbonyl oxygen of the fourth amino acid upstream from the Arg. The favorable positively charged guanidinium ion interaction with the carbonyl oxygen atom also stabilizes the shielded conformation." Note that the second sentence indicates a possible sidechain-mainchain hydrogen bond.

Since wrapping is of interest because of its implications for hydrophobicity, one could attempt to model hydrophobicity directly as a scalar quantity. Such an approach using a sidechain-based (cf. Section 8.3) evaluation has been taken [56, 57] based on estimates of hydrophobicity provided earlier [257] (see the values in Table 7.1). We have defined wrapping as an integer quantity defined for each bond, but this could (by interpolation) be extended as a function defined everywhere, and the use of a cut-off function [56, 57, 257] essentially does that. But the scalar quantity of real interest with regard to electrostatic bonds is the dielectric, which is described in Section 8.5.

8.1.2 Defining hydrophobicity

There have been several attempts to define a hydrophobicity scale for protein sidechains as a guide to protein-ligand binding [112, 246, 291, 339]. The numbers for two of these are listed in Table 7.1. Also included in Table 7.1 is a scale for sidechains based on datamining protein interfaces that turns out to correlate closely with the amount of wrapping [147].

Although the scales are designated as hydrophobicity measures, they are really intended to be proxies for the local dielectric environment [291]. One characteristic feature of most hydrophobicity measures is that the scale attempts to balance hydrophobicity with hydrophilicity, in such a way that amphiphilic residues tend to be in the middle of the scale. However, a hydrophilic residue does not cancel the hydrophobic effect in a simple way, at least regarding its impact on the local dielectric. A hydrophilic residue surrounded by hydrophobic groups will not have a strong effect on the dielectric environment. It is both the abundance and the mobility of water molecules that contributes to the dielectric effect. A small number of (confined) water molecules hydrogen bonded to a singular polar group on a protein sidechain will not cause a significant increase in the local dielectric.

For a protein structure to persist in water, its electrostatic bonds must be shielded from water attack [28, 140, 148, 274, 331, 404]. This can be achieved through wrapping by nonpolar groups (such as CH_n, $n = 0, 1, 2, 3$) in the vicinity of electrostatic bonds to exclude surrounding water [140]. Such desolvation enhances the electrostatic energy contribution and stabilizes backbone hydrogen bonds [33]. Any amide and carbonyl partners in backbone hydrogen bonds can become separated temporarily due to thermal fluctuations or other movements of a protein. If such groups remain well wrapped, they are protected from being hydrated and more easily return to the bonded state [91], as depicted in Figure 2.4.

The thermodynamic benefit associated with water removal from pre-formed structure makes under-wrapped proteins adhesive [128, 141, 143]. As shown in [140], under-wrapped hydrogen bonds (UWHB's) are determinants of protein associations. In Section 9.1, we describe the average adhesive force exerted by an under-wrapped hydrogen bond on a test hydrophobe.

The dielectric environment of a chemical bond can be modified in different ways, but wrapping is a common factor. There are different ways to quantify wrapping. Here we explore two that involve simple counting. One way of assessing a local environment around a hydrogen bond involves just counting the number of 'hydrophobic' residues in the vicinity of a hydrogen bond. This approach is limited for two reasons.

The first difficulty of a 'residue-based' approach relates to the taxonomy of residues being used. The concept of 'hydrophobic residue' appears to be ambiguous for several residues. In some taxonomies, Arg, Lys, Gln, and Glu are listed as hydrophilic. However, we will see that they contribute substantially to a hydrophobic environment. On the other hand, Gly, Ala, Ser, Thr, Cys and others are often listed variously as hydrophobic or hydrophilic or amphiphilic. We have identified these five residues in Chapter 4 as among the most likely to be neighbors of underwrapped hydrogen bonds, as discussed at more length in Chapter 7. As noted in Section 4.4.1, glycine, and to a lesser extent alanine, can be viewed as polar, and hence hydrophilic, but alanine has only a nonpolar group in its sidechain and thus would often be viewed as hydrophobic.

A second weakness of the residue-counting method is that it is based solely on the residue level and does not account for more subtle, 'sub-residue' features. We will see that these limitations can be overcome to a certain extent with the right taxonomy of residues. However, we will also consider (Section 8.4) a measure of wrapping that looks into the sub-residue structure by counting all neighboring non-polar groups.

The residue-counting method is included both for historical and pedagogical reasons, although we would not recommend using it in general. It provides an example of how models are developed over time, with refinements made once better understanding is available. If an effect is important, then simple models ought to be able to represent it, even if later more refined models are preferred. Thus we consider in detail some of the predictions from the residue-counting method to see how well they are holding up.

In the residue-counting measure of wrapping, we define precisely two classes of residues relevant to wrapping. This avoids potential confusion caused by using taxonomies of residues based on standard concepts. One could think of this dichotomy as defining hydrophilic versus hydrophobic residues, but that is not intended. In Section 8.3.2, we show that this definition is sufficient to give some insight into protein aggregation and to make predictions about protein behaviors.

atomic symbol	H	C	N	O	F	Na	Mg	P	S	Se
electronegativity	2.59	2.75	3.19	3.66	4.0	0.56	1.32	2.52	2.96	2.55
nuclear charge	1	6	7	8	9	11	12	15	16	34
outer electrons	1	4	5	6	7	1	2	5	6	6
missing electrons	1	4	3	2	1	7	6	5	2	2

Table 8.1: Electronegativity scale [249, 325, 356] of principal atoms in biology. The 'outer electrons' row lists the number of electrons needed to complete the outer shell.

However, it is also possible to provide a more refined measure that looks below the level of the residue abstraction and instead counts all non-polar groups, independent of what type of sidechain they inhabit. We present this more detailed approach in Section 8.4. We will show in Section 9.1 that there is a measurable force associated with an UWHB that can be identified by the second definition. Later we will define this force rigorously and use that as part of the definition of dehydron in Section 8.5. In Section 8.5, we will review a more sophisticated technique that incorporates the geometry of nonpolar groups as well as their number to assess the extent of protection via dielectric modulation.

8.2 Assessing polarity

The key to understanding hydrophobicity is polarity. Nonpolar groups repel water molecules (or at least do not attract them strongly) and polar groups attract them. We have already discussed the concept of polarity, e.g., in the case of dipoles (Section 3.3). Similarly, we have noted that certain sidechains, such as glutamine, are polar, even though there the net charge on the sidechain is zero. Here we explain how such polarity can arise due to more subtle differences in charge distribution.

8.2.1 Electronegativity scale

The key to understanding the polarity of certain molecules is the **electronegativity scale** [249, 325, 356], part of which is reproduced in Table 8.1. Atoms with similar electronegativity tend to form nonpolar groups, such as CH_n and to a lesser extent C-S. Atomic pairs with significant differences in electronegativity tend to form polar groups, such as C-O and N-H. The scaling of the electronegativity values is arbitrary, and the value for fluorine has been taken to be exactly four by convention [325].

Let us show how the electronegativity scale can be used to predict polarity. In a C-O group, the O is more electronegative, so it will pull charge (electrons) from C, yielding a pair with a negative charge associated with the O side of the group, and a positive charge associated with the C side of the pair. Similarly, in an N-H group, the N is more electronegative, so it pulls charge from the H, leaving a net negative charge near the N and a net positive charge near the H. This is depicted in Figure 8.1. In Section 8.2.2, we will see that molecular dynamics codes assign such partial charges.

Only the differences in electronegativity have any chemical significance. But these differences can be used to predict the polarity of atomic groups, as we now illustrate for the carbonyl and amide groups. For any atom X, let $\mathcal{E}(X)$ denote the electronegativity of X as defined in Table 8.1. Since $\mathcal{E}(O) > \mathcal{E}(C)$, we conclude that the dipole of the carbonyl group C-O can be represented by a positive charge on the carbon and a negative charge on the oxygen. Similarly, because $\mathcal{E}(N) > \mathcal{E}(H)$, the dipole of the amide group N-H can be represented by a positive charge on the hydrogen and a negative charge on the nitrogen. A more detailed comparison of the electronegativities of C, O, N, and H gives

$$\mathcal{E}(O) - \mathcal{E}(C) = 3.66 - 2.75 = 0.91 > 0.60 = 3.19 - 2.59 = \mathcal{E}(N) - \mathcal{E}(H). \tag{8.1}$$

Thus we conclude that the charge difference in the dipole representation of the carbonyl group (C-O) should be larger than the charge difference in the dipole representation of the amide (N-H) group. Thus, it would be expected to find larger partial charges for C-O than for N-H, as we will see. Of course, the net charge for both C-O and N-H must be zero.

It is beyond our scope to explain electronegativity here, but there is a simple way to comprehend the data. Electronegativity represents the power of an atom to attract electrons in a covalent bond [325]. Thus a stronger positive charge in the nucleus would lead to a stronger attraction of electrons, which is reflected in the correlation between nuclear charge and electronegativity shown in Table 8.1. More precisely, there is a nearly linear relationship between the electronegativity scale and the number of electrons in the outer shell. The value for hydrogen can be explained by realizing that the outer shell is half full, as it is for carbon.

The atoms with a complete outer shell (helium, neon, argon, etc.) are not part of the electronegativity scale, since they have no room to put extra electrons that might be attracted to them. Similarly, atoms with just a few electrons in the outer shell seem to be more likely to donate electrons than acquire them, so their electronegativity is quite small, such as sodium and magnesium. Hydrogen and carbon are in the middle of the scale, not surprisingly, since they are exactly in the middle between being empty and full of electrons.

Figure 8.1: Caricature showing how an atom A with greater electronegativity draws electrons from atom B to which it is covalently bonded: $\mathcal{E}(A) > \mathcal{E}(B)$. The circles indicate the extent of the electron distribution, and the dots indicate the atom nuclei. The $\pm$ signs indicate the resulting polarity of the AB pair.

8.2.2   Polarity of groups

Using the electronegativity scale, we can now estimate the polarity of groups of atoms. For example, the near match of electronegativity of carbon and hydrogen leads to the correct conclusion that the carbonaceous groups CH_n, $n = 0, 1, 2, 3$ are not polar, at least in appropriate contexts. The typically symmetric arrangement of hydrogens also decreases the polarity of a carbonaceous group, at least when the remaining $4-n$ atoms bonded to it are other carbons or atoms of similar electronegativity.

If a carbon is not covalently attached exclusively to carbon or hydrogen then it is likely polarized and carries a partial charge. Thus, C_α carbons and the carbons in the carbonyl (C-O) group in the peptide bonds of all residues are polar. Sidechain carbons are polar if they are covalently attached to

Full name of amino acid	three letter	single letter	The various PDB codes for the nonpolar carbonaceous groups
Alanine	Ala	A	CB
Arginine	Arg	R	CB, CG
Asparagine	Asn	N	CB
Aspartate	Asp	D	CB
Cysteine	Cys	C	(CB)
Glutamine	Gln	Q	CB, CG
Glutamate	Glu	E	CB, CG
Glycine	Gly	G	na
Histidine	His	H	CB
Isoleucine	Ile	I	CB, CG1, CG2, CD1
Leucine	Leu	L	CB, CG, CD1, CD2
Lysine	Lys	K	CB, CG, CD
Methionine	Met	M	CB (CG, CE)
Phenylalanine	Phe	F	CB, CG, CD1, CD2, CE1, CE2, CZ
Proline	Pro	P	CB, CG
Serine	Ser	S	na
Threonine	Thr	T	CG2
Tryptophan	Trp	W	CB, CG, CD2, CE3, CZ2, CZ3, CH2
Tyrosine	Tyr	Y	CB, CG, CD1, CD2, CE1, CE2
Valine	Val	V	CB, CG1, CG2

Table 8.2: PDB codes for nonpolar carbonaceous groups. The carbonaceous groups (CG, CE) surrounding the sulfur in Met and (CB) adjacent to sulfur in Cys may be considered polar. The notation 'na' indicates there are no nonpolar carbonaceous groups.

Full name of compound	PDB code	The various PDB codes for the nonpolar carbonaceous groups
pyroglutamic acid	PCA	CB, CG
phosphorylated tyrosine	PTR	CB, CG, CD1, CD2, CE1, CE2
staurosporine	STU	Ci, $i = 1, \ldots, 7$; $i = 11, \ldots, 16$; C24, C26

Table 8.3: Sample PDB codes and nonpolar carbonaceous groups for some nonstandard amino acids and other compounds.

Residues	atom type	PDB codes	charge
ASP (GLU)	C	CG (CD)	0.27
	OM	ODi (OEi) $i = 1, 2$	-0.635
ASN (GLN)	NT	ND2 (NE2)	-0.83
	H	HD2i (HE2i), $i = 1, 2$	0.415
	C	CG (CD)	0.38
	O	OD1 (OE1)	-0.38
CYS	S	SG	-0.064
	H	HG	0.064
THR	CH1	CB	0.15
	OA	OG1	-0.548
	H	HG1	0.398
SER	CH2	CB	0.15
	OA	OG	-0.548
	H	HG	0.398

Table 8.4: Partial charges from the Gromos force field for polar and negatively charged amino acids.

heteroatoms such as N or O. Sulfur (S) is a closer electronegative match with carbon and polarizes carbon to a lesser extent. The case CH_n with $n = 0$ occurs in the aromatic sidechains in the C_γ position, and there are molecules (e.g., beta-Carotene) in which carbons are bonded only to other carbons. The number of carbon neighbors can be either three or four (e.g., in Fucoxanthin).

To illustrate the polarity of the atoms not listed in Table 8.2, we present the partial charges of the remaining atoms as utilized in the Gromos code in Table 8.4 and Table 8.5. In Table 12.1, partial charges for aromatic sidechains are listed. See [108] for partial charges in the Amber code.

In addition to the the charges shown for the individual sidechain atoms, the backbone is assigned partial charges as follows: the charges of the amide group are ± 0.28 and the carbonyl group are ± 0.38 [17]. That is, in the amide (N-H) group, the N is given a partial charge of -0.28 and the H is given a partial charge of $+0.28$. Similarly, in the carbonyl (C-O) group, the O is given a partial charge of -0.38 and the C is given a partial charge of $+0.38$. Note that the partial charges for C-O are larger than the partial charges for N-H, in accord with our prediction using the electronegativity scale in (8.1).

The N-terminal and C-terminal groups also have appropriate modifications. The C-terminal oxygens have a charge of -0.635, and the attached carbon has a charge of 0.27. The N-terminal nitrogen has a charge of 0.129, and the attached three hydrogens have a charge of 0.248. All of the groups listed in Table 8.2 have zero partial charge.

8.3   Counting residues

In [135], the definition of 'well-wrapped' was based on the proximity of certain residues and defined in relation to the observed distribution of wrapping among a large sample set of proteins. The extent

Residue	atom type	PDB codes	charge
ARG	CH2	CD	0.09
	NE	NE	-0.11
	C	CZ	0.34
	NZ	NHi, $i = 1, 2$	-0.26
	H	HE, HHij, $i, j = 1, 2$	0.24
LYS	CH2	CE	0.127
	NL	NZ	0.129
	H	HZi, $i = 1, 3$	0.248
HIS (A/B)	C	CD2/CG	0.13
	NR	NE2/ND1	-0.58
	CR1	CE1	0.26
	H	HD1/HE2	0.19

Table 8.5: Partial charges from the Gromos force field for positively charged amino acids. The partial charges for His represent two possible ionized states which carry neutral charge.

of hydrogen-bond desolvation was defined by the number of residues ρ_R with at least two *nonpolar* carbonaceous groups (CH$_n$, $n = 0, 1, 2, 3$) whose β-carbon is contained in a specific desolvation domain, as depicted in Figure 8.2. In Section 8.2.2, we explained how to determine the polarity of groups using the electronegativity scale. The nonpolar carbonaceous groups are listed in Table 8.2.

The C_α carbons in all residues are covalently bonded to a nitrogen atom. The mismatch in electronegativity between carbon and nitrogen (Table 8.1) implies that the C_α carbons are polar and thus do not contribute to repelling water. Sidechain carbons are counted only if they are not covalently attached to heteroatoms such as N or O. The CH groups in serine and threonine are attached to an oxygen, which renders them polar. Similarly, a lone carbon that is attached to oxygens is also polar. Thus the seven residues listed in Figure 4.3(a) are eliminated from the group of wrappers, as well as Met and His, in the residue-counting method.

8.3.1   Desolvation domain

The desolvation domain was chosen in [135] to be the union of two (intersecting) 7Å-radius spheres centered at the C_α-carbons of the residues paired by the hydrogen bond, as shown in Figure 8.2. The desolvation circles in Figure 8.2 are drawn artificially large (corresponding to roughly 9Å) in this two-dimensional depiction to show various possibilities.

The choice of the C_α carbons as the centers of the desolvation spheres is justified in Figure 8.3. These figures show that the center of the line joining the centers of the desolvation spheres is often the center of the hydrogen bonds in typical secondary structures. In the case of a parallel β-sheet, the desolvation domain is the same for two parallel hydrogen bonds. The radius represents a typical cutoff distance to evaluate interactions between nearby residues. C_α-carbons which are neighboring in protein sequence are about 3.8Å apart (cf. Exercise 2.2). The distance between other C_α-carbons

is easily determined by datamining in the PDB (cf. Exercise 2.3).

An amide-carbonyl hydrogen bond was defined in [135] by an N-O (heavy-atom) distance within the range 2.6–3.4Å (typical extreme bond lengths) and a 60-degree latitude in the N-H-O angle (cf. Section 6.3). As a scale of reference, at maximum density, water occupies a volume that corresponds to a cube of dimension just over 3.1Å on a side (cf. Section 17.7.3).

The average extent of desolvation, ρ_R, over all backbone hydrogen bonds of a monomeric structure can be computed from any set of structures. In [135], a nonredundant sample of 2811 PDB-structures was examined. The average ρ_R over the entire sample set was found to be 6.6 [135]. For any given structure, the dispersion (standard deviation) σ from the mean value of ρ_R for that structure can be computed. The dispersion averaged over all sampled structures was found to be $\sigma = 1.46$ [135]. These statistics suggested a way to identify the extreme of the wrapping distribution as containing three or fewer wrapping residues in their desolvation domains. This can be interpreted as defining underwrapped as ρ_R values that are more than two standard deviations from the mean.

Figure 8.2: Caricature showing desolvation spheres with various side chains. The open circles denote the nonpolar carbonaceous groups, and the solid circles represent the C_α carbons. The hydrogen bond between the amide (N-H) and carbonyl (O-C) groups is shown with a dashed line. Glycines appear without anything attached to the C_α carbon. There are 22 nonpolar carbonaceous groups in the union of the desolvation spheres and six sidechains with two or more carbonaceous groups whose C_β carbon lie in the spheres.

The distribution of the selected proteins as a function of their average wrapping as measured by ρ_R is shown in Fig. 5 in [135]. The probability distribution has a distinct inflection point at $\rho = 6.2$. Over 90% of the proteins studied have $\rho_R > 6.2$, and none of these are yet known to yield amyloid aggregation under physiological conditions. In addition, individual sites with low wrapping on selected proteins were examined and found to correlate with known binding sites.

In Section 8.3.2, we will see that the known disease-related amyloidogenic proteins are found in the relatively under-populated $3.5 < \rho_R < 6.2$ range of the distribution, with the cellular prion proteins located at the extreme of the spectrum ($3.5 < \rho_R < 3.75$). We discuss there the implications regarding a propensity for organized aggregation. Approximately 60% of the proteins in the critical region $3.5 < \rho_R < 6.2$ which are not known to be amyloidogenic are toxins whose structures are

Figure 8.3: The hydrogen bond (dashed line) configuration in (a) α-helix, (b) antiparallel β-sheet, and (c) parallel β-sheet. A dotted line connects the C_α carbons (squares) that provide the centers of the spheres forming the desolvation domains in Figure 8.2. The amide (N-H) groups are depicted by arrow heads and the carbonyl (O-C) groups are depicted by arrow tails.

stabilized mostly by disulfide bonds.

To further assess the effectiveness of the residue-based assessment of wrapping, we review additional results and predictions of [135].

8.3.2 Predicting aggregation

Prediction of protein aggregation can be based on locating regions of the protein surface with high density of defects which may act as aggregation sites [197, 239, 288]. Figure 3a of [135] depicts the (many) UWHB's for the human cellular prion protein (PDB file 1QM0) [341, 347, 433]. Over half of the hydrogen bonds are UWHB's, indicating that many parts of the structure must be open to water attack. For example, α-helix 1 has the highest concentration of UWHB's, and therefore may be prone to structural rearrangement.

In helix 1 (residues 143 to 156), all of the hydrogen bonds are UWHB's, and this helix has been identified as undergoing an α-helix to β-strand transition [341, 347, 433]. Furthermore, helix 3 (residues 199 to 228) contains a significant concentration of UWHB's at the C-terminus, a region assumed to define the epitope for protein-X binding [341]. The remaining UWHB's occur at the helix-loop junctures and may contribute to flexibility required for rearrangement.

The average underwrapping of hydrogen bonds in an isolated protein may be a significant indicator of aggregation, but it is not likely to be sufficient to determine amyloidogenic propensity. For instance, protein L (PDB file 2PTL) is not known to aggregate even though its $\rho_R = 5.06$ value is outside the standard range of sufficient wrapping. Similarly, trp-repressor (PDB file 2WRP) has $\rho_R = 5.29$, and the factor for inversion stimulation (PDB file 3FIS) has $\rho_R = 4.96$. Many neurotoxins (e.g., PDB file 1CXO with $\rho_R = 3.96$) are in this range as well.

The existence of short fragments endowed with fibrillogenic potential [23, 97, 117, 176, 197, 288, 239] suggests a localization or concentration of amyloid-related structural defects. In view of this, a local wrapping parameter, the maximum density $\delta_{\max}$ of UWHB's on the protein surface, was introduced [135]. A statistical analysis involving $\delta_{\max}$ [135] established that a threshold $\delta_{\max} > 0.38/\text{nm}^2$ distinguishes known disease-related amyloidogenic proteins from other proteins with a

low extent of hydrogen bond wrapping. On the basis of a combined assessment, identifying both low average wrapping and high maximum density of underwrapping, it was predicted [135] that six proteins might posses amyloidogenic propensity. Four of them,

- angiogenin (cf. PDB files 1B1E and 2ANG),

- meizothrombin (cf. PDB file 1A0H),

- plasminogen (cf. PDB file 1B2I), and

- anti-oncogene A (cf. PDB file 1A1U),

are involved in some form of blood clotting, wound healing, or cell adhesion.

Not all protein aggregation is related to disease. Angiogenesis refers to the growth of new capillaries from an existing capillary network, and many processes involve this, including wound healing. Angiogenin is only one of many proteins involved in the angiogenesis process, but it appears to have certain unique properties [256]. Meizothrombin is formed during prothrombin activation, and is known to be involved in blood clotting [220] and is able to bind to procoagulant phospholipid membranes [328]. Plasminogen has been identified as being a significant factor in wound healing [350].

An extreme in dehydron concentration could be a contributor to cell adhesion. Surprisingly, "cell-cell adhesiveness is generally reduced in human cancers" [193]. Thus anti-oncogene A could use its lack of wrapping as a mechanism to foster cell adhesion.

8.4 Counting nonpolar groups

A more refined measure of hydrogen-bond protection has been proposed based on the number of vicinal nonpolar groups [128, 140]. The desolvation domain for a backbone hydrogen bond is defined again as the union of two intersecting spheres centered at the α-carbons of the residues paired by the hydrogen bond, as depicted in Figure 8.2. In this case, all of the dark circles are counted, whether or not the base of the sidechain lies within the desolvation domain. The extent of intramolecular desolvation of a hydrogen bond, ρ_G, is defined by the number of sidechain nonpolar groups (CH_n, $n = 0, 1, 2, 3$) in the desolvation domain (see Table 8.2).

The distribution of wrapping for a large sample of non-redundant proteins is given in Figure 15.1 for a radius of 6Å for the definition of the desolvation domain. In [141], an UWHB was defined by the inequality $\rho_G < 12$ for this value of the radius. Statistical inferences involving this definition of ρ_G were found to be robust to variations in the range 6.4 ± 0.6 Å for the choice of desolvation radius [140, 148]. In Figure 8.4 the distribution of wrapping is presented for a particular PDB file.

The 'group' definition of wrapping is similar to the definition of **buried** groups [260]. This provides a way of defining the difference between entities at the 'surface' of a protein versus the 'core' of the protein. The definition of buried utilized a sphere of radius 15.5Å around each atom. If this sphere contains more than 400 heavy (non-hydrogen) atoms, then the atom is declared to be buried. We can think of this in terms of heavy-atom density, which allows us to compare with the

known sizes (Section 5.4). Roughly speaking, when the local density of heavy atoms is greater than one per 39 $\mathring{A}^3$ (corresponding to a box of side about 3.4 $\mathring{A}$), that region is considered to be buried. For comparison, the average density of water is about one water per 30 $\mathring{A}^3$ (corresponding to a box of side about 3.1 $\mathring{A}$) (cf. Section 17.7.3), whereas we see in Tables 5.4–5.6 that most protein atom groups have a volume less than 39$\mathring{A}^3$, and thus a density of greater than one per 39 $\mathring{A}^3$.
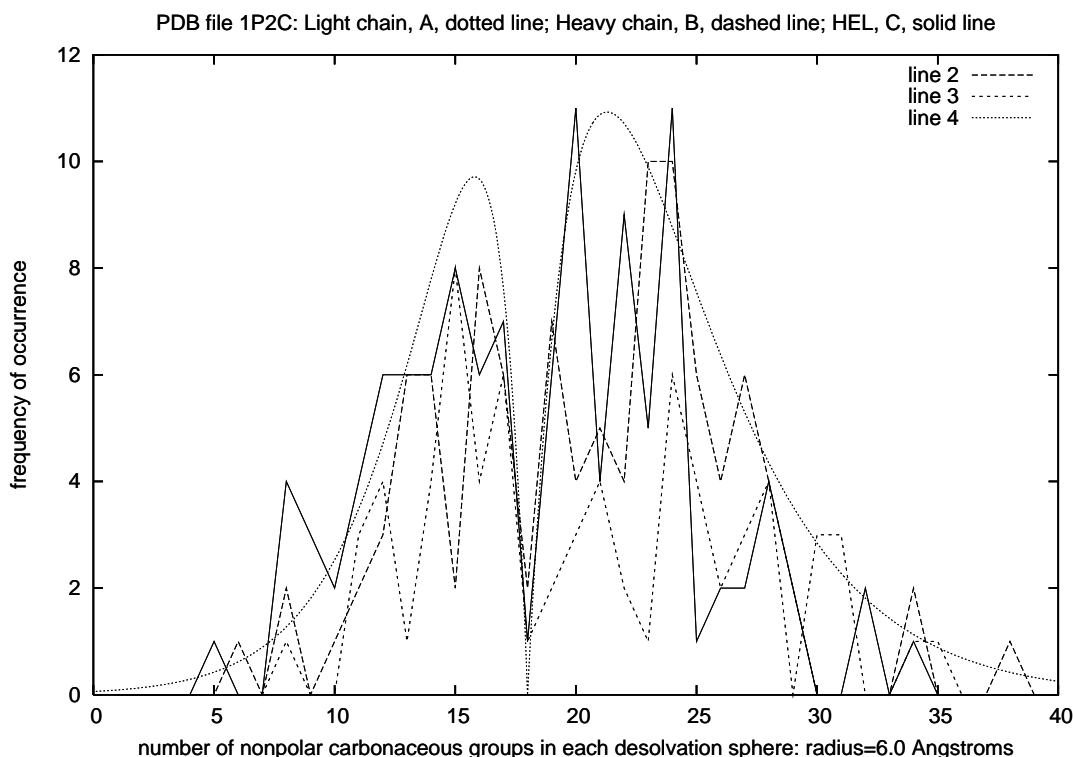
Figure 8.4: Distribution of wrapping for PDB file 1P2C. There are three chains: light, heavy chains of the antibody, and the antigen (HEL) chain. The desolvation radius is 6.0$\mathring{A}$. Smooth curves (8.2) are added as a guide to the eye.

8.4.1   Distribution of wrapping for an antibody complex

It is instructive to consider wrapping of hydrogen bonds from a more detailed statistical point of view. In Figure 8.4 the distribution of wrapping is presented for the antibody complex whose structure is recorded PDB file 1P2C. There are three chains, two in the antibody (the light and heavy chains), and one in the antigen, hen egg-white lysozyme (HEL).

What is striking about the distributions is that they are bi-modal, and roughly comparable for all three chains. We have added a smooth curve representing the distributions

$$d_i(r) = a_i |r - r_0| e^{-|r - r_0|/w_i} \tag{8.2}$$

to interpolate the actual distributions. More precisely, d_1 represents the distribution for $r < r_0$, and d_2 represents the distribution for $r > r_0$. The coefficients chosen were $w_1 = 2.2$ and $w_2 = 3.3$.

The amplitude coefficients were $a_1 = 12$ and $a_2 = 9$, and the offset $r_0 = 18$ for both distributions. In this example, there seems to be a line of demarcation at $\rho = 18$ between hydrogen bonds that are well wrapped and those that are underwrapped.

The distributions in Figure 8.4 were computed with a desolvation radius of 6.0Å. Larger desolvation radii were also used, and the distributions are qualitatively similar. However the sharp gap at $\rho = 18$ becomes blurred for larger values of the desolvation radius.

8.4.2   Residues versus polar groups

The two measures considered here for determining UWHB's share some important key features. Both count sidechain indicators which fall inside of desolvation domains that are centered at the C_α backbone carbons. The residue-based method counts the number of residues (of a restricted type) whose C_β carbons fall inside the desolvation domain. The group-based method counts all of the carbonaceous groups that are found inside the desolvation domain, independent of which residue that they come from.

We observed that the average measure of wrapping based on counting residues was $\rho_R = 6.6$, whereas the average measure of wrapping based on counting non-polar groups is $\rho_G = 15.9$. The residues in the former count represent at least two non-polar groups, so we would expect that $\rho_G > 2\rho_R$. We see that this holds, and that the excess corresponds to the fact that some residues have three or more non-polar groups. Note that these averages were obtained with different desolvation radii, 6.0Å for ρ_G and 7.0Å for ρ_R. Adjusting for this difference would make ρ_G even larger, indicating an even greater discrepancy between the two measures. This implies that ρ_G provides a much finer estimation of local hydrophobicity.

The structural analysis in [135] identified site mutations which might stabilize the part of the cellular prion protein (PDB file 1QM0) believed to nucleate the cellular-to-scrapie transition. The (Met134, Asn159)-hydrogen bond has a residue wrapping factor of only $\rho_R = 3$ and is only protected by Val161 and Arg136 locally, which contribute only a minimal number (five) of non-polar carbonaceous groups. Therefore it is very sensitive to mutations that alter the large-scale context preventing water attack. It was postulated in [135] that a factor that triggers the prion disease is the stabilization of the (Met134, Asn159) β-sheet hydrogen bond by mutations that foster its desolvation beyond wild-type levels.

In the wild type, the only nonadjacent residue in the desolvation domain of hydrogen bond (Met134, Asn159) is Val210, thus conferring marginal stability with $\rho_R = 3$. Two of the three known pathogenic mutations (Val210Ile and Gln217Val) would increase the number of non-polar carbonaceous groups wrapping the hydrogen bond (Met134, Asn159), even though the number of wrapping residues would not change. Thus we see a clearer distinction in the wrapping environment based on counting non-polar carbonaceous groups instead of just residues.

The third known pathogenic mutation, Thr183Ala, may also improve the wrapping of the hydrogen bond (Met134, Asn159) even though our simple counting method will not show this, as both Thr and Ala contribute only one nonpolar carbonaceous group for desolvation. However, Ala is four positions below Thr in Table 7.1 and is less polar than Thr. Table 7.1 reflects a more refined notion of wrapping for different sidechains, but we do not pursue this here.

8.5   Defining dehydrons via geometric requirements

The enhancement of backbone hydrogen-bond strength and stability depends on the partial structuring, immobilization or removal of surrounding water. In this section we review an attempt [142] to quantify this effect using a continuous representation of the local solvent environment surrounding backbone hydrogen bonds [55, 128, 140, 148, 195, 318, 411]. The aim is to estimate the changes in the permittivity (or dielectric coefficient) of such environments and the sensitivity of the Coulomb energy to local environmental perturbations caused by protein interactions [128, 148]. However, induced-fit distortions of monomeric structures are beyond the scope of these techniques.

The new ingredient is a sensitivity parameter M_k assessing the net decrease in the Coulomb energy contribution of the k-th hydrogen bond which would result from an exogenous immobilization, structuring or removal of water due to the approach by a hydrophobic group. This perturbation causes a net decrease in the permittivity of the surrounding environment which becomes more or less pronounced, depending on the pre-existing configuration of surrounding hydrophobes in the monomeric state of the protein. In general, nearby hydrophobic groups induce a structuring of the solvent needed to create a cavity around them, and the net effect of this structuring is a localized reduction in the solvent polarizability with respect to reference bulk levels. This structuring of the solvent environment should be reflected in a decrease of the local dielectric coefficient ϵ. This effect has been quantified in recent work which delineated the role of hydrophobic clustering in the enhancement of dielectric-dependent intramolecular interactions [128, 148].

We now describe an attempt to estimate ϵ as a function of the fixed positions $\left\{ r_j \mid j = 1, \ldots, n_k \right\}$ of surrounding nonpolar hydrophobic groups (CH_n, with $n = 1, 2, 3$, listed in Table 8.2). The simpler estimates of wrapping considered so far could fail to predict an adhesive site when it is produced by an uneven distribution of desolvators around a hydrogen bond, rather than an insufficient number of such desolvators. Based on the fixed atomic framework for the monomeric structure, we now identify Coulomb energy contributions from intramolecular hydrogen bonds that are most sensitive to local environmental perturbations by subsuming the effect of the perturbations as changes in ϵ.

Suppose that the carbonyl oxygen atom is at $\mathbf{r}_O$ and that the partner hydrogen net charge is at $\mathbf{r}_H$. The electrostatic energy contribution $E_{\text{COUL}}(k, \mathbf{r})$ for this hydrogen bond in a dielectric medium with dielectric permittivity $\epsilon(\mathbf{r})$ is approximated (see Chapter 16) by

$$E_{\text{COUL}}(\mathbf{r}) = \frac{-1}{4\pi\epsilon(\mathbf{r})} \frac{qq'}{|\mathbf{r}_O - \mathbf{r}_H|}, \tag{8.3}$$

where q, q' are the net charges involved and where $|\cdot|$ denotes the Euclidean norm.

Now suppose that some agent enters in a way to alter the dielectric field, e.g., a hydrophobe that moves toward the hydrogen bond and disrupts the water that forms the dielectric material. This movement will alter the Coulombic energy as it modifies ϵ, and we can use equation (8.3) to determine an equation for the change in ϵ in terms of the change in E_{COUL}. Such a change in E_{COUL} can be interpreted as a force (cf. Chapter 3). We can compute the resulting effect as a derivative with respect to the position R of the hydrophobe:

$$\nabla_R(1/\epsilon(\mathbf{r})) = \frac{4\pi|\mathbf{r}_O - \mathbf{r}_H|}{qq'} \left(-\nabla_R E_{\text{COUL}}(\mathbf{r}) \right) = \frac{4\pi|\mathbf{r}_O - \mathbf{r}_H|}{qq'} F(\mathbf{r}), \tag{8.4}$$

where $F(\mathbf{r}) = -\nabla_R E_{\text{COUL}}(\mathbf{r})$ is a net force exerted on the hydrophobe by the fixed pre-formed hydrogen bond. This force represents a net 3-body effect [128], involving the bond, the dielectric material (water) and the hydrophobe. If E_{COUL} is decreased in this process, the hydrophobe is attracted to the hydrogen bond because in so doing, it decreases the value of $E_{\text{COUL}}(\mathbf{r})$.

To identify the 'opportune spots' for water exclusion on the surface of native structures we need to first cast the problem within the continuous approach, taking into account that $1/\epsilon$ is the factor in the electrostatic energy that subsumes the influence of the environment. Thus to identify the dehydrons, we need to determine for which Coulombic contributions the exclusion or structuring of surrounding water due to the proximity of a hydrophobic 'test' group produces the most dramatic increase in $1/\epsilon$. The quantity M_k was introduced [142] to quantify the sensitivity of the Coulombic energy for the k-th backbone hydrogen bond to variations in the dielectric. For the k-th backbone hydrogen bond, this sensitivity is quantified as follows.

Define a desolvation domain D_k with border ∂D_k circumscribing the local environment around the k-th backbone hydrogen bond, as depicted in Figure 8.2. In [142], a radius of 7Å was used. The set of vector positions of the n_k hydrophobic groups surrounding the hydrogen bond is extended from $\{\mathbf{r}_j \mid j = 1, 2, \ldots, n_k\}$ to $\{\mathbf{r}_j \mid j = 1, 2, \ldots, n_k; R\}$ by adding the test hydrophobe at position R. Now compute the gradient $\nabla_R(1/\epsilon)|_{R=R_o}$, taken with respect to a perpendicular approach by the test hydrophobe to the center of the hydrogen bond at the point $R = R_o$ located on the circle consisting of the intersection C of the plane perpendicular to the hydrogen bond with the boundary ∂D_k of the desolvation domain. Finally, determine the number

$$M_k = \max\left\{ |\nabla_R(1/\epsilon(\{\mathbf{r}_j\}, R_o, r_H - r_O))| \ \big| \ R_o \in C \right\}. \tag{8.5}$$

The number M_k quantifies the maximum alteration in the local permittivity due to the approach of the test hydrophobe in the plane perpendicular to the hydrogen bond, centered in the middle of the bond, at the surface of the desolvation domain.

The quantity M_k may be interpreted in physical terms as a measure of the maximum possible attractive force exerted on the test hydrophobic group by the pre-formed hydrogen bond. The only difficulty in estimating M_k is that it requires a suitable model of the dielectric permittivity ϵ as a function of the geometry of surrounding hydrophobic groups. We will consider the behavior of the dielectric permittivity more carefully in Chapter 16, but for now we consider a heuristic model used in [142].

The model in [142] for the dielectric may be written

$$\epsilon^{-1} = (\epsilon_o^{-1} - \epsilon_w^{-1})\Omega(\{\mathbf{r}_j\})\Phi(\mathbf{r}_H - \mathbf{r}_O) + \epsilon_w^{-1}, \tag{8.6}$$

where ϵ_w and ϵ_o are the permittivity coefficients of bulk water and vacuum, respectively, and

$$\Omega(\{\mathbf{r}_j\}) = \prod_{j=1,\ldots,n_k} \left(1 + e^{-|\mathbf{r}_O - \mathbf{r}_j|/\Lambda}\right)\left(1 + e^{-|\mathbf{r}_H - \mathbf{r}_j|/\Lambda}\right) \tag{8.7}$$

provides an estimate of the change in permittivity due to the hydrophobic effects of the carbonaceous groups. In [142], a value of $\Lambda = 1.8$Å was chosen to represent the characteristic length associated with the water-structuring effect induced by the solvent organization around the hydrophobic
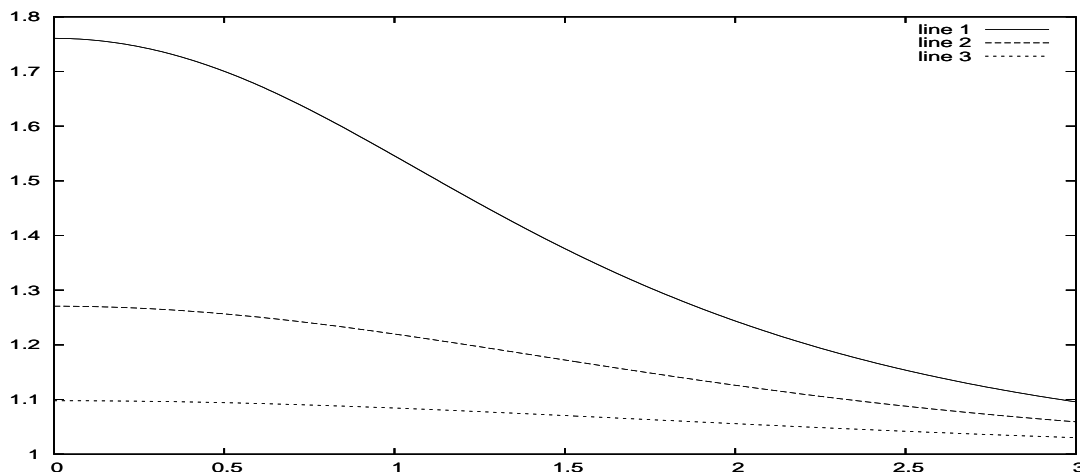
Figure 8.5: The function $\omega(x, y)$ plotted as a function of the distance along the x-axis connecting r_H and r_O, for three different values of the distance y from that axis: $y = 1$ (solid line), $y = 2$ (dashed line), $y = 3$ (dotted line). The coordinates have been scaled by Λ and the value of $|\mathbf{r}_O - \mathbf{r}_H| = 1$ was assumed.

groups. Further, a cut-off function

$$\Phi(\mathbf{r}) = (1 + |\mathbf{r}|/\xi) \, e^{-|\mathbf{r}|/\xi}, \tag{8.8}$$

where $\xi = 5\text{Å}$ is a water dipole-dipole correlation length, approximates the effect of hydrogen bond length on its strength [142].

We can write the key expression Ω in (8.7) as

$$\Omega(\{\mathbf{r}_j\}) = \prod_{j=1,\ldots,n_k} \omega(\mathbf{r}_j), \tag{8.9}$$

where the function ω is defined by

$$\omega(\mathbf{r}) = \left(1 + e^{-|\mathbf{r}_O - \mathbf{r}|/\Lambda}\right) \left(1 + e^{-|\mathbf{r}_H - \mathbf{r}|/\Lambda}\right). \tag{8.10}$$

The function ω is never smaller than one, and it is maximal in the plane perpendicular to the line connecting r_H and r_O. Moreover, it is cylindrically symmetric around this axis. The values of ω are plotted in Figure 8.5 as a function of the distance from the perpendicular bisector of the axis connecting $\mathbf{r}_H$ and $\mathbf{r}_O$, for three different values of the distance y from the line connecting $\mathbf{r}_H$ and $\mathbf{r}_O$.

We see that the deviation in ω provides a strong spatial dependence on the dielectric coefficient in this model. Thus hydrophobes close to the plane bisecting the line connecting r_H and r_O are counted more strongly than those away from that plane, for a given distance from the axis, and those closer to the line connecting r_H and r_O are counted more strongly than those further away. When the product $\Phi\Omega = 1$, we get $\epsilon = \epsilon_o$ reflecting the maximal amount of water exclusion possible.

Correspondingly, if $\Phi\Omega$ tends to zero, then ϵ tends to ϵ_w, yielding a dielectric similar to bulk-water. Thus bigger values of Ω correspond to the effect of wrapping.

The definition (8.6) of the dielectric has not been scaled in a way that assures a limiting value of $\epsilon = \epsilon_o$. However, since we are only interested in comparing relative dielectric strength, this scaling is inessential. What matters is that larger values of Ω correspond to a lower dielectric and thus stronger bonds.

The computation of M_k involves computing the gradient of

$$\Omega(\{\mathbf{r}_1, \ldots, \mathbf{r}_{n_k}, R\}) = \omega(R)\Omega(\{\mathbf{r}_1, \ldots, \mathbf{r}_{n_k}\}) \tag{8.11}$$

with respect to R. Due to the cylindrical symmetry of ω, $|\nabla_R \omega|_{R=R_o}$ is a constant depending only on the desolvation radius $|R_o|$ and the hydrogen bond length $|\mathbf{r}_O - \mathbf{r}_H|$ for all $R_o \in C$. Thus, for a fixed desolvation radius $|R_o|$, M_k may be written as a function of $|\mathbf{r}_O - \mathbf{r}_H|$ times $\Omega(\{\mathbf{r}_1, \ldots, \mathbf{r}_{n_k}\})$ when using the model (8.6).

A sensitivity threshold for hydrogen bonds was established in [142] by statistical analysis on a sample of native structures for soluble proteins. Only 8% of backbone hydrogen bonds from a sample of 702 proteins, of moderate sizes ($52 < N < 110$) and free from sequence redundancies [194], were found to be highly sensitive in the sense that

$$M_k > \lambda/10, \tag{8.12}$$

where λ was defined to be

$$\lambda = \frac{\epsilon_o^{-1} - \epsilon_w^{-1}}{2\mathring{A}}. \tag{8.13}$$

On the other hand, 91.6% of backbone hydrogen bonds were found to be relatively insensitive to water removal, namely,

$$0 < M_k < \lambda/100 \tag{8.14}$$

This remarkable separation in the (nearly bimodal) distribution of sensitivities led [142] to the definition of a dehydron as a backbone hydrogen bond satisfying (8.12).

8.6 Dynamic models

A dehydron is, by definition, a hydrogen bond that benefits from dehydration, e.g., upon binding by a ligand. This is an inherently dynamic description. In Section 8.5, an attempt [142] was described to approximate this dynamic picture. Instead of simulating the approach of a dehydrating group, a derivative was defined to estimate the force associated with the change in dielectric due to the approaching hydrophobic group. A dynamic assessment for a model problem was carried out in [134]. This provides a prototype for a potentially improved prediction of dehydrons.

An alternate approach to quantifying hydrophobicity is to consider the behavior of water directly. When doing molecular dynamics simulations, it is possible to measure the residence times for water molecules near a hydrogen bond [91, 92, 93, 133, 138, 292] or other site of interest. The residence times are distincly different for well wrapped and underwrapped hydrogen bonds [91]. Thus such

residence times may be used as predictors of sites of interest [138, 292]. It has similarly been found that water behavior near 'wet' residues, namely, those involved in intermolecular interactions mediated by a water molecule, is also distinct [353].

8.7 Genetic code

The genetic code, depicted in Figure 8.6, describes the mapping from DNA (or RNA) to protein residues. DNA is first translated (in a one-to-one fashion) to RNA, and so we use the RNA code letters A, C, G, U to describe the code. A DNA sequence of an RNA sequence consists of sequences of triplets, called **codons**, and each codon specifies a particular residue, with the exception of three that indicate that translation of DNA/RNA to proteins is to stop. There are 64 such codon values, so there is significant redundancy, but two letter codons (with only 16 values) would not be sufficint to describe 20 residues. To read a gene sequence, one must first determine the **reading frame**, that is, where the codons begin, and there are three possible reading frames, each shifted by one letter.

The initial discovery of the code was not easy, and there were many false steps in the process [184]. One proposed code by Crick and co-workers was a **comma free code** with the property that there was only one possible reading frame for the code [184]. Such codes were mathematically appealing but would be too brittle: most single-letter mutations would produce codons that did not code for a protein residue. Evolution requires mutation of DNA, and resulting mutation of residues, and the simplest mutation would be to a single letter in the DNA (and resulting RNA) sequence. The genetic code insures that most such mutations lead from one residue to another, although some would lead to a stop code, which could be fatal.

The topology of the genetic code is fairly simple using the standard representation used in Figure 8.6. As a first approximation, it is a two-dimensional toroidal grid, and single-letter mutations in codons represent a shift in the horizontal or vertical directions (with wraparound). Thus Leu to Pro is a mutation that can be done with a single-letter mutation, as well as Pro to Thr and Thr to Ala. But there are some grid boxes divided into two sub-boxes, and these should be visualized as extending vertically in a third coordinate perpendicular to the page. Some single-letter mutations will thus be vertical in direction, and mutations between vertical boxes require appropriate interpretation. A movement in the coordinate directions of the plane of the page is a single point mutation if it goes between two upper boxes, or between two lower boxes, but not between an upper and a lower box in neighboring squares. Thus His to Lys requires a two-letter mutation. Boxes without dashed lines separating them (like the one for Pro) should be visualized as extending vertically as well, so that a horizontal step represents single-letter mutations between Pro and His as well as between Pro and Gln.

The individual residue cells are all connected with the exception of Ser. The Arg cell may appear at first to be disconnected, but it is an L shaped domain. Not surprisingly, the cell for Trp is one of the smallest, but so is that of Met, whose codon plays the dual role of a start codon to signal the initiation of DNA/RNA translation.

Second Position

First Position	u	c	a	g	Third Position
u	uuu uuc ⌉ Phe 7 || uua uug ⌉ Leu 4 ||	ucu ucc uca ucg — Ser 0 +−	uau uac ⌉ Tyr 6 +− uaa stop uag stop	ugu ugc ⌉ Cys 0 +− uga stop ugg Trp 7 +−	u c a g
c	cuu cuc cua cug — Leu 4 ||	ccu ccc cca ccg — Pro 2 ||	cau cac ⌉ His 1 +− caa cag ⌉ Gln 2 +−	cgu cgc cga cgg — Arg 2 ++	u c a g
a	auu auc aua — Ile 4 || aug Met 1 +−	acu acc aca acg — Thr 1 +−	aau aac ⌉ Asn 1 +− aaa aag ⌉ Lys 3 ++	agu agc ⌉ Ser 0 +− aga agg ⌉ Arg 2 ++	u c a g
g	guu guc gua gug — Val 3 ||	gcu gcc gca gcg — Ala 1 ||	gau gac ⌉ Asp 1 −− gaa gag ⌉ Glu 2 −−	ggu ggc gga ggg — Gly 0 +−	u c a g

Figure 8.6: The genetic code. The first digit after the residue name is the amount of wrapping and the second indicator is polarity: $||$ is nonpolar, $+-$ is polar, $--$ is negatively charged and $++$ is positively charged. The code segregates residues by polarity, nonpolars on the left, strongly polar residues on the right. Wrapping is more diversely distributed.

8.7.1   Interpretation of the genetic code

Once the actual code was determined, there was extensive effort to explain why it might have evolved into an "optimal" code according to some criteria [11, 101, 168, 426]. Certain properties of the genetic code have been long recognized, such as the segregation of residues into polar and hydrophobic groups [254]. In Figure 8.6, we see that most of the residues on the left half of the code are hydrophobic, and all of the residues on the right half have some polar component. This property of the genetic code implies that most single-codon mutations will not change the polarity of the residue drammatically. To measure polarity, we propose a simple system consisting of four states: nonpolar, polar and charged, where there are two signs for charged residues. This uses a simple interpretation of the polar scale as indicated in Figure 8.7.
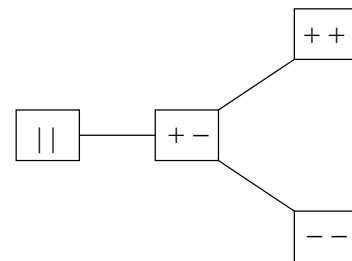
Figure 8.7: A polarity scale for sidechainss: $||$ is nonpolar, $+-$ is polar, $--$ is negatively charged and $++$ is positively charged.

The genetic code has only a few places where a single-letter codon mutation can change the polarity more than one step in the polarity scale in Figure 8.7. Mutations between Lys and Glu occur when the first codon changes between a and g. Similarly, Ala can mutate to Asp and Glu, and Arg can mutate to Ile and Leu. Such mutations might be catastrophic, but they represent only a small fraction of the possible mutations. Otherwise, all single-letter codon mutations either change between nonpolar and polar ($+-$) residues, or between polar and charged residues. Moreover, mutations occurring on the left hand side of the code involve mostly nonpolar residues.

Although the genetic code has the property that single-letter codon mutations cause minimal change in polarity, they can cause signifcant change in wrapping. For example, the three residues with maximal wrapping (Phe, Tyr, Trp) are isolated in the genetic code (requiring more than a single-letter mutation to move between them). By contrast, Arg and Lys are neighbors, as are Asp and Glu. Thus the genetic code allows radical changes in wrapping to be explored while keeping changes in polarity to a minimum.

There are only 50 possible mutations from one residue to another (or 100 if you count both directions) based on single-letter mutations in DNA/RNA. These mutations are listed in Table 8.6. The rest of the 190 pairs of possible transitions from one residue to another (or 380 if you count both directions) require at least two-letter mutations, and thus would require a more coordinated mutation strategy, or a sequnce of changes. The sparsity of the genetic code in terms of possible residue mutations is a significant feature.

The number of possible single-letter DNA/RNA mutations is of course much larger, on the order of 64 squared. Thus most single-letter DNA/RNA mutations are **synonomous mutations** in that they leave the residue unchanged. This is another important property of the genetic code. Since the genetic code is a three-letter code, any residue can be mutated to any other by a three-letter mutation. However, not all pairs of residues can be linked by a two-letter mutation (Exercise 8.9).

Of the 50 possible mutations from one residue to another based on single-letter mutations in DNA/RNA, 23 of them involve no change in polarity. We might call them **polarity-synonomous mutations**. Thus only 27 mutations based on single-letter mutations in DNA/RNA, of the possible 190 pairs of possible transitions from one residue to another, have any polarity change, and 22 of

Res	Mutations
Ala	Asp(0,↑↑), Glu(+1,↑↑), Ser(−1,↑), Thr(0,↑), Val(+2,◯)
Arg	Gln(0,↓), Gly(−2,↓), His(−1,↓), Ile(+2,↓↓),
	Leu(+2,↓↓), Lys(+1,◯), Met(−1,↓), Ser(−2,↓)
Asn	Asp(0,↑), His(0,◯), Ser(−1,◯), Thr(0,◯)
Asp	Glu(+1,◯), Gly(−2,↓), Tyr(+6,◯)
Cys	Gly(0,◯), Phe(+7,◯), Trp(+7,◯), Tyr(+6,◯)
Gln	His(−1,◯), Lys(+1,↑), Pro(0,↓)
Glu	Gly(−2,↓), Lys(+1,↑↑)
Gly	Ser(0,◯), Val(+3,↓), Trp(+7,◯)
His	Pro(0,↓), Tyr(+5,◯)
Ile	Leu(0,◯), Met(−3,↑), Ser(−2,↓), Thr(−3,↑)
Leu	Phe(+3,◯), Pro(−2,◯), Trp(+3,◯), Val(-1,◯)
Lys	Thr(−2,↓)
Met	Thr(0,◯), Val(+2,↓)
Phe	Ser(−7,↑), Val(−4,◯)
Pro	Ser(−2,↑), Thr(−1,↑)
Ser	Tyr(+6,◯)

Table 8.6: Amino acid mutations caused by single-letter DNA/RNA mutations. Only mutations to residues later in the alphabet are listed. The table can be extended antisymmetrically (Exercise 8.8). The symbol ↑ indicates a move to the right in the polarity scale in Figure 8.7, and ↓ indicates a move to the left. The symbol ◯ indicates no polarity change.

these involve just one step in our simple polarity scale depicted in Figure 8.7. On the other hand, only 12 of the mutations listed in Table 8.6 are neutral with respect to wrapping change. The overwhelming majority of mutations involve a change in wrapping, and over half of the total involve a change of two or more units.

Thus we can describe the genetic code according to three attributes. First, it minimizes the effect of most single-letter codon mutations by making them synonomous, not changing the residue at all. This provides basic protection against any kind of DNA/RNA mutation. Second, the code minimizes polarity changes due to single-letter codon mutations, limiting substantial chemical changes due to single-letter codon mutations. Finally, the genetic code enables wrapping changes via single-letter codon mutations, so that the inevitable changes due to natural mutations in DNA/RNA are directed toward a useful but less dangerous path.

These properties of the genetic code are easy to understand. The first property relaxes constraints on the system that translates DNA/RNA to protein residues, and it allows simpler systems to function adequately and with less energetic cost. Changes in polarity have a strong impact on chemical changes, and these can likely cause an unwanted change in function. Thus the genetic code makes it hard to make radical changes in polarity. The only remaining mechanism to explore mutation space is changes in wrapping. As we know, these do cause important changes in function, but they do so in a much more subtle way.

8.7.2   Size matters

Another variable that is of interest is the size of sidechains. A mutation that changes the size of a sidechain could have a significant effect. Fortunately, the number of wrappers provides a very simply proxy for the size of the sidechain. For nonpolar residues, including Ala and Gly, the number of wrappers is the number of heavy atoms in the sidechain. For polar and charged residues, there is a difference, but we see in Table 8.7 that the differences are largely predictable. That is, the difference between the number heavy atoms and the number of wrappers, which is the number of polar atoms or charged atoms, is for the most part either two or three. The exceptions are Arg and His. Therefore, we can take the change in the number of wrappers as a good proxy for the size variation upon mutation among polar and charged residues.

The genetic code is not a simple translation table [43]. There is an amino acid residue called selenocysteine (with three-letter code Sec and single-letter code U) which is a cysteine with the sulfur atom replaced by selenium. The genetic code for this is UGA, which is typically a stop code, and the interpretation of UGA is then context dependent. Similarly, the amino acid residue pyrrolysine (abbreviated by Pyl and O) uses the the (normally stop) codon UAG, and the translation is again context dependent. Pyl is related to Lys in structure, but has a significant modification, including a pentagonal ring, at the terminal end of the sidechain. Finally, the codon CUG can be translated ambiguously as Ser or Lys depending on context in some organisms [43]. Further complexities in code translation, including the time-dependence of different synonymous codes (codons that code for the same residue), are discussed in [43].

3-letter code	1-letter code	Nonpolar Carbons	Heavy Atoms	Polar Atoms
Asp	D	1	4	3
Asn	N	1	4	3
Cys	C	0	2	2
Glu	E	2	5	3
Gln	Q	2	5	3
Lys	K	3	5	2
Met	M	1	4	3
Ser	S	0	2	2
Thr	T	1	3	2
Tyr	Y	6	8	2
Arg	R	2	7	5
His	H	1	6	5

Table 8.7: The size of residues (the number of heavy atoms) compared with the number of nonpolar carbons. The difference is the number of polar and charged amino acids residues, and is typically only two or three.

8.8   Exercises

Exercise 8.1 *It was observed [135] that the two proteins*

- *RADR zinc finger peptide (PDB file 1A1K) and*

- *rubredoxin (PDB file 1B20).*

have extreme concentrations of dehydrons. Investigate these two proteins to see why this might influence their function.

Exercise 8.2 *In Section 5.2.4, we noted that sidechains have different conformations. Determine and list the number of different rotameric states possible for each sidechain (hint: read [272]). Explain your reasoning. Compare the number of rotameric degrees of freedom (number of states) for the seven residues listed in Figure 4.3(a) with the remaining group of thirteen sidechains. (Hint: review Exercises 5.5 and 5.6.)*

Exercise 8.3 *In Figure 8.3(a), it appears that the dotted line joining the two C_α cabons intersects the dashed line joining the amide and carbonyl groups. By searching the PDB, determine the distribution of distances between the midpoints of these two lines for α-helices.*

Exercise 8.4 *The residues S, T, and Y can become phoshorylated. Various web sites record the locations where these residues can be expected to become phoshorylated. Investigate any possible correlation between wrapping and phosphorylation sites.*

Exercise 8.5 *Explain whether you would expect methyl flouride to have a polar carbon, based on the electronegativity scale.*

Exercise 8.6 *Determine the extent to which the wrapping in a given protein is bimodal, as depicted in Figure 8.4. Determine a way to measure consistently the degree to which a distribution is bimodal, and survey a large set of PDB structures with this measure.*

Exercise 8.7 *Count the number of synonomous mutations of DNA/RNA, that is single-letter mutation in DNA/RNA, that do not change the residue.*

Exercise 8.8 *Make a table of all residue mutations that can arise from a single-letter mutation using the genetic code topology. (Hint: complete Table 8.6 by including the mutations to residues lower in the alphabet; the wrapping number should switch sign, and the direction of the polarity change should invert.)*

Exercise 8.9 *Make a list of all residue pairs that can be linked only by three-letter codon mutations, such as Met and His, Ile and Trp, Lys and Phe, and so on. That is, determine the complement of pairs that can be reached by either one-letter or two-letter codon mutations. To be precise, we say that two residues can be linked by an n-letter mutation if, for any three letter code representing one residue, there is a three letter code representing the other residue that differs by n letters. Thus any residue can be linked to any other by an n-letter mutation for some $n \leq 3$. We are asking for pairs where there is no link for $n \leq 2$. (Hint: look at diagonal moves in the genetic code that are two squares away, and the moves that a knight can make on a chess board, e.g., two over and one up/down. Examine whether these can be reached by a two-letter codon mutation.)*

Exercise 8.10 *Make a 20×20 matrix indexed by the 20 standard amino acids whose entries are the wrapping change arising from a single-letter mutation using the genetic code topology. If no such mutation exists, let the matrix entry be NaN. Make sure that matrix is antisymmetric. Leave the diagonal blank. (Hint: capture the wrapping data from Table 8.6 to generate the upper-triangular part of the matrix; using octave, set the non-existing entries to be 1/0. Extend antisymmetrically.)*

Chapter 9

Stickiness of dehydrons

We have explained why under-wrapped hydrogen bonds benefit from the removal of water. This makes them susceptible to interaction with molecules that can replace water molecules in the vicinity of the hydrogen bond. Conceptually, this implies that under-wrapped hydrogen bonds attract entities that can dehydrate them. Thus they must be sticky. If so, it must be possible to observe this experimentally. Here we review several papers that substantiate this conclusion. One of them involves a mesoscopic measurement of the force associated with a dehydron [141]. A second presents data on the direct measurement of the dehydronic force using atomic force microscopy [124]. Another paper examines the effect of such a force on a deformable surface [143].

9.1 Surface adherence force

We defined the notion of an under-wrapped hydrogen bond by a simple counting method in Chapter 8 and have asserted that there is a force associated with UHWB's. Here we describe measurements of the adhesion of an under-wrapped hydrogen bond by analyzing the flow-rate dependence of the adsorption uptake of soluble proteins onto a phospholipid bilayer.

9.1.1 Biological surfaces

The principal biological surface of interest is the cell membrane. This is a complex system, but a key component is what is called a **phospholipid bilayer**. The term **lipid** refers to a type of molecule that is a long carbonaceous polymer with a polar (phospho) group at the 'head.' This it is hydrophobic at one end and hydrophilic at the other. These molecules align to form a complex that could be described as a bundle of pencils, with the hydrophilic head group (the eraser) at one side of the surface and the hydrophobic 'tail' on the other side. These bundles can grow to form a surface when enough pencils are added. A second surface can form in the opposite orientation, with the two hydrophobic surfaces in close proximity. This results in a membrane that is hydrophilic on both sides, and thus can persist in an aqueous environment.

One might wonder what holds together a lipid bilayer. We have noted that there is a significant volume change when a hydrophobic molecule gets removed from water contact in Section 5.4. The

volume change causes self-assembly of lipids and provides a substantial pressure that holds the surface together. The architecture of a lipid bilayer is extremely adaptive. For example, a curved surface can be formed simply by allocating more lipid to one side than the other. Moreover, it easily allows insertion of other molecules of complex shape but with other composition. Much of a cell membrane is lipid, but there are also proteins with various functions as well as other molecules such as cholesterol. However, a simple lipid bilayer provides a useful model biological surface.

9.1.2 Soluble proteins on a surface

One natural experiment to perform is to release soluble proteins in solution near a lipid bilayer and to see to what extent they attach to the bilayer. Such an experiment [129] indicated a significant correlation between the under-wrapping of hydrogen bonds and bilayer attachment. The results were explained by assuming that the probability of successful landing on the liquid-solid interface is proportional to the ratio of UWHB's to all hydrogen bonds on the protein surface. Here, the number of surface hydrogen bonds is taken simply as a measure of the surface area. Thus the ratio can be thought of as an estimate of the fraction of the surface of the protein that is under-wrapped. The experiments in [129] indicated that more dehydrons lead to more attachments, strongly suggesting that dehydrons are sticky. However, such indications were only qualitative.

A more refined analysis of lipid bilayer experiments was able to quantify a force of attachment [141]. The average magnitude of the attractive force exerted by an UWHB on a surface was assessed based on measuring the dependence of the adsorption uptake on the flow rate of the ambient fluid above the surface. The adhesive force was measured via the decrease in attachment as the flow rate was increased.

Six proteins were investigated in [141], as shown in Table 9.1, together with their numbers of well-wrapped hydrogen bonds as well as dehydrons. The UWHB's for three of these are shown in Fig. 1a-c in [141]. The particular surface was a Langmuir-Blodgett bilayer made of the lipid DLPC (1,2 dilauroyl-sn-glycero-3 phosphatidylcholine) [349]. We now review the model used in [141] to interpret the data.

9.2 A two-zone model

In [141], a two-zone model of surface adhesion was developed. The first zone deals with the experimental geometry and predicts the number of proteins that are likely to reach a fluid boundary layer close to the lipid bylayer. The probability Π of arrival is dependent on the particular experiment, so we only summarize the model results from [141]. The second zone is the fluid boundary layer close to the lipid bilayer, where binding can occur. In this layer, the probability P of binding is determined by the thermal oscillations of the molecules and the solvent as well as the energy of binding.

The number M of adsorbed molecules is given by

$$M = \Pi P(n_{UW}, n_W, T)N \tag{9.1}$$

where Π is the fraction of molecules that reach the (immobile) bottom layer of the fluid, $P(n_{UW}, n_W, T)$ is the conditional probability of a successful attachment at temperature T given that the bottom layer has been reached, and N is the average number of protein molecules in solution in the cell. The quantities n_{UW} and n_W are the numbers of underwrapped and well-wrapped hydrogen bonds on the surface of the protein, respectively. These will be used to estimate the relative amount of protein surface area related to dehydrons. The fraction Π depends on details of the experimental design, so we focus initially on on the second term P.

9.2.1 Boundary zone model

Suppose that ΔU is the average decrease in Coulombic energy associated with the desolvation of a dehydron upon adhesion. It is the value of ΔU that we are seeking to determine. Let ΔV be the Coulombic energy decrease upon binding at any other site. Let f be the fraction of the surface covered by dehydrons. As a simplified approximation, we assume that

$$f \approx \frac{n_{UW}}{n_{UW} + n_W}. \tag{9.2}$$

Then the probability of attachment at a dehydron is predicted by thermodynamics as

$$P(n_{UW}, n_W, T) = \frac{fe^{\Delta U/k_B T}}{(1-f)e^{\Delta V/k_B T} + fe^{\Delta U/k_B T}} \approx \frac{n_{UW}e^{\Delta U/k_B T}}{n_W e^{\Delta V/k_B T} + n_{UW}e^{\Delta U/k_B T}}, \tag{9.3}$$

with $k_B =$ Boltzmann's constant. In [141], ΔV was assumed to be zero. In this case, (9.3) simplifies to

$$P(n_{UW}, n_W, T) = \frac{fe^{\Delta U/k_B T}}{(1-f) + fe^{\Delta U/k_B T}} \approx \frac{n_{UW}e^{\Delta U/k_B T}}{n_W + n_{UW}e^{\Delta U/k_B T}} \tag{9.4}$$

(cf. equation (2) of [141]). Note that this probability is lower if $\Delta V > 0$.

9.2.2 Diffusion zone model

The probability Π in (9.1) of penetrating the bottom layer of the fluid is estimated in [141] by a model for diffusion via Brownian motion in the plane orthogonal to the flow direction. This depends on the solvent bulk viscosity μ, and the molecular mass m and the **hydrodynamic radius** [226] or **Stokes radius** [185] of the protein. This radius R associates with each protein an equivalent sphere that has approximately the same flow characteristics at low Reynolds numbers. This particular instance of a 'spherical cow' approximation [106, 240] is very accurate, since the variation in flow characteristics due to shape variation is quite small [226]. The drag on a sphere of radius R, at low Reynolds numbers, is $F = 6\pi R \mu v$ where v is the velocity. The drag is a force that acts on the sphere through a viscous interaction. The coefficient

$$\xi = 6\pi R\mu/m = F/mv \tag{9.5}$$

where m is the molecular mass, is a temporal frequency (units: inverse time) that characterizes Brownian motion of a protein. The main non-dimensional factor that appears in the model is

$$\alpha = \frac{m\xi^2 L^2}{2k_B T} = \frac{L^2(6\pi R\mu)^2/m}{2k_B T}, \tag{9.6}$$

which has units of energy in numerator and denominator. We have [4]

$$
\begin{aligned}
\Pi(v, R, m) &= \int_\Lambda \int_{\Omega \setminus \Lambda} \int_{[0,\tau]} \frac{\alpha L^{-2}}{\pi \Gamma(t)} e^{-\alpha L^{-2}|\mathbf{r}-\mathbf{r}_0|^2/\Gamma(t)} \, dt d\mathbf{r}_0 d\mathbf{r} \\
&= \int_{\tilde\Lambda} \int_{\tilde\Omega \setminus \tilde\Lambda} \int_{[0,L/v]} \frac{\alpha}{\pi \Gamma(t)} e^{-\alpha|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}_0|^2/\Gamma(t)} \, dt d\tilde{\mathbf{r}}_0 d\tilde{\mathbf{r}}
\end{aligned}
\tag{9.7}
$$

where $\mathbf{r}$ is the two-dimensional position vector representing the cell cross-section Ω, $|\mathbf{r}|$ denotes the Euclidean norm of $\mathbf{r}$, Λ is the 6Å×10^8Å cross-section of the bottom layer, and $\Gamma(t) = 2\xi t - 3 + 4e^{-\xi t} - e^{-2\xi t}$. Note that Γ grows like $2\xi t$ for t large, but initially there is a different behavior that corresponds to a correction to account for the discrete nature of physical diffusion of particles of finite size [401].

The domains $\tilde\Lambda$ and $\tilde\Omega$ represent domains scaled by the length L, and thus the variables $\tilde{\mathbf{r}}$ and $\tilde{\mathbf{r}}_0$ are non-dimensional. In particular, the length of $\tilde\Lambda$ and $\tilde\Omega$ is one in the horizontal coordinate. Note that $\Gamma(t) = \frac{2}{3}(\xi t)^3 + \mathcal{O}((\xi t)^4)$ for ξt small. Also, since the mass m of a protein tends to grow with the radius cubed, α actually decreases like $1/R$ as the Stokes radius increases.

9.2.3 Model validity

The validity of the model represented by equations (9.1—9.7) was established by data fitting. The only parameter in the model, ΔU, was varied, and a value was found that consistently fits within the confidence band for the adsorption data for the six proteins (see Fig. 3 of [141]) across the entire range of flow velocities v. This value is

$$\Delta U = 3.91 \pm 0.67\,\text{kJ/mole} = \Delta U = 0.934 \pm 0.16\,\text{kcal/mole}. \tag{9.8}$$

This value is within the range of energies associated with typical hydrogen bonds. Thus we can think of a dehydron as a hydrogen bond that gets turned 'on' by the removal of water due to the binding of a ligand.

Using the estimate (9.8) of the binding energy for a dehydron, an estimate was made [141] of the force

$$|F| = 7.78 \pm 1.5\text{pN} \tag{9.9}$$

exerted by the surface on a single protein molecule at a 6Å distance from the dehydron.

9.3 Direct force measurement

The experimental techniques reviewed in the previous section suggest that the density of dehydrons correlates with protein stickiness. However, the techniques are based on measuring the aggregate

protein name	PDB code	residues	WWHB	dehydrons
apolipoprotein A-I	1AV1	201	121	66
β lactoglobulin	1BEB	150	106	3
hen egg-white lysozyme	133L	130	34	13
human apomyoglobin	2HBC	146	34	3
monomeric human insulin	6INS	50	30	14
human β_2-microglobulin	1I4F	100	17	9

Table 9.1: Six proteins and their hydrogen bond distributions. WWHB=well-wrapped hydrogen bonds.

behavior of a large number of proteins. One might ask for more targeted experiments seeking to isolate the force of a dehydron, or at least a small group of dehydrons. Such experiments were reported in [124] based on atomic force microscopy (AFM).

We will not give the details of the experimental setup, but just describe the main points. The main concept was to attach hydrophobic groups to the tip of an atomic force microscope. These were then lowered onto a surface capable of forming arrays of dehydrons. This surface was formed by a self-assembling monolayer of the molecules SH-$(CH_2)_{11}$-OH. The OH "head" groups are capable of making OH-OH hydrogen bonds, but these will be exposed to solvent and not well protected.

The data obtained by lowering a hydrophobic probe on such a monolayer are complex to interpret. However, they become easier when they are compared with a similar monolayer not containing dehydrons. In [124], the molecule SH-$(CH_2)_{11}$-Cl was chosen.

The force-displacement curve provided by the AFM have similarities for both monolayers [124]. For large displacements, there is no force, and for very small displacements the force grows substantially as the tip is driven into the monolayer. However, in between, the characteristics are quite different.

For the OH-headed monolayer, as the displacement is decreased to the point where the hydrophobic group on the tip begins to interact with the monolayer, the force on the tip decreases, indicating a force of attraction. Near the same point of displacement, the force on the tip increases for the chlorine-headed monolayer. Thus we see the action of the dehydronic force in attracting the hydrophobes to the dehydron-rich OH-headed layer. On the other hand, there is a resistance at the similar displacement as the hydrophobic tip begins to dehydrate the chlorine-headed monolayer. Ultimately, the force of resistance reaches a maximum, and then the force actually decreases to a slightly negative (attractive) value as the monolayer becomes fully dehydrated. It is significant that the displacement for the force minimum is approximately the same for both monolayers, indicating that they both correspond to a fully dehydrated state.

The force-displacement curves when the tip is removed from the surface also provide important data on the dehydronic force. The force is negative for rather large displacements, indicating the delay due to the requirements of rehydration. Breaking the hydrophobic bond formed by the hydrophobic groups on the tip and the monolayer requires enough force to be accumulated to completely rehydrate the monlayer. This effect is similar to the force that is required to remove

sticky tape, in which one must reintroduce air between the tape and the surface to which it was attached. For the chlorine-headed monolayer, there is little change in force as the displacement is increased by four Ångstroms from the point where the force is minimal. Once the threshold is reached then the force returns abruptly to zero, over a distance of about one Ångstrom. For the OH-headed monolayer, the threshold is delayed by another two Ångstroms, indicating the additional effect of the dehydronic force.

The estimation of the dehydronic force is complicated by the fact that one must estimate the number of dehydrons that will be dehydrated by the hydrophobic groups on the tip. But the geometry of AFM tips is well characterized, and the resulting estimate [124] of

$$|F| = 5.9 \pm 1.2 \text{pN} \tag{9.10}$$

at a distance of 5Å is in close agreement with the estimate (9.9) of 7.78 ± 1.5pN at a distance of 6Å in [141]. Part of the discrepancy could be explained by the fact that in [141] no energy of binding was attributed to the attachment to areas of a protein lacking dehydrons. If there were such an energy decrease, due e.g. to the formation of intermolecular interactions, the estimate of the force obtained in [141] would be reduced.

9.4 Membrane morphology

Since dehydrons have an attractive force that causes them to bind to a membrane, then the equal and opposite force must pull on the membrane. Since membranes are flexible, then this will cause the membrane to deform.

The possibility of significant morphological effect of dehydrons on membranes was suggested by the diversity of morphologies [368] of the inner membranes of cellular or subcellular compartments containing soluble proteins [143]. These vary from simple bag-like membranes [113] (e.g., erythrocytes, a.k.a. red blood cells) to highly invaginated membranes [407] (e.g., mitochondrial inner membranes). This raises the question of what might be causing the difference in membrane structure [235, 262, 303, 410].

Some evidence [143] suggests that dehydrons might play a role: hemoglobin subunits (which comprise the bulk of erythrocyte contents) are generally well wrapped, whereas two mitochondrial proteins, cytochrome c and pyruvate dehydrogenase, are less well wrapped. The correlation between the wrapping difference and the morphology difference provided motivation to measure the effect experimentally [143].

9.4.1 Protein adsorption

Morphology induction was tested in fluid phospholipid (DLPC) bilayers (Section 9.1) coating an optical waveguide [143]. The density of bilayer invaginations was measured by a technology called evanescent field spectroscopy which allowed measurement of both the thickness and refractive index of the adlayer [345, 387]. DLPC was added as needed for membrane expansion, with the portion remaining attached to the waveguide serving as a nucleus for further bilayer formation. Stable invaginations in the lipid bilayer formed after 60-hour incubation at T=318K.

9.4.2 Density of invaginations

The density of invaginations correlates with the extent of wrapping, ρ, of the soluble protein structure (Fig. 1, 2a in [143]). Greater surface area increase corresponds with lack of wrapping of backbone hydrogen bonds. The density of invaginations as a function of concentration (Figure 2b in [143]) shows that protein aggregation is a competing effect in the protection of solvent-exposed hydrogen bonds ([140, 128, 129, 121, 148]): for each protein there appears to be a concentration limit beyond which aggregation becomes more dominant.

9.5 Kinetic model of morphology

The kinetics of morphology development suggest a simple morphological instability similar to the development of moguls on a steep ski run. When proteins attach to the surface, there is a force that binds the protein to the surface. This force pulls upward on the surface (and downward on the protein) and will increase the curvature in proportion to the local density of proteins adsorbed on the surface [129]. The rate of change of curvature $\frac{dg}{dt}$ is an increasing function of the force f:

$$\frac{dg}{dt} = \phi(f) \tag{9.11}$$

for some increasing function ϕ. Note that

$$\phi(0) = 0; \tag{9.12}$$

if there is no force, there will be no change. The function ϕ represents a material property of the surface.

The probability p of further attachment increases as a function of the curvature at that point since there is more area for attachment where the curvature is higher. That is, $p(g)$ is also an increasing function.

Of course, attachment also reduces surface area, but we assume this effect is small initially. However, as attachment grows, this neglected term leads to a 'saturation' effect. There is a point at which further reduction of surface area becomes the dominating effect, quenching further growth in curvature. But for the moment, we want to capture the initial growth of curvature in a simple model. We leave as Exercise 9.2 the development of a more complete model.

Assuming equilibrium is attained rapidly, we can assert that the force f is proportional to $p(g)$: $f = cp(g)$ at least up to some saturation limit, which we discuss subsequently. If we wish to be conservative, we can assert only that

$$f = \psi(p(g)) \tag{9.13}$$

with ψ increasing. In any case, we conclude that f may be regarded as an increasing function of the curvature g, say

$$f = F(g) := \phi(\psi(p(g))). \tag{9.14}$$

To normalize forces, we should have no force for a flat surface. That is, we should assume that $p(0) = 0$. This implies, together with the condition $\phi(0) = 0$, that

$$F(0) = 0. \tag{9.15}$$

The greater attachment that occurs locally causes the force to be higher there and thus the curvature to increase even more, creating an exponential runaway (Fig. 4 in [143]). The repeated interactions of these two reinforcing effects causes the curvature to increase in an autocatalytic manner until some other process forces it to stabilize.

The description above can be captured in a semiempirical differential equation for the curvature g at a fixed point on the bilayer. It takes the form

$$\frac{dg}{dt} = F(g), \tag{9.16}$$

where F is the function in (9.14) that quantifies the relationships between curvature, probability of attachment and local density of protein described in the previous paragraph. Abstractly, we know that F is increasing because it is the composition of increasing functions. Hence F has a positive slope s at $g = 0$. Moreover, it is plausible that $F(0) = 0$ using our assumptions made previously.

Thus the curvature should grow exponentially at first with rate s. In the initial stages of interface development, F may be linearly approximated by virtue of the mean value theorem, yielding the autocatalytic equation:

$$\frac{dg}{dt} = sg. \tag{9.17}$$

Figure 4 in [143] indicates that the number of invaginations appears to grow exponentially at first, and then saturates.

We have observed that there is a maximum amount of protein that can be utilized to cause morphology (Figure 2b in [143]) beyond which aggregation becomes a significantly competitive process. Thus, a 'crowding problem' at the surface causes the curvature to stop increasing once the number of adsorbed proteins gets too high at a location of high curvature.

9.6 Exercises

Exercise 9.1 *Determine the minimal distance between a hydrophobe and a backbone hydrogen bond in protein structures. That is, determine the number of wrappers as a function of the desolvation radius, and determine when, on average, this tends to zero.*

Exercise 9.2 *Derive a more refined model of morphological instability accounting for the reduction of surface area upon binding. Give properties of a function F as in (9.14) that incorporate the effect of decreasing surface area, and show how it would lead to a model like (9.16) which would saturate (rather than grow exponentially forever), reflecting the crowding effect of the molecules on the lipid surface.*

Exercise 9.3 *The logistic equation*

$$\frac{dg}{dt} = g(1 - g), \tag{9.18}$$

naturally includes both an exponential rise and a saturation effect. Show how this might be used to include a saturation effect in the model in Section 9.5.

Exercise 9.4 *Determine the distribution of distances between waters in PDB files and (a) the nearest nonpolar carbonaceous group and (b) the nearest of all other heavy atoms. Also refine your datamining to compute different distributions in (a) for the different nonpolar carbonaceous groups CH_n for different values of $n = 0, 1, 2, 3$. Similary, refine your datamining to compute different distributions in (b) for different atom classes (e.g., mainchain atoms versus ones in sidechains, and C versus N versus O).*

Chapter 10

Electrostatic force details

In Section 3.3, we introduced some basic electronic interactions. Here we look at electronic interactions in more detail. Our objective is to understand the expected configuration of interacting charged and polar side chains. We make the assumption that the minimum energy configuration will be informative. Since we cannot know what the global electrostatic environment will be in general, we use local electrostatic energy as the quantity to be minimized. The resulting configurations provide only a guide to what we might expect in practice, but we will see that there are some suprising results.

The basic electronic entities are groups of charges that are constrained to be together, such as dipoles. In Section 10.2.1 we study dipole-dipole interactions. In Section 10.3.1, we consider charge-dipole interactions such as arise in cation-pi pairs such as Arg-Tyr or Lys-Phe. We also consider like-charge repulsion such as occurs with Arg-His or Asp-Glu pairs in Section 10.3.2.

There is a natural hierarchy of charged groups. These can be ranked by the rate of decay of their potentials, and thus by how localized they are. At the highest (most global) level is the single charge, with a potential r^{-1}. The dipole is a combination of opposite charges at nearby locations, with a potential r^{-2}. The quadrupole is a collection of four charges arranged in appropriate positions with a potential r^{-3}. Some important entities, such as water, are often modeled as being four charges at positions with substantial symmetry, and it is important to know whether they constitute quadruples or just dipoles. This determines the global accumulation of charge and thus has significant implications as we now discuss. We subsequently return to the question of whether water is a dipole or quadrupole.

10.1 Global accumulation of electric force

The reason that we need to know the order of decay of the potential, or the associated force, for various types of charged groups is quite simple to explain. Suppose that we have a material made of an assembly of electrostatic entities, such as water. We would like to understand the locality of forces exerted by the entities on each other. In particular, are they local, or do global contributions have a significant effect?

To quantify this question, suppose we try to estimate the potential or force due to an organized
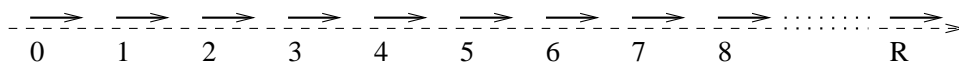
Figure 10.1: A one dimensional array of dipoles.

collection molecular groups each of whose effects at a distance r is proportional to r^{-n} for some n. In Figure 10.1 we depict a simple one-dimensional array of dipoles. Summing over all space, we determine the total force. In the one dimensional case, as depicted in Figure 10.1, the resulting force or potential at the origin is proportional to

$$\sum_{r=1}^{R} r^{-n} \leq C \quad \text{provided } n > 1. \tag{10.1}$$

The potential for dipoles has $n = 3$, and the result remains bounded as $R \to \infty$. On the other hand, if instead of dipoles at the points $1, 2, \ldots, R$ we had point charges of all the same sign, then the resulting potential as in (10.1) would diverge logarithmically.

In Figure 10.2, we depict a two-dimensional arrangement of dipoles. One each circle of radius r, we place r dipoles pointing toward the origin, at a spacing of 2π units along the circles. We can also imagine that instead of dipoles we place point charges or other entities. In the two dimensional case, there are r such entities all at a distance r from the origin, so the resulting potential is

$$\sum_{r=1}^{R} r\, r^{-n} \leq C \quad \text{provided } n > 2. \tag{10.2}$$

Figure 10.2: A two dimensional array of dipoles.

Again, the potential for dipoles has $n = 3$, and the result remains bounded as $R \to \infty$.

Now let us imagine a three-dimensional arrangement of dipoles arrayed on concentric spherical shells of radius r, all pointing towards the origin in space, that is, the center of the concentric spheres. We leave the depiction of this to the imagination. Placing r^2 points on each sphere of radius r is not trivial, similar to the problem of putting dimples on a golf ball. But as r gets larger, it becomes possible to arrangement them in an asymptotically symmetric arrangement.

We can estimate the resulting potential or force at the origin by computing sums over expanding spherical shell sets

$$\left\{ \mathbf{r} \in \mathbb{R}^3 \;\middle|\; r - 1 \leq |\mathbf{r}| < r \right\} \text{ for } r = 1, 2, 3, \ldots.$$

In each spherical shell region, the sum of all contributions, ignoring possible cancellations, would be approximately cr^{2-n} since all values of $\mathbf{r}$ in the set would be comparable to r, and there would be approximately cr^2 of them (assuming as we do that they are uniformly distributed). Then the total result would be proportional to

$$\sum_{r=1}^{R} r^{2-n} \tag{10.3}$$

charge groups	force name	power law	equation
charge-charge	salt bridge	r^{-1}	(3.1)
charge-dipole	Keesom force	$\cos\theta\, r^{-2}$	(10.28)
dipole-dipole	polar bond	r^{-3}	(3.11),(10.23)
dipole-quadrupole	van der Waals	r^{-4}	(3.37)
induced dipole	London-Debye	r^{-6}	(3.30)

Table 10.1: Different power law behaviors for various forces, together with common names frequently used for them. The 'equation' column indicates where they can be found in the text.

which is divergent (as R increases) for $n \leq 3$.

Note that R is the size of a macroscopic system in microscopic units, so it is related to Avogadro's number, hence R should be viewed as nearly infinite. The borderline case $n = 3$, for which the divergence is only logarithmic, corresponds to the dipole potential and thus the electric force in the charge-dipole interaction. For the dipole-dipole interaction, $n = 4$, the first exponent where the forces can be said to be local, but the convergence rate is rather slow: $\mathcal{O}(1/R_{\mathrm{cut}})$ if we take R_{cut} to be a cut-off radius beyond which we ignore external effects. This explains to some extent why molecular dynamics simulations have to expend so much computational effort to compute electrostatic interactions in order to represent the forces accurately.

When the electrostatic force is proportional to r^{-n} for some n, the electrostatic potential is proportional to r^{1-n}. Thus the electrostatic potential exhibits a logarithmic divergence for a system of dipoles if no further organization (i.e., formation of quadrupoles by groups of dipoles) obtains. This would imply that dipoles *must* form such structures in larger aggregates, or otherwise infinite potentials would arise. However, we do not attempt a more detailed analysis of such systems here.

We list in Table 10.1 several different forces that we encounter together with the different power laws associated with them.

10.2 Dipole-dipole interactions

We have seen that certain bonds can be modeled by simple interactions between charge groups. For example, polar groups can be modeled simply by placing partial charges appropriately at atom centers, as described in Section 8.2.2. Typically, these groups can be represented as dipoles (Section 3.3). Here we investigate in detail the angular dependence of some of these models. In Section 3.5.3, we saw that dipoles can interact even with neutrally charged groups; this is one manifestation of the van der Waals force. However, we will see that this can really be rationalized as a quadrupole-dipole interaction.

In this section, we will restrict most of our attention to charge groups in two dimensions. In some cases, we will explicitly include the third coordinate $z = 0$ in the representations (see Figure 3.12), but in others we will simply omit it (see Figure 10.3).

10.2.1 Dipole-dipole configurations

Let us consider the effect of angular orientation on the strength of interaction of two dipoles. Since the possible set of configurations has a high dimension, we break down into special cases.

Figure 10.3: Dipole-dipole interaction: rotated in-line configuration. The positive (resp., negative) charge center of the dipole on the right is at $(0,0)$ (resp., at $(\cos\phi, \sin\phi)$). The positive (resp., negative) charge center of the dipole on the left is at $(-R-1, 0)$ (resp., at $(-R, 0)$).

In-line interaction configuration

Suppose we have two dipoles as indicated in Figure 10.3. The exact positions of the charges are as follows. The position of the positive charge on the right we take as the origin, and we assume the separation distance of the charges is one. The separation between the positive charge on the right and the negative charge on the left is R. Thus the charge centers of the dipole on the left are at $(-R-1, 0)$ (positive charge) and $(-R, 0)$ (negative charge). The negative charge on the right is at $(\cos\phi, \sin\phi)$.

The distances between the various charges are easy to compute. The distance between the negative charge on the left and the positive charge on the right is R, and the distance between the two positive charges is $R+1$. The distance between the two negative charges is

$$|(\cos\phi, \sin\phi) - (-R, 0)| = \sqrt{(R + \cos\phi)^2 + \sin^2\phi}$$
$$= \sqrt{1 + R^2 + 2R\cos\phi}, \tag{10.4}$$

and the distance between the positive charge on the left and the negative charge on the right is

$$|(\cos\phi, \sin\phi) - (-R-1, 0)| = \sqrt{(1 + R + \cos\phi)^2 + \sin^2\phi}$$
$$= \sqrt{1 + (R+1)^2 + 2(R+1)\cos\phi}. \tag{10.5}$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$\frac{1}{R+1} - \frac{1}{R} + \frac{1}{\sqrt{1 + R^2 + 2R\cos\phi}} - \frac{1}{\sqrt{1 + (R+1)^2 + 2(R+1)\cos\phi}}. \tag{10.6}$$

A plot of the interaction energy (10.6) is given in Figure 10.4 as a function of ϕ for various values of R. Since we know (cf. (3.11)) that the interaction energy will decay like R^{-3}, we have scaled the energy in Figure 10.4 by R^3 to keep the plots on the same scale. The value of $R = 1000$ indicates the asymptotic behavior; see Exercise 10.1 for the analytical expression of the asymptotic limit. Indeed, there is little difference between $R = 100$ (not shown) and $R = 1000$. The flatter curve is the smallest value of R (=2) and shows only limited angular dependence. Thus

<p style="text-align:center">modeling a hydrogen bond using a simple dipole-dipole
interaction does not yield a very strong angular dependence.</p>

Figure 10.4: Dipole-dipole (in-line) interaction energy (vertical axis), scaled by R^3, for $R = 2, 4, 10, 1000$. The energies correspond to the dipole-dipole configuration in Figure 10.3. Horizontal ϕ-axis measured in radians. The flattest curve corresponds to $R = 2$.

Parallel interaction configuration

Let us consider the effect of a different angular orientation on the strength of interaction of two dipoles. Suppose we have two dipoles as indicated in Figure 10.5. Here the dipoles stay parallel, but the one on the right is displaced by an angle ϕ from the axis through the dipole on the left. The exact positions of the charges are as follows.

The position of the negative charge on the left we take as the origin, and we assume the separation distance of the charges is one. The separation between the positive charge on the right and the negative charge on the left is R. Thus the charge centers of the dipole on the right are at $R(\cos\phi, \sin\phi)$ (positive charge) and $(1 + R\cos\phi, R\sin\phi)$ (negative charge).

The distance between the positive charges is the same as the distance between the negative charges because the dipoles are parallel:

$$|(1 + R\cos\phi, R\sin\phi)| = \sqrt{1 + R^2 + 2R\cos\phi}. \tag{10.7}$$

Figure 10.5: Dipole-dipole interaction: rotated parallel configuration. The negative (resp., positive) charge center of the dipole on the left is at $(0, 0)$ (resp., at $(-1, 0)$). The positive (resp., negative) charge center of the dipole on the right is at $R(\cos\phi, \sin\phi)$ (resp., at $(1 + R\cos\phi, \sin\phi)$).

Similarly, the distance between the positive charge on the left and the negative charge on the right is

$$|(2 + R\cos\phi, R\sin\phi)| = \sqrt{4 + R^2 + 4R\cos\phi}. \tag{10.8}$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$-\frac{1}{R} + \frac{2}{\sqrt{1 + R^2 + 2R\cos\phi}} - \frac{1}{\sqrt{4 + R^2 + 4R\cos\phi}}. \tag{10.9}$$
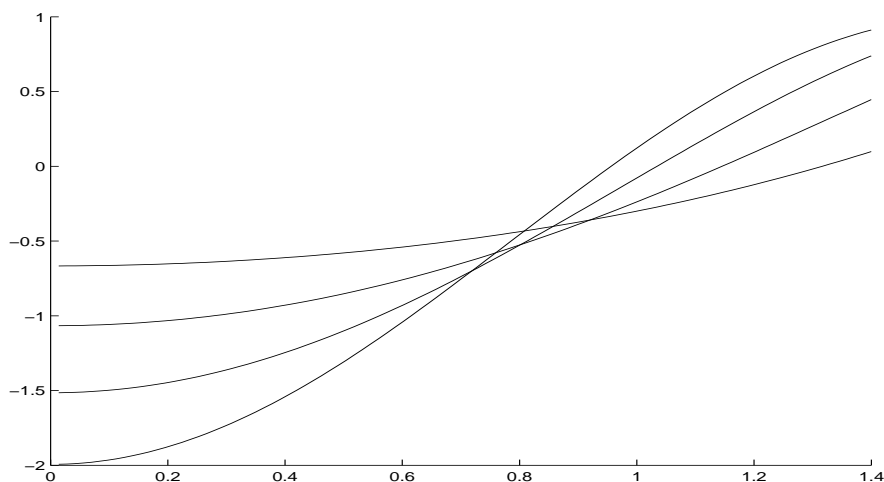
Figure 10.6: Dipole-dipole (parallel) interaction energy (vertical axis), scaled by R^3, for $R = 2, 4, 10, 1000$. The energies correspond to the dipole-dipole configuration in Figure 10.5. Horizontal ϕ-axis measured in radians.

A plot of the interaction energy (10.9) is given in Figure 10.6 as a function of ϕ for various values of R. Since we know (cf. (3.11)) that the interaction energy will decay like R^{-3}, we have scaled the energy in Figure 10.4 by R^3 to keep the plots on the same scale. The value of $R = 1000$ indicates the asymptotic behavior; see Exercise 10.2 for the analytical expression. Again, there is little difference between $R = 100$ (not shown) and $R = 1000$. The flatter curve is the smallest value of R (=2) and shows only limited angular dependence.

10.2.2   Two-parameter interaction configuration

Now we consider the effect of a dual angular orientation on the strength of interaction of two dipoles. We do this as a first step to understanding the problem of hydrogen placement for serine, cf. Section 6.2. It also has direct bearing on the energy associated with different orientations of the charge groups (e.g., amide and carbonyl) in hydrogen bonds, cf. Section 6.3.

Suppose we have two dipoles as indicated in Figure 10.7. The exact positions of the charges are as follows. The position of the negative charge on the left we take as the origin, and we assume the separation distance of the charges is one. The separation between the positive charge on the right and the negative charge on the left is R. Thus the charge centers of the dipole on the right are at $R(\cos\theta, \sin\theta)$ (positive charge) and $R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi)$ (negative charge).

The distance between the negative charge on the left and the positive charge on the right is R, and the separation between the positive charge on the right and the positive charge on the left is

$$|(1 + R\cos\theta, R\sin\theta)| = \sqrt{1 + R^2 + 2R\cos\theta}. \tag{10.10}$$
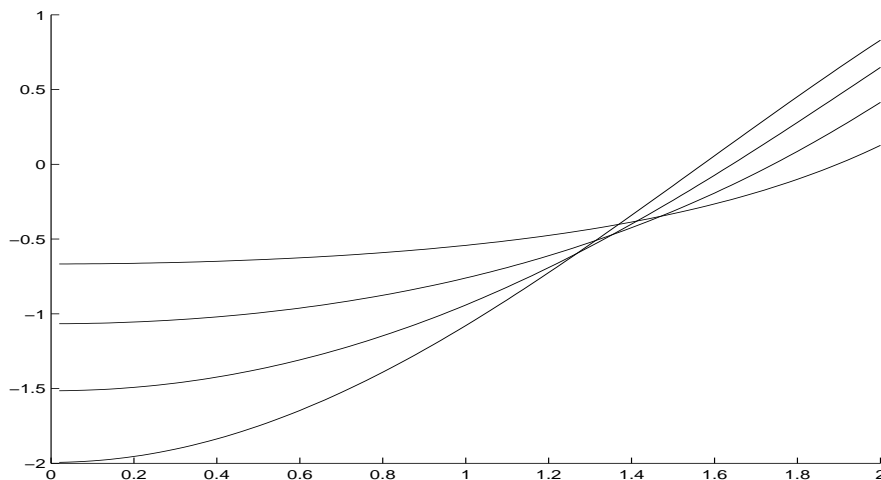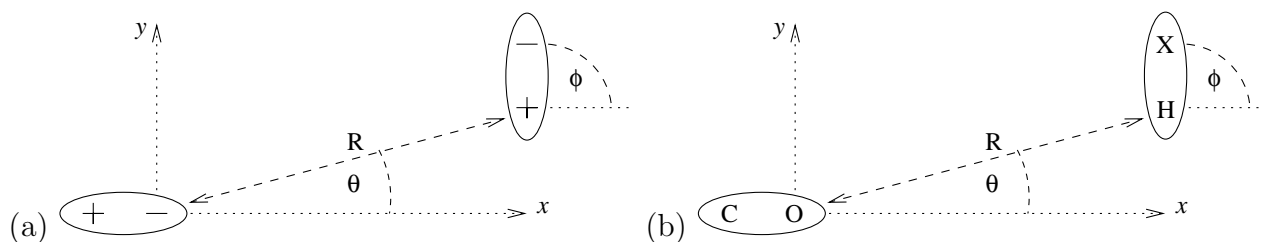
Figure 10.7: Dipole-dipole (two-angle) interaction configuration. (a) The abstract case. The negative (resp., positive) charge center of the dipole on the left is at $(0,0)$ (resp., at $(-1,0)$). The positive (resp., negative) charge center of the dipole on the right is at $R(\cos\theta, \sin\theta)$ (resp., at $R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi)$). (b) Examples: in the serine-hydrogen placement problem, the atom X is oxygen (O), and for a carbonyl-amide backbone hydrogen bond, the atom X is nitrogen (N).

The separation between the positive charge on the right and the negative charge on the left is

$$
\begin{aligned}
|R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi)| &= \sqrt{(R\cos\theta + \cos\phi)^2 + (R\sin\theta + \sin\phi)^2} \\
&= \sqrt{(R^2 + 1 + 2R(\cos\theta\cos\phi + \sin\theta\sin\phi)}.
\end{aligned}
\tag{10.11}
$$

Finally, the distance (squared) between the positive charge on the left and the negative charge on the right is

$$
\begin{aligned}
|R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi) - (-1,0)|^2 &= |(1 + R\cos\theta + \cos\phi, R\sin\theta + \sin\phi)|^2 \\
&= (1 + \cos\phi)^2 + 2R\cos\theta(1 + \cos\phi) + R^2 + 2R\sin\theta\sin\phi + \sin^2\phi \\
&= 2(1 + \cos\phi) + 2R\cos\theta(1 + \cos\phi) + R^2 + 2R\sin\theta\sin\phi \\
&= 2(1 + \cos\phi)(1 + R\cos\theta) + R^2 + 2R\sin\theta\sin\phi.
\end{aligned}
\tag{10.12}
$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$
\begin{aligned}
V(R) := & -\frac{1}{R} + \frac{1}{\sqrt{1 + R^2 + 2R\cos\theta}} + \frac{1}{\sqrt{1 + R^2 + 2R(\cos\theta\cos\phi + \sin\theta\sin\phi)}} \\
& -\frac{1}{\sqrt{R^2 + 2(1 + \cos\phi)(1 + R\cos\theta) + 2R\sin\theta\sin\phi}}.
\end{aligned}
\tag{10.13}
$$

Minimum energy configuration

Since there are now two angles to vary, it is not so clear how to display the energy in a useful way. But one question we may ask is: what is the minimum energy configuration if we allow ϕ to vary for a given θ? We might think that the dipole on the right would always point at the negative charge at the left. This would correspond to having the minimum energy configuration at $\phi = \theta$. This is clearly true at $\theta = 0$, but say at $\theta = \pi/2$, we might expect the minimum energy configuration to occur when the dipole on the right is flipped, that is at $\phi = \pi = 2\theta$. We plot the energy as a function of ϕ for various values of θ in Figure 10.8. As an aid to the eye, we plot a circle at a
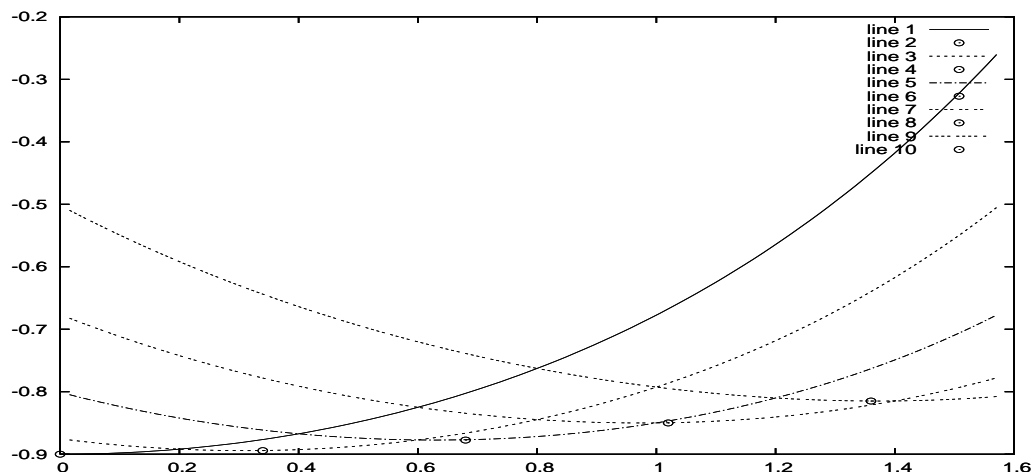
Figure 10.8: Dipole-dipole (two-angle) interaction energy (vertical axis), scaled by R^3, for $R = 3$, as a function of ϕ for various fixed values of $\theta = 0, 0.2, 0.4, 0.6, 0.8$. Approximate minimum values of the energy are indicated by circles at the points $\phi = 1.7\,\theta$. The horizontal ϕ-axis is measured in radians.

point close the minimum in energy, as a way to see how the optimum ϕ varies as a function of θ. In particular, we have plotted the point not at $\phi = \theta$, nor at $\phi = 2\theta$, but rather $\phi = 1.7\,\theta$. This is convincing evidence that the relationship between the optimum value of ϕ for a fixed value of θ is complicated.

In the case that $\theta = \pi/2$, the expression (10.13) simplifies to

$$-\frac{1}{R} + \frac{1}{\sqrt{1 + R^2}} - \frac{1}{\sqrt{R^2 + 2(1 + \cos\phi) + 2R\sin\phi}} + \frac{1}{\sqrt{1 + R^2 + 2R\sin\phi}}. \tag{10.14}$$

Then if $\phi = \pi$, this further simplifies to $-2R^{-1} + 2(1+R^2)^{-1/2}$ as we would expect. However, the minimum of the expression (10.14) does not occur at $\phi = \pi$, due to the asymmetry of the expression around this value. We leave as Exercise 10.4 to plot (10.14) as a function of ϕ for various values of R to see the behavior.

When R is large, we might expect that $\phi_{\text{opt}} \approx \theta$, since the dipole should point in the general direction of the other dipole. However, this is not the case; rather there is a limiting behavior that is different. In Figure 10.9, the optimal ϕ is plotted as a function of θ, and we note that it is very nearly equal to 2θ, but not exactly. For θ small, it behaves more nearly like $\phi \approx 1.7\,\theta$, but for larger values of θ the optimal ϕ increases to, and then exceeds, 2θ, before returning to the value of 2θ near $\theta = \pi$.

The minimum ϕ has been determined by computing the energies for discrete values of ϕ and then interpolating the data by a quadratic around the discrete minimum. Necessary adjustments at the ends of the computational domain are evident. Limited resolution in the computations contributes to the visible jaggedness of the curves in the plot. We leave as an exercise to produce smoother plots, as well as to explore the asymptotic behavior as $R \to \infty$.
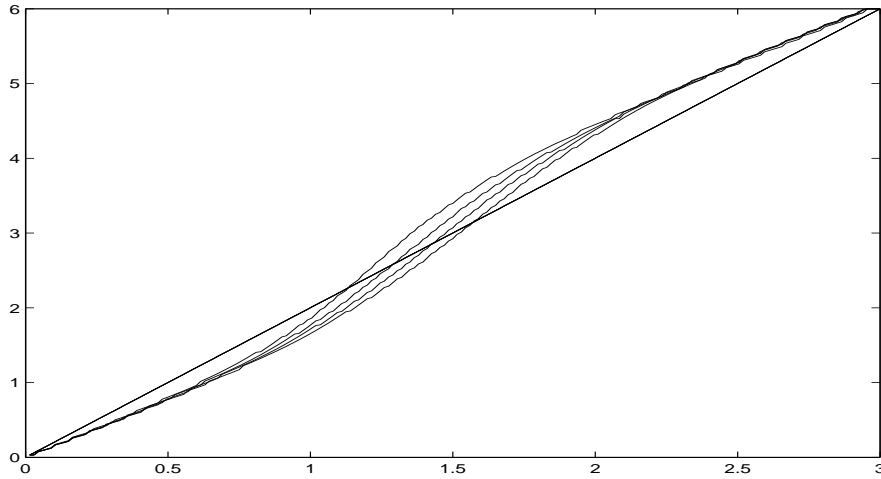
Figure 10.9: Optimal ϕ angle (vertical axis) corresponding to the minimum interaction energy as a function of θ (horizontal axis) for the dipole-dipole (two-angle) interaction, for $R = 3, 5, 10, 1000$ (the left-most curve corresponds to $R = 3$, and they move to the right with increasing R). Both the horizontal θ-axis and the vertical ϕ-axis are measured in radians. The line $\phi = 2\theta$ has been added as a guide.

The energy, again scaled by R^3, at the optimal value of ϕ is plotted as a function of θ in Figure 10.10. Since the curves in this figure are not horizontal, the dipole system has a torque that would tend to move them to the $\theta = 0$ position if θ were not fixed (as we assume it is, due to some external geometric constraint).

Limiting expression

To examine the limiting behavior of the energy minimum in detail, we use a general formula valid in three dimensions. We return to the two-angle case in Section 10.2.4.

10.2.3   Three dimensional interactions

We now consider the interaction between two dipoles in arbitrary orientation. For simplicity, we place the center of the two dipoles along the x-axis, one at the origin and the other at $\mathbf{r} = (r, 0, 0)$. Thus assume that there is are charges ± 1 at $\pm \mathbf{u}$, and correspondingly charges ± 1 at points $\mathbf{r} \pm \mathbf{v}$ where $\mathbf{u}$ and $\mathbf{v}$ are unit vectors, that is, $u_1^2 + u_2^2 + u_3^2 = v_1^2 + v_2^2 + v_3^2 = 1$. The interaction potential is then

$$V(r, \mathbf{u}, \mathbf{v}) = \frac{1}{|\mathbf{u} - \mathbf{r} - \mathbf{v}|} + \frac{1}{|-\mathbf{u} - \mathbf{r} + \mathbf{v}|} - \frac{1}{|\mathbf{u} - \mathbf{r} + \mathbf{v}|} - \frac{1}{|-\mathbf{u} - \mathbf{r} - \mathbf{v}|}. \tag{10.15}$$

We use the relation

$$|\mathbf{w} + \mathbf{x} + \mathbf{y}|^2 = |\mathbf{w}|^2 + |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2(\mathbf{w} \cdot \mathbf{x} + \mathbf{w} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{x}) \tag{10.16}$$
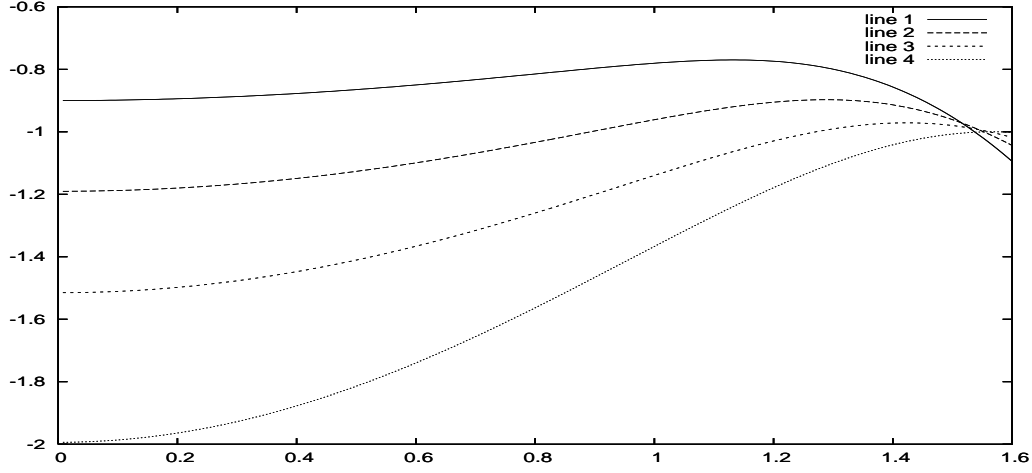
Figure 10.10: Dipole-dipole (two-angle) interaction energy minimum, scaled by R^3, for $R = 3, 5, 10, 1000$ (top to bottom), as a function of θ. Plotted is the energy (vertical axis) at the optimal value of ϕ that minimizes the energy as a function of ϕ for fixed θ. Horizontal θ-axis measured in radians.

to get

$$V(r, \mathbf{u}, \mathbf{v}) = \frac{1}{\sqrt{2 + r^2 - 2\mathbf{u} \cdot \mathbf{v} - 2r(u_1 - v_1)}} + \frac{1}{\sqrt{2 + r^2 - 2\mathbf{u} \cdot \mathbf{v} + 2r(u_1 - v_1)}}$$
$$- \frac{1}{\sqrt{2 + r^2 + 2\mathbf{u} \cdot \mathbf{v} - 2r(u_1 + v_1)}} - \frac{1}{\sqrt{2 + r^2 + 2\mathbf{u} \cdot \mathbf{v} + 2r(u_1 + v_1)}} \qquad (10.17)$$
$$= \frac{1}{\sqrt{2 + r^2}} \left(\frac{1}{\sqrt{1 - a - b}} + \frac{1}{\sqrt{1 - a + b}} - \frac{1}{\sqrt{1 + a - c}} - \frac{1}{\sqrt{1 + a + c}} \right)$$

where

$$a = \frac{2\mathbf{u} \cdot \mathbf{v}}{2 + r^2}, \quad b = \frac{2r(u_1 - v_1)}{2 + r^2} = \frac{2\mathbf{r} \cdot (\mathbf{u} - \mathbf{v})}{2 + r^2}, \quad \text{and} \quad c = \frac{2r(u_1 + v_1)}{2 + r^2} = \frac{2\mathbf{r} \cdot (\mathbf{u} + \mathbf{v})}{2 + r^2}. \qquad (10.18)$$

Expanding using the expression

$$\boxed{(1 - \epsilon)^{-1/2} = 1 + \tfrac{1}{2}\epsilon + \tfrac{3}{8}\epsilon^2 + \tfrac{5}{16}\epsilon^3 + \mathcal{O}\left(\epsilon^4\right),} \qquad (10.19)$$

we find for large r that

$$V(r, \mathbf{u}, \mathbf{v})\sqrt{2 + r^2} = \frac{1}{\sqrt{1 - a - b}} + \frac{1}{\sqrt{1 - a + b}} - \frac{1}{\sqrt{1 + a - c}} - \frac{1}{\sqrt{1 + a + c}}$$
$$\approx \tfrac{1}{2} \left(a + b + a - b + a - c + a + c\right)$$
$$+ \tfrac{3}{8} \left((a + b)^2 + (a - b)^2 - (a - c)^2 - (a + c)^2\right)$$
$$+ \tfrac{5}{16} \left((a + b)^3 + (a - b)^3 + (a - c)^3 + (a + c)^3\right) + \mathcal{O}\left(r^{-4}\right) \qquad (10.20)$$
$$= 2a + \tfrac{3}{4} \left(b^2 - c^2\right) + \tfrac{5}{16} \left(4a^3 + 2a(b^2 + c^2)\right) + \mathcal{O}\left(r^{-4}\right)$$
$$= 2a + \tfrac{3}{4} \left(b^2 - c^2\right) + \mathcal{O}\left(r^{-4}\right)$$

because $a = \mathcal{O}\left(r^{-2}\right)$, $b = \mathcal{O}\left(r^{-1}\right)$, and $c = \mathcal{O}\left(r^{-1}\right)$. But

$$b^2 - c^2 = \frac{4r^2}{(2+r^2)^2}\left((u_1 - v_1)^2 - (u_1 + v_1)^2\right) = \frac{-16u_1 v_1 r^2}{(2+r^2)^2} = \frac{-16(\mathbf{u}\cdot\mathbf{r})(\mathbf{v}\cdot\mathbf{r})}{(2+r^2)^2} \tag{10.21}$$

Therefore

$$V(r,\mathbf{u},\mathbf{v})\sqrt{2+r^2} = \frac{4\mathbf{u}\cdot\mathbf{v}}{2+r^2} - \frac{12(\mathbf{r}\cdot\mathbf{u})(\mathbf{r}\cdot\mathbf{v})}{(2+r^2)^2} + \mathcal{O}\left(r^{-4}\right). \tag{10.22}$$

Thus

$$\boxed{V(r,\mathbf{u},\mathbf{v}) \approx \frac{4\mathbf{u}\cdot\mathbf{v} - 12(\tilde{\mathbf{r}}\cdot\mathbf{u})(\tilde{\mathbf{r}}\cdot\mathbf{v})}{r^3},} \tag{10.23}$$

where $\tilde{\mathbf{r}} = (1/r)\mathbf{r}$ is the unit vector in the direction of $\mathbf{r}$.

10.2.4   Application to two-angle problem

As an application, we consider the limiting behavior of the two-angle problem considered in Section 10.2.2. We modify coordinates slightly as indicated in Figure 10.11.

Thus we take $\mathbf{u} = (1,0)$, $\mathbf{v} = (\cos\phi, \sin\phi)$, and $\tilde{\mathbf{r}} = (\cos\theta, \sin\theta)$. Therefore

$$\begin{aligned}
r^3 V(r,\phi,\theta) \approx U(\phi,\theta) &:= 4\cos\phi - 12\cos\theta(\cos\phi\,\cos\theta + \sin\phi\,\sin\theta)\\
&= 4\cos\phi(1 - 3\cos^2\theta) - 12\cos\theta\sin\phi\,\sin\theta \qquad (10.24)\\
&= 4\cos\phi(1 - 3\cos^2\theta) - 6\sin\phi\,\sin 2\theta.
\end{aligned}$$

We are interesting in finding, for each θ, the value of ϕ that minimizes $U(\phi,\theta)$. Thus we compute

$$\frac{\partial U}{\partial \phi}(\phi,\theta) = -4\sin\phi(1 - 3\cos^2\theta) - 6\cos\phi\sin 2\theta. \tag{10.25}$$

Therefore, at a minimum, we have

$$\tan\phi = \frac{\sin 2\theta}{2\cos^2\theta - \frac{2}{3}}, \tag{10.26}$$

from which we get $\phi(\theta)$. Expanding for small θ, we find $\phi = \frac{3}{2}\theta + \mathcal{O}\left(\theta^2\right)$. It also follows that when $\theta = \pi/2$, then $\phi = \pi$. We leave to Exercise 10.17 further investigation of this curious function.

Figure 10.11: Dipole-dipole (two-angle) interaction configuration. The $\mp$ charge centers of the dipole on the left are at $(\pm 1, 0)$. The positive (resp., negative) charge center of the dipole on the right is at $R(\cos\theta, \sin\theta)$ (resp., at $R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi)$).

10.3   Charged interactions

Interactions involving charged groups may seem conceptually simpler than those involving only dipoles. The interaction between two charges is given in (3.1); it depends only on the distance between them. We will see that the interaction between a single charge and a dipole is not much more complex. But there are more complex interactions between charged groups that require careful analysis. These have direct bearing on the expected orientation of neighboring charged sidechains (e.g., Asp-Glu) in particular.

10.3.1   Charge-dipole interactions

Charge-dipole interactions are simpler to analyze, and we have already anticipated their asymptotic strength in (3.8). On the other hand, this forms a very important class of interactions. Although mainchain-mainchain (hydrogen bond) interactions do not involve such pairs, all of the three other interactions among sidechains and mainchains can involve charge-dipole interactions. In addition, more complex interactions, such as cation-π interactions (Section 12.2) are of this form. Thus we develop the basics of charge-dipole interactions in some detail.

Figure 10.12: Charge-dipole interaction configuration. The dipole on the left has its negative charge at the origin and its positive charge center at $(-1, 0, 0)$. The positive charge $(+a)$ center on the right is at $R(\cos\theta, \sin\theta, 0)$.

By choosing coordinates appropriately, we can assume that the positive and negative sites of the dipole align on the x-axis, and that the charge is located in the x, y plane, as depicted in Figure 10.12. Assume that the negative charge of the dipole is at the origin and that the isolated charge is positive, located at $r(\cos\theta, \sin\theta, 0)$. We choose scales such that the charges of the dipole are of unit size (± 1), located at $(\mp 1, 0, 0)$, that is, the positive charge of the dipole is at $(-1, 0, 0)$. If a is the charge of the isolated charge, then the interaction energy of the system is

$$V(r, \theta) = -\frac{a}{\sqrt{(1 - r\cos\theta)^2 + r^2\sin^2\theta}} + \frac{a}{\sqrt{(1 + r\cos\theta)^2 + r^2\sin^2\theta}}. \tag{10.27}$$

We leave as Exercise 10.8 to show that

$$V(r, \theta) \approx -\frac{2a\cos\theta}{r^2} \tag{10.28}$$

for large r and fixed θ.

We will be interested in the force field that the dipole exerts on the charge as well. It is easier to compute the gradient of V in Cartesian coordinates (note that we can ignore the z direction in our computations):

$$
\begin{aligned}
V(x, y) &= -\frac{a}{\sqrt{(1-x)^2 + y^2}} + \frac{a}{\sqrt{(1+x)^2 + y^2}} \\
&= a\frac{\sqrt{(1-x)^2 + y^2} - \sqrt{(1+x)^2 + y^2}}{\sqrt{(1-x)^2 + y^2}\,\sqrt{(1+x)^2 + y^2}} \\
&= a\frac{\sqrt{1 - 2x + r^2} - \sqrt{1 + 2x + r^2}}{\sqrt{1 - 2x + r^2}\,\sqrt{1 + 2x + r^2}} \\
&= \frac{a}{r}\frac{\sqrt{(1-2x)r^{-2} + 1} - \sqrt{(1+2x)r^{-2} + 1}}{\sqrt{(1-2x)r^{-2} + 1}\,\sqrt{(1+2x)r^{-2} + 1}} \\
&\approx -\frac{2a\cos\theta}{r^2}
\end{aligned}
\tag{10.29}
$$

Figure 10.13: Charge-charge interaction configuration similar to what is found in an interaction between Asp and Glu. The pair of negative charges on the right is at $\pm(\cos\phi, \sin\phi)$ and the pair of negative charges on the left is at $(R \pm 1, 0)$.

for large $r = \sqrt{x^2 + y^2}$.

To improve readability, we will use the notation $[x, y]$ to denote the vector with components x and y. Thus we find

$$
\begin{aligned}
\nabla V(x, y) &= a((1-x)^2 + y^2)^{-3/2}[x-1, y] - a((1+x)^2 + y^2)^{-3/2}[x+1, y] \\
&= a\frac{(1+2x+r^2)^{3/2} - (1-2x+r^2)^{3/2}}{(1-2x+r^2)^{3/2}(1+2x+r^2)^{3/2}}[x, y] \\
&\quad - a\frac{(1+2x+r^2)^{3/2} + (1-2x+r^2)^{3/2}}{(1-2x+r^2)^{3/2}(1+2x+r^2)^{3/2}}[1, 0] \\
&= \frac{a}{r^2}\frac{(r^{-2} + 2r^{-1}\cos\theta + 1)^{3/2} - (r^{-2} - 2r^{-1}\cos\theta + 1)^{3/2}}{(r^{-2} - 2r^{-1}\cos\theta + 1)^{3/2}(r^{-2} + 2r^{-1}\cos\theta + 1)^{3/2}}[\cos\theta, \sin\theta] \\
&\quad - \frac{a}{r^3}\frac{(r^{-2} + 2r^{-1}\cos\theta + 1)^{3/2} + (r^{-2} - 2r^{-1}\cos\theta + 1)^{3/2}}{(r^{-2} - 2r^{-1}\cos\theta + 1)^{3/2}(r^{-2} + 2r^{-1}\cos\theta + 1)^{3/2}}[1, 0] \\
&\approx \frac{6a\cos\theta}{r^3}[\cos\theta, \sin\theta] - \frac{2a}{r^3}[1, 0]
\end{aligned}
\tag{10.30}
$$

for large r, where we used $(1 + \epsilon)^{3/2} \approx 1 + \frac{3}{2}\epsilon$. Finally, we recall that these calculations are fully valid in three dimensions, so we have derived expressions valid for all z as well. In all cases, the z-component of ∇V is zero.

10.3.2   Charge-charge interactions

We now consider the preferred angular orientation for two like charged groups as one finds in residues such as Asp and Glu. Suppose we have two charge groups as indicated in Figure 10.13. The exact positions of the charges are as follows. We assume the separation distance of the charges is two, and we assume that the origin is the center of the two negative charges on the right. Thus there are negative charges at $(\cos\phi, \sin\phi)$ and $(-\cos\phi, -\sin\phi)$. The separation between the charge groups is R; the negative charges on the left are fixed at $(-R \pm 1, 0)$. Thus the interaction energy for the

charged pairs (assuming unit charges) depend on the distances

$$
\begin{aligned}
r_{++} &= |+(\cos\phi, \sin\phi) - (R+1, 0)|, \\
r_{-+} &= |-(\cos\phi, \sin\phi) - (R+1, 0)|, \\
r_{+-} &= |+(\cos\phi, \sin\phi) - (R-1, 0)|, \quad \text{and} \\
r_{--} &= |-(\cos\phi, \sin\phi) - (R-1, 0)|.
\end{aligned}
\tag{10.31}
$$

With the denotations $C = \cos\phi, S = \sin\phi$, we find

$$
\begin{aligned}
r_{++}^2 &= (C - R - 1)^2 + S^2 = 2 + R^2 - 2\cos\phi(R+1) + 2R, \\
r_{-+}^2 &= (C + R + 1)^2 + S^2 = 2 + R^2 + 2\cos\phi(R+1) + 2R, \\
r_{+-}^2 &= (C - R + 1)^2 + S^2 = 2 + R^2 - 2\cos\phi(R-1) - 2R, \quad \text{and} \\
r_{--}^2 &= (C + R - 1)^2 + S^2 = 2 + R^2 + 2\cos\phi(R-1) - 2R.
\end{aligned}
\tag{10.32}
$$

We can write these expressions succinctly as

$$
r_{\pm_1 \pm_2} = \sqrt{2 + R^2 \pm_1 2(\cos\phi(R \pm_2 1)) \pm_2 2R}.
\tag{10.33}
$$

The energy (of repulsion) for the charge groups is

$$
\frac{1}{r_{++}} + \frac{1}{r_{-+}} + \frac{1}{r_{+-}} + \frac{1}{r_{--}},
\tag{10.34}
$$

and we seek to find the value of ϕ that minimizes it. We leave as Exercise 10.5 to plot the expression in (10.34) which is symmetric around $\phi = \pi/2$ and has a simple minimum there.

A more realistic model of the charge group for Asp and Glu is depicted in Figure 10.14. The configuration is now three-dimensional, with the carbon joining the oxygens below the plane. We assume a positive charge on the left at $(-R, 0, -z)$ and on the right at $(0, 0, -z)$. The negative charges are at $(\cos\phi, \sin\phi, 0)$, $(-\cos\phi, -\sin\phi, 0)$, and $(R \pm 1, 0, 0)$. We leave as Exercise 10.6 to investigate the minimum energy configuration.

10.4   General form of a charge group

We now develop some technology to allow us to analyze charge groups in general contexts. Our main objective will be to understand the asymptotic decay rate of the corresponding potential (or resulting force field, the derivative of the potential). We have seen in Section 10.1 that the exponent of the decay rate is crucial in determining the global effect of the charge group. Thus our main objective is to develop only qualitative comparisons of different charge groups. Some quantitative comparisons are presented in Figure 10.15, but even these are only intended to clarify the qualitative relationsip between different types of charge groups.

The general form of a potential for a charged system can be written as a sum of point charge potentials

$$
V(\mathbf{r}) = \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r} - \mathbf{r}_k|},
\tag{10.35}
$$

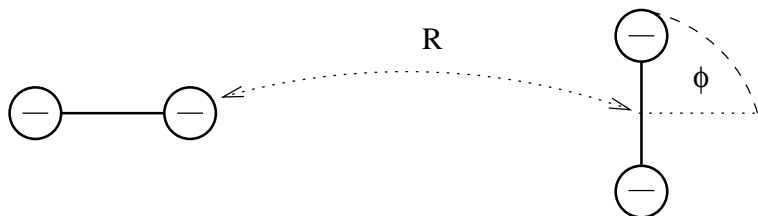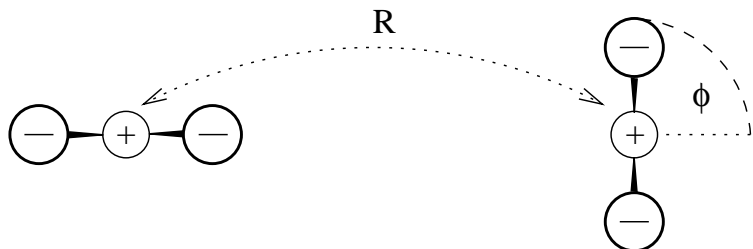Figure 10.14: Charge-charge interaction configuration similar to what is found in an interaction between Asp and Glu, but with a more refined model. The pair of negative charges on the right are at $(\cos\phi, \sin\phi, 0)$ and $(-\cos\phi, -\sin\phi, 0)$ and the associated positive charge is at $(0, 0, -z)$. The pair of negative charges on the left are at $(R \pm 1, 0, 0)$, and the associated positive charge is at $(R, 0, -z)$.

where the charges q_k are at $\mathbf{r}_k$. When the net charge of the system is zero, we can interpret V as being defined by a difference operator applied to the fundamental charge potential

$$W(\mathbf{r}) = 1/|\mathbf{r}| \tag{10.36}$$

as follows. Define a translation operator $T_\mathbf{x}$ by

$$(T_\mathbf{x} f)(\mathbf{r}) = f(\mathbf{r} - \mathbf{x}) \tag{10.37}$$

for any function f. Then we can interpret the expression (10.35) as

$$V = \sum_{k=1}^{K} q_k T_{\mathbf{r}_k} W, \tag{10.38}$$

where W is defined in (10.36). In view of (10.38), we define the operator

$$\mathcal{D} = \sum_{k=1}^{K} q_k T_{\mathbf{r}_k}. \tag{10.39}$$

We will see that this corresponds to a difference operator when the net charge of the system is zero.

10.4.1   Asymptotics of general potentials

The decay of $V(\mathbf{r})$ for simple dipoles can be determined by algebraic manipulations as in Section 3.3. However, for more complex arrangements, determining the rate is quite complicated. Multipole expansions become algebraically complex as the order increases. Here we offer an alternative calculus to determine asymptotic behavior of general potentials. We begin with some more precise notation.

Let as assume that there is a small parameter ϵ that defines the distance scale between the charge locations. That is, we define

$$V_\epsilon(\mathbf{r}) = \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r} - \epsilon\mathbf{r}_k|}, \tag{10.40}$$

where we now assume that the $\mathbf{r}_k$'s are fixed and are of order one in size. There is a dual relationship between the asymptotics of $V_\epsilon(\mathbf{r})$ as $\mathbf{r} \to \infty$ and $\epsilon \to 0$, as follows:

$$\boxed{V_\epsilon(\mathbf{r}) = |\mathbf{r}|^{-1} V_{\epsilon/|\mathbf{r}|}\left(|\mathbf{r}|^{-1}\mathbf{r}\right).} \tag{10.41}$$

The proof just requires changing variables in (10.40):

$$V_\epsilon(\mathbf{r}) = \frac{1}{|\mathbf{r}|} \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r}|^{-1}|\mathbf{r} - \epsilon\mathbf{r}_k|} = \frac{1}{|\mathbf{r}|} \sum_{k=1}^{K} \frac{q_k}{|(|\mathbf{r}|^{-1}\mathbf{r}) - (\epsilon/|\mathbf{r}|)\mathbf{r}_k|} = |\mathbf{r}|^{-1} V_{\epsilon/|\mathbf{r}|}\left(|\mathbf{r}|^{-1}\mathbf{r}\right). \tag{10.42}$$

Given a general potential V of the form (10.35), we can think of this as having $\epsilon = 1$, that is, $V = V_1$. Using (10.42), we can derive the asymptotic form

$$V(\mathbf{r}) = V_1(\mathbf{r}) = |\mathbf{r}|^{-1} V_{|\mathbf{r}|^{-1}}\left(|\mathbf{r}|^{-1}\mathbf{r}\right) = \epsilon V_\epsilon\left(\omega\right), \tag{10.43}$$

where we now define $\epsilon = |\mathbf{r}|^{-1}$ and $\omega = |\mathbf{r}|^{-1}\mathbf{r}$ satisfies $|\omega| = 1$. This says that we can determine asymptotics of V as $\mathbf{r} \to \infty$ by considering instead the behavior of V_ϵ on bounded sets (e.g., ω with $|\omega| = 1$) as $\epsilon \to 0$.

The reason that V_ϵ is useful is that we can write it in terms of a difference operator applied to W. Recalling (10.39), we define

$$\mathcal{D}_\epsilon = \sum_{k=1}^{K} q_k T_{\epsilon\mathbf{r}_k}, \tag{10.44}$$

and observe from (10.35) and (10.36) that

$$V_\epsilon = \mathcal{D}_\epsilon W. \tag{10.45}$$

We will see in typical cases that, for some $k \geq 0$,

$$\lim_{\epsilon \to 0} \epsilon^{-k} \mathcal{D}_\epsilon = \mathcal{D}_0, \tag{10.46}$$

where $\mathcal{D}_0$ is a differential operator of order k. The convergence in (10.46) is (at least) weak convergence, in the sense that for any smooth function f in a region $\Omega \subset \mathbb{R}^3$,

$$\lim_{\epsilon \to 0} \epsilon^{-k} \mathcal{D}_\epsilon f(\mathbf{x}) = \mathcal{D}_0 f(\mathbf{x}) \tag{10.47}$$

uniformly for $\mathbf{x} \in \Omega$. In particular, we will be mainly interested in sets Ω that exclude the origin, where the potentials are singular. Thus we conclude that

$$\lim_{\epsilon \to 0} \epsilon^{-k} V_\epsilon(\omega) = \mathcal{V}(\omega), \qquad |\omega| = 1, \tag{10.48}$$

where the limiting potential is defined by

$$\mathcal{V}(\mathbf{r}) = \mathcal{D}_0 W(\mathbf{r}). \tag{10.49}$$

Applying (10.43), (10.48), and (10.49), we conclude that

$$V(\mathbf{r}) \approx \frac{1}{|\mathbf{r}|^{k+1}} \mathcal{D}_0 W(|\mathbf{r}|^{-1}\mathbf{r}), \tag{10.50}$$

for large $\mathbf{r}$. More precisely, we will typically show that

$$\epsilon^{-k} \mathcal{D}_\epsilon \phi(\mathbf{r}) = \mathcal{D}_0 \phi(\mathbf{r}) + \mathcal{O}(\epsilon) \tag{10.51}$$

in which case we can assert that

$$V(\mathbf{r}) = \frac{1}{|\mathbf{r}|^{k+1}} \mathcal{D}_0 W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-k-2}). \tag{10.52}$$

10.4.2   Application of (10.52)

Let us show how (10.52) can be used in practice by considering a known situation, that of a dipole. Thus take $\mathbf{r}_1 = (\frac{1}{2}, 0, 0)$ and $\mathbf{r}_2 = (-\frac{1}{2}, 0, 0)$. We can compute the action of $\mathcal{D}_\epsilon$ on smooth functions via

$$\mathcal{D}_\epsilon \phi(x, y, z) = \phi(x + \tfrac{1}{2}\epsilon, y, z) - \phi(x - \tfrac{1}{2}\epsilon, y, z). \tag{10.53}$$

By Taylor's theorem, we can expand a function ψ to show that

$$\psi(x + \xi) - \psi(x - \xi) = 2\xi \psi'(x) + \tfrac{1}{3}\xi^3 \psi^{(3)}(x) + \mathcal{O}(\xi^5). \tag{10.54}$$

Applying (10.54) to $\psi(x) = \phi(x, y, z)$, we have

$$\mathcal{D}_\epsilon \phi(x, y, z) = \epsilon \frac{\partial}{\partial x} \phi(x, y, z) + \mathcal{O}(\epsilon^3). \tag{10.55}$$

Taking limits, we see that

$$\epsilon^{-1} \mathcal{D}_\epsilon \to \frac{\partial}{\partial x} \tag{10.56}$$

as $\epsilon \to 0$. Thus we conclude that the potential for a dipole is $\mathcal{O}(|\mathbf{r}|^{-2})$ for large $\mathbf{r}$, in keeping with the derivation in Section 3.3. More precisely, applying (10.52) we have

$$V(\mathbf{r}) = |\mathbf{r}|^{-2} \left(\frac{\partial}{\partial x} W \right) (|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-3}). \tag{10.57}$$

10.5   Quadrupole potential

The most important potential after the dipole is the quadrupole. As the name implies, it typically involves four charges. For this reason, the geometry can be quite complex. This provides an opportunity to apply the techniques developed in Section 10.4. We begin with a simple case.

(a)

(b)
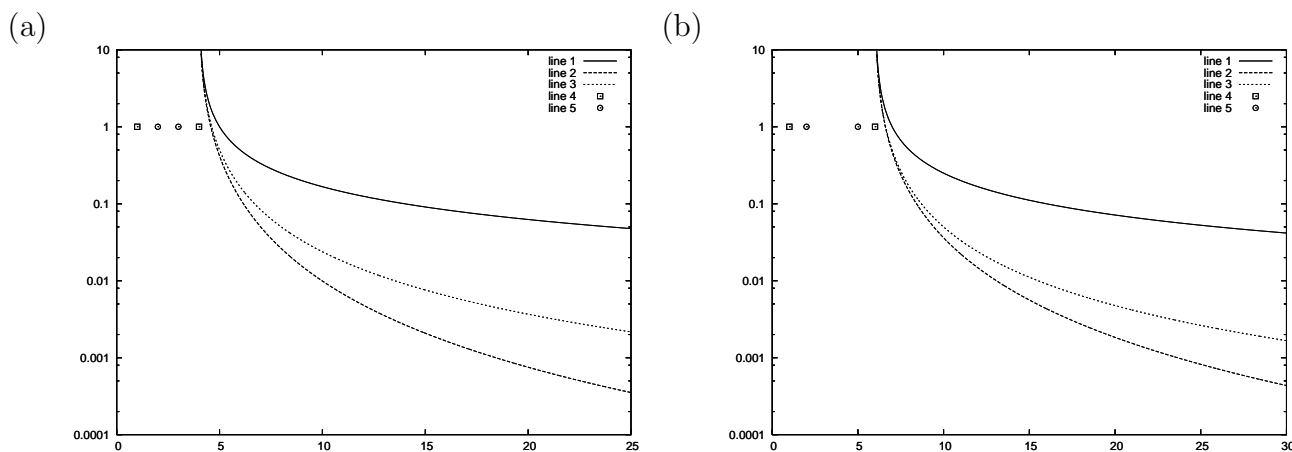
Figure 10.15: Comparison of single charge, dipole and quadrupole potentials. Dipole separation (a) two units and (b) four units. The locations of the negative charges are indicated by circles and the locations of the positive charges are indicated by squares. The upper solid line is the potential for a single positive charge whose horizontal position is indicated by the right-most square. The middle, short-dashed line is the potential for the dipole corresponding to the right-most dipole. The lower, longer-dashed line is the potential for the dipole corresponding to the quadrupole formed by the pair of dipoles.

10.5.1 Opposing dipoles

Two opposing dipoles tend to cancel each other out, but the result is not zero, rather it is a quadrupole. For example, suppose there unit negative charges at $\pm(a, 0, 0)$, where a is some (positive) distance parameter, with unit positive charges at $\pm(a + 1, 0, 0)$, viz.:

$$+ \quad - \qquad\qquad - \quad +$$

These four charges can be arranged as two dipoles, one centered at $a + \frac{1}{2}$ and the other centered at $-a - \frac{1}{2}$. Thus the separation distance S between the two dipoles is

$$S = 2a + 1. \tag{10.58}$$

The partial charges for a benzene ring as modeled in Table 12.1 consist of three sets of such paired dipoles, arranged in a hexagonal fashion.

The potential for such a charge group can be estimated by algebraic means, as we did in Chapter 3, or we can utilize the technology of Section 10.4. We define

$$\mathcal{D}_\epsilon = T_{\epsilon(a+1,0,0)} - T_{\epsilon(a,0,0)} + T_{\epsilon(-a-1,0,0)} - T_{\epsilon(-a,0,0)}. \tag{10.59}$$

In evaluating $\mathcal{D}_\epsilon \phi$, we may as well assume that ϕ is only a function of x, cf. Section 10.4.2. Applying (10.54) to ϕ and ϕ' we find that

$$
\begin{aligned}
\mathcal{D}_\epsilon \phi(x) &= \phi(x - \epsilon(a+1)) - \phi(x - \epsilon a) + \phi(x + \epsilon(a+1)) - \phi(x + \epsilon a) \\
&= \epsilon \phi'(x - \epsilon(a + \tfrac{1}{2})) - \epsilon \phi'(x + \epsilon(a + \tfrac{1}{2})) + \mathcal{O}(\epsilon^3) \\
&= \epsilon^2 (2a+1)\phi''(x) + \mathcal{O}(\epsilon^3) \\
&= \epsilon^2 S \phi''(x) + \mathcal{O}(\epsilon^3),
\end{aligned}
\tag{10.60}
$$

where S is the separation distance between the dipoles. Thus

$$
\lim_{\epsilon \to 0} \epsilon^{-2} \mathcal{D}_\epsilon = (2a+1)\frac{\partial^2}{\partial x^2} = S\frac{\partial^2}{\partial x^2},
\tag{10.61}
$$

where $S = 2a + 1$ is the separation distance (10.58) between the dipoles. Applying (10.52), we find

$$
V(\mathbf{r}) = |\mathbf{r}|^{-3} S \frac{\partial^2}{\partial x^2} W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-4})
\tag{10.62}
$$

for large $\mathbf{r}$, where $W \ (= 1/r)$ is defined in (10.36).

The potential for opposing dipoles is depicted in Figure 10.15 for two separation distances, $S = 2$ (a) and $S = 4$ (b). For the larger value of the separation, there is little difference between the dipole and quadrupole potentials near the right-most charge. There is a much greater difference between the potentials for a single charge and that of a dipole. Thus the separation distance affects substantially the cancellation of the second dipole, at least locally. If the distance units are interpreted as Ångstroms, then the separation $S = 4$ (b) is roughly comparable to the partial charge model of a benzene ring (cf. Table 12.1) consisting of three sets of such paired dipoles.

10.5.2 Four-corner quadrupole

The four-corner arrangement provides a two-dimensional arrangement of opposing dipoles, as follows:

$$
\begin{array}{cc}
+ & - \\
\\
- & +
\end{array}
$$

This quadrupole system has positive charges $q_1 = q_2 = 1$ at $\mathbf{r}_1 = (-1, 1, 0)$ and $\mathbf{r}_2 = (1, -1, 0)$ and negative charges $q_3 = q_4 = -1$ at $\mathbf{r}_3 = (1, 1, 0)$ and $\mathbf{r}_4 = (-1, -1, 0)$. A plot of the potential along a diagonal where it is maximal is given in Figure 10.16. Note that it dies off a bit more rapidly than the potential for the opposing dipoles (cf. Figure 10.15). Defining

$$
\mathcal{D}_\epsilon = \sum_{k=1}^{K} q_k T_{\epsilon r_k}
\tag{10.63}
$$
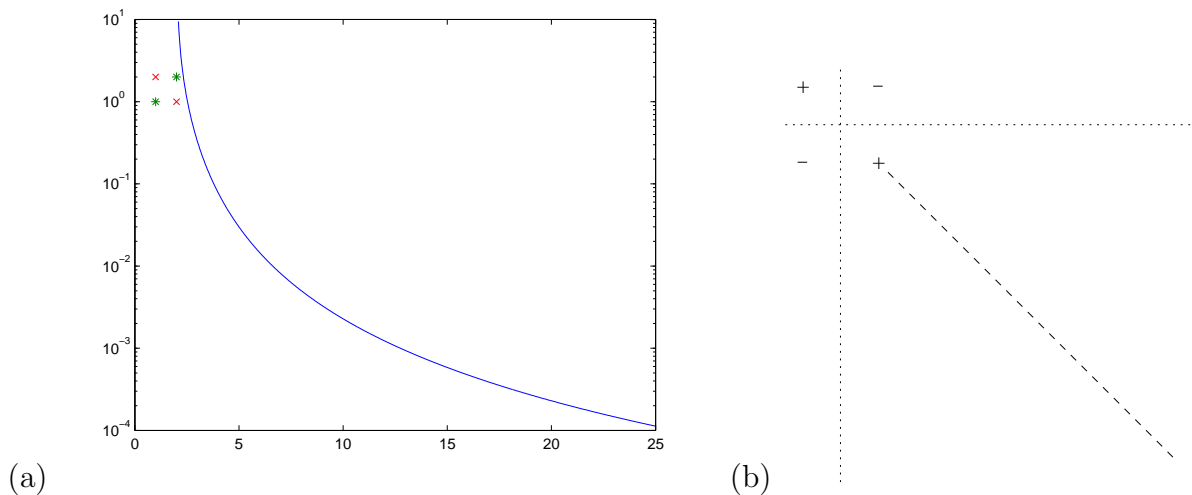
Figure 10.16: Four corner quadrupole potential. (a) The potential is plotted as a function of distance s along the line $(x(x), y(s)) = ((2 + s/\sqrt{2}, 1 - s/\sqrt{2})$ which emanates from the lower-right corner of the quadrupole. The locations of the negative charges are indicated by circles and the locations of the positive charges are indicated by squares. (b) Schematic representation. The line used for the plot in (a) is indicated as a dashed line. The potential vanishes, by symmetry, on the dotted lines.

and applying (10.54) twice, we see that

$$
\begin{aligned}
\mathcal{D}_\epsilon \phi(\mathbf{r}) &= \sum_{k=1}^{K} q_k \phi(\mathbf{r} - \epsilon \mathbf{r}_k) \\
&= 4 \frac{\partial}{\partial x} \frac{\partial}{\partial y} \phi(0) \epsilon^2 + \mathcal{O}(\epsilon^3).
\end{aligned}
\tag{10.64}
$$

Thus

$$
V(\mathbf{r}) = |\mathbf{r}|^{-3} 4 \frac{\partial}{\partial x} \frac{\partial}{\partial y} W(|\mathbf{r}|^{-1} \mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-4}).
\tag{10.65}
$$

It is not hard to generalize these results to the case where the opposing charges are located at the four corners of any parallelogram.

10.5.3 Quadrupole example

An example of a (near) quadrupole is found in the human prion (PDB file 1I4M) in the motif DRYYRE [146]. This is shown in Figure 10.17. The charges closely approximate the 'four corner' arrangement for a suitable parallelogram. The DRYYRE residue group forms a helical structure. Note that the four charged sidechains are nearly planar, with the tyrosines transverse to this plane. The detail Figure 10.17(b) shows the skewness of the two opposing dipoles.

(a)   (b)
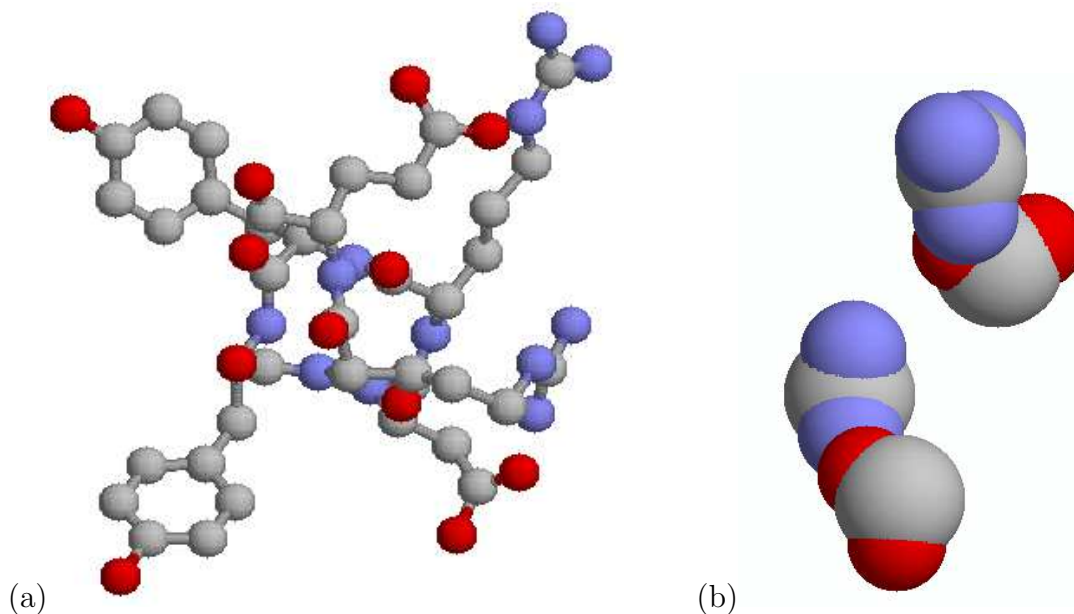
Figure 10.17: A near quadrupole found in the PDB file 1I4M of the human prion. (a) The four charged groups are nearly aligned on the right side of the figure. Shown is the residue sequence DRYYRE. (b) Detail of the charged groups indicating the alignment of the opposing dipoles.

10.5.4   Water: dipole or quadrupole?

Water can be written as a combination of two dipoles, following the general pattern of Section 10.4. So is water a quadrupole or just a dipole? The answer is crucial to determine the locality or globality of water–water interaction.

We can approximate the electronic structure of water as system with positive charges

$$q_1 = q_2 = a \quad \text{at} \quad \mathbf{r}_{1,2} = (\pm c, -1, 0) \tag{10.66}$$

and negative charges

$$q_3 = q_4 = -b \quad \text{at} \quad \mathbf{r}_{3.4} = (0, y^0, \pm d), \tag{10.67}$$

where $y^0 > 0$ denotes the position above the x-axis of the lone-pair oxygen charges. Note that we have chosen the spatial unit so that the hydrogens are exactly one unit below the x-axis (and the charge center is the origin), but otherwise all positions are arbitrary. This is exactly the model of water that is known as Tip5P [279], with $a = b$. We would like to show that this system is a dipole; by that, we mean two things, one of which is that it is *not* a quadrupole.

To discover the exact multipole nature of the water model encoded in (10.66) and (10.67), we modify it to form a quadrupole. We extend the system (10.66–10.67) to involve two more charges:

$$q_5 = -2a \quad \text{at} \quad \mathbf{r}_5 = (0, -1, 0) \quad \text{and}$$
$$q_6 = 2b \quad \text{at} \quad \mathbf{r}_6 = (0, y_0, 0). \tag{10.68}$$

The configuration of charges is depicted in Figure 10.18.

The extended system (10.66),(10.67),(10.68) is a quadrupole due to the cancellations leading to an expression such as (10.64). More precisely, note that the charges at locations 1, 2 and 5 correspond to a second difference stencil centered at point 5 for approximating

$$\frac{\partial^2 \phi}{\partial x^2}(0, -1, 0) \qquad (10.69)$$

(with suitable scaling). Similarly, the charges at locations 3, 4 and 6 correspond to a second difference stencil centered at point 6 for approximating

$$\frac{\partial^2 \phi}{\partial z^2}(0, y^0, 0) \qquad (10.70)$$

(with suitable scaling). Therefore

$$\mathcal{D}_\epsilon \phi(0) = \sum_{k=1}^{6} q_k \phi(\epsilon \mathbf{r}_k)$$

$$= ac^2 \epsilon^2 \frac{\partial^2 \phi}{\partial x^2}(0, -1, 0) - \qquad (10.71)$$

$$bd^2 \epsilon^2 \frac{\partial^2 \phi}{\partial z^2}(0, y^0, 0)\epsilon^2 + \mathcal{O}(\epsilon^4),$$

Figure 10.18: Configuration of charges in water model. Open circles indicate negative charge locations; shaded circles indicate locations of positive charge.

and a similar result would hold when expanding about any point $\mathbf{r}$.

Let V^D denote the potential of the system with charges as indicated in (10.68). We leave as Exercise 10.10 to show that this is a dipole provided $a = b$. Let V_Q denote the quadrupole potential associated with (10.71), and let V^W be the water potential using the model (10.66–10.67). Thus we have written the water potential as

$$V^W = V^D + V^Q \qquad (10.72)$$

for an explicit dipole potential V^D, with charges at $\mathbf{r}_5$ and $\mathbf{r}_6$, and a quadrupole. Thus the water model (10.66–10.67) is asymptotically a dipole, and not a quadrupole. Moreover, we see that the axis of the dipole is the y-axis, the bisector of the angle $\angle HOH$.

10.6   Multipole summary

Let us summarize the asymptotic behavior of the various potentials that can arise. We have seen in (10.52) that the order of decay of the potential can be determined by the arrangement of the charges (10.44). When the net charge is non-zero, we have $k = 0$, but when the net charge is zero then $k \geq 1$. The dipole is the case $k = 1$, and $k = 2$ is called a quadrupole. Similarly, $k = 3$ is

charge group	nonzero net charge	dipole	quadrupole	octapole
decay rate	r^{-1}	r^{-2}	r^{-3}	r^{-4}

Table 10.2: Asymptotic decay rates for potentials of various charge groups.

charge group	nonzero	dipole	quadrupole	octapole
nonzero	r^{-1}	r^{-2}	r^{-3}	r^{-4}
dipole	r^{-2}	r^{-3}	r^{-4}	r^{-5}
quadrupole	r^{-3}	r^{-4}	r^{-5}	r^{-6}
octapole	r^{-4}	r^{-5}	r^{-6}	r^{-7}

Table 10.3: Asymptotic decay rates for interaction energies of various charge groups.

an octapole, and so on. We summarize in Table 10.2 the different powers for the potential of these different charge groups.

The interaction energy between different charge groups has been worked out in specific cases. We summarize the general case in Table 10.3. We leave as Exercise 10.11 to verify the additional cases not already covered.

10.7 Further results

We collect here some further results about electrostatic interactions.

10.7.1 Dipole induction by dipoles

Water has both a fixed dipole and an inducible dipole. That is, water is both polar and polarizable. The dipole strength of water in the gas phase $\mu \approx 0.5$e-Å (cf. Section 17.7.2), and the polarizability $\alpha \approx 1.2$Å^3. Thus an electric field strength of only one-tenth of an electron per square Ångstrom $(0.1$e-Å$^{-2})$ could make a substantial modification to the polarity of water, since the change in polarity is approximated by the product of the polarizability and the electric field strength (see (3.27)).

10.7.2 Modified dipole interaction

Since the dipole-dipole interaction does not reproduce the sort of angular dependence we expect for certain bonds, e.g., hydrogen bonds, it is reasonable to try to modify the model. We ask the question: if the hydrogen charge density is represented in a more complex way, will a stronger angular dependence appear? To address this question, we introduce a negative charge to represent the electron density beyond the hydrogen. The exact positions of the charges are as follows. The position of the negative charge on the right we take as the origin, and we assume the separation
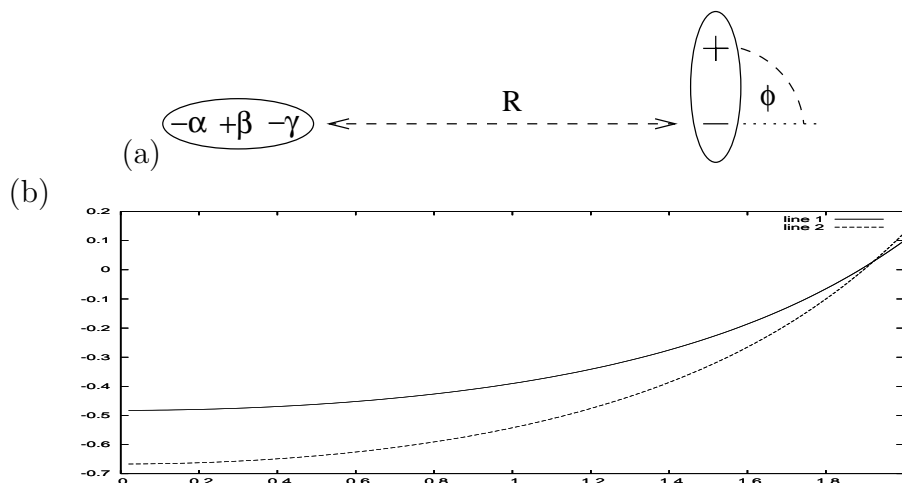
Figure 10.19:   (a) A modified model for dipole-dipole interaction. The negative (resp., positive) charge center of the dipole on the right is at $(0,0)$ (resp., at $(\cos\phi, \sin\phi)$). The charge centers of the multipole on the left are at $(-R-1,0)$ (negative charge $-\alpha$), $(-R,0)$ (positive charge $+\beta$) and $(-R+\delta,0)$ (negative charge $-\gamma$). (b) Dipole-dipole (in-line) interaction energy, scaled by R^3, for $R=2$ for the modified model versus the conventional model, depicted in Figure 10.3. Horizontal ϕ-axis measured in radians. The solid curve corresponds to $\alpha = 0.8, \beta = 1.0, \gamma = 0.2, \delta = 0.2$, whereas the flattest (dashed) curve corresponds to $\alpha = 1.0, \beta = 1.0, \gamma = 0.0$, which is the same as the model depicted in Figure 10.3.

distance between the charges is one. The separation of the positive charge on the left and the negative charge on the right is R. Thus the charge centers of the multipole on the left are at $(-R-1,0)$ (negative charge $-\alpha$), $(-R,0)$ (positive charge $+\beta$) and $(-R+\delta,0)$ (negative charge $-\gamma$). The positive charge on the right is at $(\cos\phi, \sin\phi)$. The original model depicted in Figure 10.3 corresponds to the choices $\alpha = 1$, $\beta = 1$, $\gamma = 0$, in which case the value of δ does not matter.

The distances between the various charges are easy to compute. The distance between the positive charge on the left and the negative charge on the right is R, and the distance between the main (α) negative charge on the left and the negative charge on the right is $R+1$. The distance between the minor (γ) negative charge on the left and the negative charge on the right is $R-\delta$.

The distance between the positive charge on the right and the minor (γ) negative charge on the left is

$$|(\cos\phi, \sin\phi) - (-R+\delta, 0)| = \sqrt{(R-\delta+\cos\phi)^2 + \sin^2\phi}$$
$$= \sqrt{1 + (R-\delta)^2 + 2(R-\delta)\cos\phi}. \tag{10.73}$$

The distance between the two positive charges is

$$|(\cos\phi, \sin\phi) - (-R, 0)| = \sqrt{(R+\cos\phi)^2 + \sin^2\phi}$$
$$= \sqrt{1 + R^2 + 2R\cos\phi}, \tag{10.74}$$

and the distance between the main (α) negative charge on the left and the positive charge on the right is

$$
\begin{aligned}
|(\cos\phi, \sin\phi) - (-R - 1, 0)| &= \sqrt{(1 + R + \cos\phi)^2 + \sin^2\phi} \\
&= \sqrt{1 + (R + 1)^2 + 2(R + 1)\cos\phi}.
\end{aligned}
\tag{10.75}
$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$
\begin{aligned}
&\frac{\alpha}{R + 1} - \frac{\beta}{R} + \frac{\gamma}{R - \delta} - \frac{\gamma}{\sqrt{1 + (R - \delta)^2 + 2(R - \delta)\cos\phi}} \\
&+ \frac{\beta}{\sqrt{1 + R^2 + 2R\cos\phi}} - \frac{\alpha}{\sqrt{1 + (R + 1)^2 + 2(R + 1)\cos\phi}}.
\end{aligned}
\tag{10.76}
$$

A plot of the interaction energy (10.76) is given in Figure 10.19 as a function of ϕ for $R = 2$, scaled by $R^{-3} = 8$. The flatter curve corresponds to the new model with a more complex dipole. Thus we see that this does not produce an improved model of the angular dependence of a hydrogen bond.

10.7.3   Hydrogen placement for Ser and Thr

Let us consider the problem of determining the angular orientation of the hydrogen in serine and threonine, depicted in Figure 6.5. We choose coordinates so that the x, y plane contains the terminal carbon and oxygen from the sidechain of Ser/Thr and the negative site of the partial charge of the moitie that is forming the hydrogen bond, as depicted in Figure 10.20. In the special case that the positive charge in the dipole forming the hydrogen acceptor is also in this plane, then we can argue by symmetry that the hydrogen must lie in this plane as well, at one of the solid dots indicated at the intersection of the circle with the plane of the page. But in general, we must assume that the location of the positive partial charge is outside of this plane.

In Figure 10.20(b), we indicate the view from the plane defined by the positions of the oxygen and the negative and positive partial charges of the dipole. The circle of possible locations for the hydrogen (see Figure 6.5) is now clearly visible, and the intersection points with the plane of the page are again indicated by black dots. Now we see it is not obvious what the optimal position for the hydrogen would be.

To determine the optimal hydrogen position, let us assume that the coordinates are as in Figure 10.20, with the origin chosen to be at the center of the circle. Thus, the plane of the page is the x, y plane, and the coordinates of the circle are $(0, \cos\theta, \sin\theta)$. The position of the negative partial charge is then $(x_0, y_0, 0)$ and the positive partial charge is (x_1, y_1, z_1). The interaction potential between the dipole and the hydrogen is thus

$$
\frac{-1}{\sqrt{x_0^2 + (y_0 - \cos\theta)^2 + \sin^2\theta}} + \frac{1}{\sqrt{x_1^2 + (y_1 - \cos\theta)^2 + (z_1 - \sin\theta)^2}}
\tag{10.77}
$$

For given x_0, y_0, x_1, y_1, z_1, this expression can be minimized to find the optimal θ.
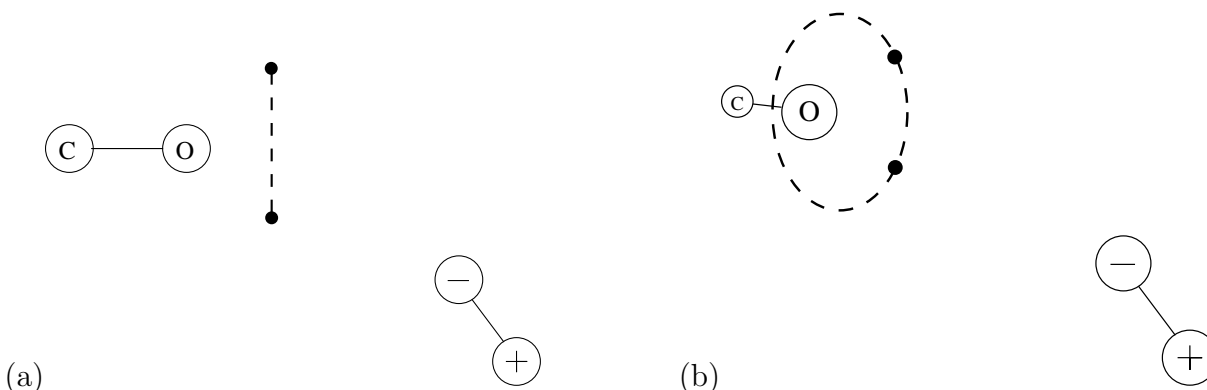
Figure 10.20: Configuration for the placement of hydrogen at the end of the sidechain of serine or threonine in response to a nearby dipole. The dashed line indicates the circle of possible hydrogen placements. (a) The plane of the circle is orthogonal to the plane of the page. (b) The plane of the circle is skew to the plane of the page.

We can also use the expression (10.30) to find the optimum θ. In coordinates determined so that the hydrogen and the dipole lie in a plane, the interaction field (10.30) has a zero component orthogonal to the plane. For the hydrogen position on the circle to be correct, the tangent to the circle must be orthogonal to the gradient of the interaction potential at that point. Suppose that we write the circle as $(x(\phi), y(\phi), z(\phi))$ in these coordinates. Then a necessary condition is that

$$\nabla V(x(\phi), y(\phi), z(\phi)) \cdot (x'(\phi), y'(\phi), z'(\phi)) = 0. \tag{10.78}$$

10.8   Exercises

Exercise 10.1 *Show that the interaction energy (10.6) tends to the asymptotic form*

$$\frac{-2\cos\phi}{R^3}. \tag{10.79}$$

Exercise 10.2 *Show that the interaction energy (10.9) tends to the asymptotic form*

$$\frac{-\frac{1}{2} - \frac{3}{2}\cos\phi}{R^3}. \tag{10.80}$$

Exercise 10.3 *Verify that the second term in the energy expression in (10.13) is indeed the same as (10.7). Also verify that the fourth term in the energy expression in (10.13) is correct.*

Exercise 10.4 *Plot the expression in (10.14) and verify that it is not symmetric around $\phi = \pi$ for finite R. Determine the limiting expression as $R \to \infty$ after scaling by R^3.*

Exercise 10.5 *Plot the expression in (10.34) and verify that it is symmetric around $\phi = \pi/2$ and has a simple minimum there.*

Exercise 10.6 *Carry out the calculations leading to the expression in (10.34) in the case that the charge group has a positive charge as well as the negative charges, as shown in Figure 10.14. Take the charges to be appropriate for Asp or Glu. Investigate the minimum energy configuration. Also consider three-dimensional configurations in which the positive charge is located below the negative charges.*

Exercise 10.7 *Investigate the optimal (minimum energy) configuration for charge-dipole pairs in which the charge is fixed at a distance r from the center of the dipole, which is free to rotate by an angle ϕ. Determine the value of ϕ at the minimum.*

Exercise 10.8 *Prove that the asymptotic expression (10.28) is valid for fixed θ and large r. (Hint: show that*

$$V(r, \theta) = \frac{a}{r} \left(\frac{\sqrt{1 - 2r^{-1}\cos\theta + r^{-2}} - \sqrt{1 + 2r^{-1}\cos\theta + r^{-2}}}{\sqrt{1 - 2r^{-1}\cos\theta + r^{-2}}\sqrt{1 + 2r^{-1}\cos\theta + r^{-2}}} \right) \tag{10.81}$$

and expand the expression in the numerator. Is this asymptotic approximation uniformly valid for all θ?)

Exercise 10.9 *Determine the percentage error in the approximation (10.28) when $\theta = \pi/4$ and $r = 3$.*

Exercise 10.10 *Show that a charge system with only the charges as indicated in (10.68) forms a dipole provided $a = b$ and examine its asymptotic behavior.*

Exercise 10.11 *Verify the interaction energies listed in Table 10.3 for the cases not already covered. (Hint: develop technology similar to Section 10.4.1. The interaction introduces an additional difference operator that multiplies the one associated with the potential. The order of the product of the two limiting differential operators is equal to the sum of the orders of the individual operators.)*

Exercise 10.12 *Plot the forces corresponding to the potentials in Figure 10.15. That is, compute and compare the forces on a single charge from an isolated charge, a dipole and a quadrupole.*

Exercise 10.13 *Suppose that the units for the horizontal axis in Figure 10.15 are chosen to be Ångstroms, and that the unit of charge corresponds to one electron. Determine the units of the vertical axis (see Chapter 17).*

Exercise 10.14 *Compare the asymptotic expression (10.23) for the general dipole-dipole interaction with the special cases considered earlier in the text, e.g., (3.11), (3.15), (10.79) and (10.80).*

Exercise 10.15 *Determine whether the force between two neutral groups as depicted in Figure 10.21 is attractive or repulsive.*
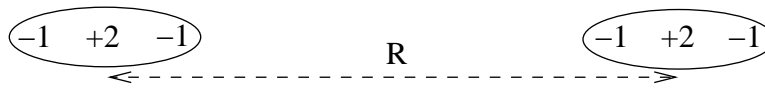
Figure 10.21: Interaction between two quadrupoles.

Exercise 10.16 *Show that the interaction energy for two quadrupoles as depicted in Figure 10.21 is*

$$\frac{1}{r-2} - \frac{4}{r-1} + \frac{6}{r} - \frac{4}{r+1} - \frac{1}{r+2} \approx \left(\frac{\partial^4}{\partial r^4}\right)\left(\frac{1}{r}\right) = \frac{24}{r^5} \tag{10.82}$$

as $r \to \infty$ *by considering the error in the difference operator represented by the left-hand side, cf. Exercise 3.7.*

Exercise 10.17 *Plot* $\phi(\theta)$ *given by (10.26). Be careful about what happens when the denominator vanishes.*

Exercise 10.18 *Repeat the computations in Section 10.2.2 using the revised coordinates given in Figure 10.11.*

Chapter 11

Case studies

We have established that dehydrons are sticky binding sites for protein ligands. Here we examine several cases where these play an important biological role. We begin by considering diverse examples in the literature where specific dehydrons can be seen to be critical in specific interactive sites. These cover different types of interactions, involving signalling, strucutral roles, and enzymatic activity. Subsequently, we study protein interactions from a more high-level view, showing that the number of dehydrons in proteins correlates positively with protein interactivity.

11.1 Basic cases

Several examples of the role of dehydrons in protein associations were given in [142]. We review them here briefly. They illustrate the way that dehydrons play a role in signaling and in protein structure formation.

11.1.1 A singular case: signaling

One striking example of a dehydron involved in protein-protein association is the binding of the light chain of antibody FAB25.3 with the HIV-1 capsid protein P24 [142], as depicted in the PDB file 1AFV. Antibodies are proteins that bind to antigens; they can in some cases directly neutralize the antigen, but in most cases they tag things for attack by the immune system. In this role, we can think of antibodies as playing the role of signalling.

Antibodies consist of two ('light' and 'heavy') chains. The smaller light chain has an important region that resembles a finger reaching out to touch the antigen at a particular location. It was found that this finger points directly at a dehydron in the antigen [142]. Thus it appears that the dehydron provides an important part of the recognition process for this antibody-antigen complex.

In Figure 11.1, a cartoon of Figure 2 in [142] depicts a portion of the capsid protein P24 (on the left) with the dehydron depicted as a grey strip. The light chain of the antibody FAB25.3 is depicted on the right, pointing directly at the dehydron on the protein P24 [142]. The binding interface for this complex is quite small and yet the single dehydron in this helix of the HIV-1 capsid protein lies precisely at the interface with the extended finger of the light chain.

The tip of the finger in the light chain in this case is the residue Tyr 32 in chain L. Five of the nonpolar carbons in this residue are in a desolvation domain (radius 6.5Å) of the hydrogen bond between Thr 108 (donor) and Ser 102 (acceptor) in the antigen (chain A), which is one of a sequence of hydrogen bonds in a helix in the HIV-1 capsid. The desolvation domain (consisting of spheres centered at the CA atoms of the donor and acceptor) contains only 9 nonpolar groups natively available from chain A. The wrapping of this dehydron by Tyr 32 is thus quite significant.

Figure 11.1: Cartoon of two fragments of proteins binding at a dehydron on one of the proteins. The dehydron is depicted as a grey strip.

It should be noted that the desolvation domain of this dehydron also contains four nonpolar groups from the B chain in 1AFV. The PDB file 1AFV consists of two copies of the antibody-antigen complex. Chains A, H, L represent one copy and chains B, K, M represent the other. Thus atoms from chain B nearby atoms in chain A represent a crystalization artifact. It is not surprising (Section 11.1.3) that a crystal contact would appear at a dehydron, but it could be of concern is that the region around this underwrapped dehydron only becomes structured due to the artificial dimer. This dehydron is part of a short helix involving only residues 101 to 104 in the A chain in the HIV-1 capsid.

11.1.2   Forming structures

Dehydrons also appear more broadly in structural roles [142]. In some cases, they guide formation of quarternary structures, such as dimers. For example, dehydrons (G49, G52), (G78, T80), and (T91, G94) in the HIV-1 protease guide the formation of the dimer structure [137]. In other cases, the structures can be quite complex. A cartoon of this kind of behavior is presented in Figure 11.2.

Virus capsids can be viewed as a model for protein-protein interaction [437, 438, 439]. The formation of the capsid in picornaviruses [162, 241] was shown to be essentially determined by the distribution of dehydrons in the individual virus-peptide (VP) subunits [142]. This distribution concentrates at the symmetry centers of the capsid and edge-to-edge subunit positioning.

There are three individual VP (virus peptide) subunits that assemble into a virus unit for the foot-and-mouth disease virus (FMDV) [162]. The atomic coordinates correspond to PDB entry 1BBT. This type of assembly of a virus capsid from several copies of a small number of VP subunits is typical [437, 439, 440]. Regions of high concentration of dehydrons are found at the symmetry centers of the capsid, and the exposure of these dehydrons to solvent is eliminated upon the associations of the units (Figure 5a in [142]). Approximately two-thirds of the dehydrons involved in domain-swapping and unit assembly become 'desensitized' upon formation of the unit. Morever, the capsid has four remaining regions with a high concentration of dehydrons.

A comparison between Figures 5a and 5b of [162] reveals that the four regions in the unit with a high dehydron density can be readily associated either with the centers of symmetry of the capsid or to the inter-unit edge-to-edge assembly of VP2 and VP3 [162]. Thus, the FMDV unit presents

two loop regions highly sensitive to water removal on VP1 and VP3 which are directly engaged in nucleating other units at the pentamer and hexamer (with three-fold symmetry) centers of the capsid. Furthermore, a helical region on VP2 is severely under-wrapped and occupies the central dimeric center (shown in Fig. 5b of [162]).

The FMDV capsid protein VP2 has a 'handle' region on the β-hairpin near the N-terminus (12-27 region) that is known to be structurally defective, and it is also underwrapped. This region is also present in the VP2 subunits of all the picornaviruses. Because of its high density of dehydrons, this region is a strong organizing center in the capsid. It has been suggested as a potential drug target [142].

Dehydrons in the human rhinovirus (PDB code 1R1A [225]) can be seen to have similar properties [142]. While most dehydrons can be attributed to the unit assembly, there are three particular sites with high dehydron density which do not become well wrapped either after the formation of the unit or after the assembly of the whole capsid. Two of them correspond to antibody binding sites, but the other site lies under the so-called canyon of the VP1 structure and has been known to be the target region for the drug WIN 51711 used to treat common cold [225].

Figure 11.2: Cartoon of two fragments of proteins binding along interfaces with multiple dehydrons on both of the proteins. The dehydrons are depicted as grey strips.

The pattern of dehydrons in the virus is shown in Fig. 7 of [142]. The viral unit has only two very strong dehydron centers on its rim: the pentamer center located in VP1, and the part of VP2 involved in the VP2-VP3 edge-to-edge contact. The remaining 15 dehydrons are listed in Table **??** are not involved in the organization of the capsid. Of the 60 residues known to be engaged in crystal contacts for this virus [241], 54 of them have sidechain carbonaceous groups in the desolvation domains of the 15 dehydrons marked in Table **??**. Thus, the dehydons not associated with the structural organization of the capsid can be seen to correlate with the crystal packing.

11.1.3   Crystal contacts

When proteins are crystalized for the purpose of imaging, the individual protein units contact each other at points called **crystal contacts** [115]. These frequently occur at dehydrons, since these promote complexation. Many PDB files involve multiple copies of proteins due the manner in which they are crystalized. Thus the crystal contacts may appear in the PDB file as interactions between different chains. Thus, they may be thought of as a type of artifact of the protein-protein interactions related to the imaging process. One example of a crystal contact was observed in Section 11.1.1.

Another view of this type of artifact can be seen in the PDB file 1P2C (see Section 14.1). This consists of two copies of an antibody-antigen complex. These complexes consist of two antibody chains and the antigen chain. In 1P2C, one of the antibody chains (chain A) makes several hydrogen bonds with the antigen (chain F) in the other copy of the complex. These are clearly nonphysical.

11.2 Enzymatic activity

Enzymes are proteins that **catalyze** (i.e., enhance) chemical reactions. We can think of them as machines, since they start with an input resource, a molecule called the **substrate**, and convert the substrate into a different molecule, the product. The process is called **catalysis**. The chemical reaction that is facilitated by an enzyme would occur naturally without a catalyst, but at a much slower rate.

Enzymes can be quite large proteins [421, 422], but the 'active site' in which catalysis takes place is localized. There are thousands of enzymes that play a role in metabolic pathways [362]. Many drugs are designed to be inhibitors of enzymes [137], disrupting the role of an enzyme in the metabolic system.

The active site of an enzyme will contain water until the substrate enters. Moreover, water removal is critical for the success of enzymatic process [137, 421]. The process of water removal can be enhanced energetically by the presence of dehydrons near the active site [137]. The target of the catalysis is frequently a small molecule, or a small protruding part of a larger molecule, and in this case, the dehydrons are found on the enzyme, as shown in the cartoon in Figure 11.3. Although most drugs were not designed to facilitate water removal, we can see retrospectively that particular drugs play a significant role in wrapping dehydrons [137, 144]. Moreover, drugs can be re-designed to enhance the interaction between hydrophobic groups on the drug and dehydrons on the target protein [96, 139, 144].

11.3 Neurophysin/vasopressin binding

In [52], a polar/aromatic interaction is suggested for the residue B-Tyr99 of the hormone vasopressin, a ligand of the bovine protein neurophysin (BNP)-II, a transporter of hormones along axons. Aromatic rings can in principle interact with any polar moiety (cf. Section 3.1.4 and Section 12.3). We analyze the interaction between neurophysin and vasopressin represented in PDB file 2BN2 discussed in [52] and exhibit some additional candidate interactions to explain the particular details of the interaction zone.

Figure 11.3: Cartoon of an enzyme with a small molecule (herringbone shading) in the active site. The dehydrons are depicted as grey strips.

Part of the reason for the suggestion [52] of a polar/aromatic interaction is the large number of disulfide bonds in neurophysin [358]. There are seven such bonds in a protein with only 79 residues; nearly a fifth of all of the residues are involved in one. Moreover, there is a disulfide bond in the ligand vasopressin. But potential polar interacting partner(s) of B-Tyr99 is(are) not made clear in [52]. For clarity, we designate the protein neurophysin as the A chain, and the ligand fragment of vasopressin the B chain, in keeping with the notation of the PDB file 2BN2. This fragment consists of only two residues, B-Phe98 and B-Tyr99. Thus B-Tyr99 denotes the tyrosine residue of vasopressin that is found in the active site

Donors	Acceptors	AW	Phe	Tyr
A-Cys 21 NH	A-Leu 11 OC	24	0	3
A-Phe 22 NH	A-Ile 26 OC	25	0	4
A-Gly 23 NH	A-Ile 26 OC	20	0	6
A-Ile 26 NH	A-Gly 23 OC	20	0	6
A-Cys 44 NH	A-Ala 41 OC	24	0	2
A-Gln 45 NH	A-Leu 42 OC	19	0	2
A-Asn 48 NH	A-Gln 45 OC	16	3	3
A-Gln 55 NH	A-Arg 8 OC	12	0	1
A-Cys 54 NH	B-Phe 98 OC	17	7	6

Table 11.1: Mainchain hydrogen bonds wrapped by B-Phe 98 and B-Tyr 99 in PDB file 2BN2. The column "AW" lists the number of wrappers coming from the A chain. The columns "Phe" and "Tyr" indicate the number of wrappers contributed by B-Phe 98 and B-Tyr 99, respectively, to the desolvation domain for the hydrogen bond. The desolvation domain radius chosen was 6.6Å.

of neurophysin.

One corollary of the large number of disulfide bonds in neurophysin is the expectation of a large number of underwrapped hydrogen bonds [127, 131]. There are 19 intramolecular mainchain hydrogen bonds in neurophysin with less than 20 wrappers using a desolvation domain radius of 6.6Å. We will see that there is correspondingly a large number of hydrogen bonds in neurophysin that are wrapped by the ligand vasopressin.

11.3.1   The role of tyrosine-99

Examination of the PDB file 2BN2 shows that B-Phe98 and B-Tyr99 wrap several under-wrapped mainchain-mainchain hydrogen bonds listed in Table 11.1. The active site, into which B-Tyr99 inserts, has a large water-exposed area without the ligand. They provide wrapping in a sector that is otherwise exposed to water attack.

The paper [52] reports binding constants for various mutations of the native B-Phe98, B-Tyr99 pair. A comparable binding constant, only 15% smaller, is obtained for B-Phe98, B-Phe99. The terminal O-H group in B-Tyr99 also forms a sidechain-mainchain hydrogen bond with A-Cys44(O). This presumably accounts for the additional binding strength that this ligand has compared to the B-Phe98,B-Phe99 variant [52]. B-Phe98,B-Phe99 is the corresponding residue pair at the binding site for the ligand phenypressin, a hormone in Australian macropods.

It is also interesting to note that the mutation B-Tyr→B-Leu would both modify wrapping patterns and eliminate the sidechain-mainchain hydrogen bond. This is reported [52] to have a significantly lower binding constant.

The CD1 and CE1 hydrophobic groups of B-Tyr99 are also in close proximity to the sidechain-mainchain bond A-Ser56(OG)—A-Cys21, which otherwise has a small number of nearby hydrophobic groups.

Donors	Acceptors	AW	Phe	Tyr	OBW
A-Cys 21 NH	A-Leu 11 OC	25	0	3	0
A-Phe 22 NH	A-Ile 26 OC	22	0	5	0
A-Ile 26 NH	A-Gly 23 OC	18	0	6	0
A-Cys 44 NH	A-Ala 41 OC	23	0	2	0
A-Gln 45 NH	A-Leu 42 OC	20	0	1	0
A-Glu 47 NH	A-Cys 44 OC	26	0	5	1
A-Asn 48 NH	A-Gln 45 OC	19	0	3	2
A-Leu 50 NH	A-Glu 47 OC	19	0	5	1
A-Gln 55 NH	A-Arg 8 OC	12	3	1	0
A-Cys 54 NH	B-Cys 1 OC	14	3	6	3
B-Phe 3 NH	A-Cys 54 OC	8	7	3	5
B-Asn 5 NH	B-Tyr 2 OC	6	2	6	5
B-Cys 6 NH	B-Tyr 2 OC	6	2	6	4

Table 11.2: Mainchain hydrogen bonds wrapped by B-Phe 3 and B-Tyr 2 in PDB file 1KJ4. The column "AW" lists the number of wrappers coming from the A chain. The columns "Phe" and "Tyr" indicate the number of wrappers contributed by B-Phe 3 and B-Tyr 2, respectively, to the desolvation domain for the hydrogen bond. The column "OBW" lists the number of other wrappers coming from the B chain. The desolvation domain radius chosen was 6.6Å.

In summary, B-Tyr99 is involved in an intermolecular hydrogen bond and in wrapping several dehydrons. The mutation B-Tyr99→B-Leu99 would eliminate both the intermolecular hydrogen bond and reduce significantly the wrapping of some of these bonds.

11.3.2   The role of phenylalanine-98

The role of B-Phe98 is less critical. It makes the mainchain-mainchain bond A-Cys54—B-Phe98 and wraps only the mainchain-mainchain bond A Asn 48–A Gln 45, shown in Figure 11.4. The mutation Leu98 binds with about half the affinity of Phe98 [52]. Presumably, the mainchain bond to A-Cys54 would be preserved, and the amount of wrapping only slightly decreased.

On the other hand, the combination B-Phe98,B-Leu99 has significantly reduced affinity compared with the native B-Phe98,B-Tyr99. This could be due to two different reasons. One could be a polar-aromatic interaction of some sort [258], but it could also be due to the factors discussed here. Compared with Tyr, Leu lacks the sidechain-mainchain bond with Cys44. Moreover, it is positions CD and CE on residue 99 that provide the wrapping. The CE positions are absent on Lue, and the CD positions are the end of the Leu sidechain, so its ability to wrap, compared with Phe, would be reduced. Similarly, two other second-position sidechains (Met and His) show minimal affinity, consistent with their lack of wrappers at the end of the sidechain and lack of ability to form a hydrogen bond at the end with Cys44.

The role of residue 48 in chain A was explored in [52] by a natural 'mutation' arising due to
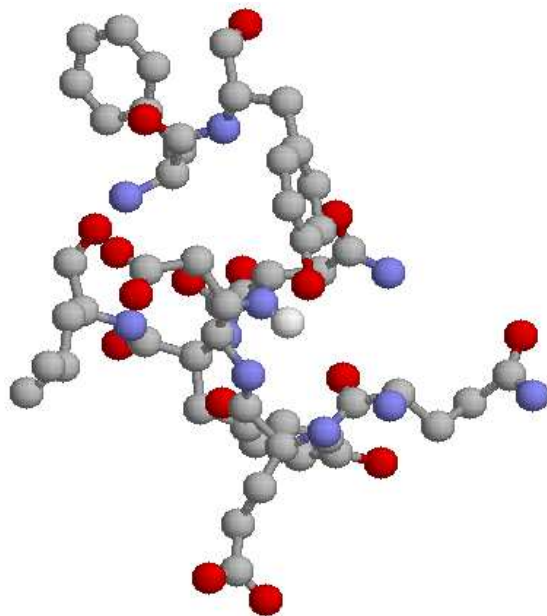
Figure 11.4: Wrapping of hydrogen bond Asn 48 N—Gln 45 O in PDB file 2BN2. The hydrogen (in white) attached to Asn48N has been added to indicate the bond, which is in the middle of the figure. The ligand pair B-Tyr99 and B-Phe98 is in the upper left of the figure. Residues 45 to 50 in the A chain are shown to indicate the bulk of the wrapping.

the fact that bovine (Asn 48) and ovine (Ile 48) differ at this location. The ovine affinity is slightly higher, but of the same order. The 'mutation' Asn 48→Ile 48 is isosteric, so it is plausible that the mainchain-mainchain hydrogen bond to Gln 45 O is maintained.

11.3.3   2BN2 versus 1JK4

The PDB structure 1JK4 also represents neurophysin bound to a larger fragment of vasopressin [428], but with the orientation of the fragment is reversed, that is, the sequence of the vasopressin fragment is B-Cys1, B-Tyr2, B-Phe3, B-Gln4, and so forth. The residues B-Tyr2, B-Phe3 in 1JK4 correspond to B-Phe98, B-Tyr99 in 2BN2. The individual bonds, listed in Table 11.2, are quite similar, and the general picture is the same. One new ingredient is the fact that dehydrons in the vasopressin fragment can now be seen that are wrapped in part by residues from neurophysin. Both B-Asn 5 — B-Tyr 2 and B-Cys 6 — B-Tyr 2 have these complex wrapping patterns.

   The main difference that becomes apparent in 1JK4 is the decreased role of B-Phe3. Indeed, upon examination, it is no longer in the active site but rather sticking out of it. B-Tyr 2 is found in 1JK4 in the same place as B-Tyr 99 in 2NB2, and in 1JK4 its role as a wrapper is more pronounced, as incicated in Table 11.2.
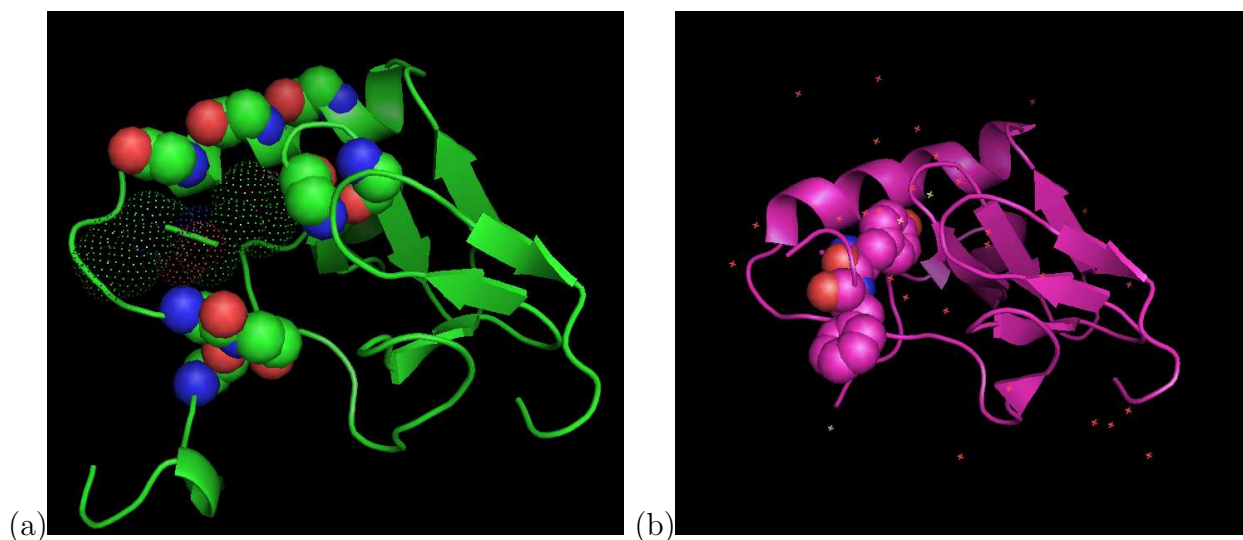
Figure 11.5: (a) The active site of the bovine protein neurophysin bound to a fragment of the hormone vasopressin, in the PDB file 2BN2. The fragment B-Phe 98, B-Tyr 99 is shown as a cloud, and several of the hydrogen bonds in Table 11.1 are indicated by showing their backbone atoms as spheres. (b) The active site of the bovine protein neurophysin bound to a fragment of the hormone vasopressin, in the PDB file 1JK4. The two residues B-Tyr 2, B-Phe 3 of the vasopressin fragment are shown as spheres.

11.4   Variations in proteomic interactivity

So far, we have considered the implications of individual dehydrons in individual proteins regarding protein-ligand interactions. Now we take a more global view to see how the number of dehydrons may correlate with interactivity generally. Our studies in Sections 11.1, 11.2 and 11.3 showed that dehydrons can account for particular interactions, and the review in the current section will show that these are not special cases, but rather are likely to represent the general trend. To verify this, we will use standard correlation analysis, but we will also include the type of graph-theory analysis introduced in Section 5.5. In addition to this mathematical technology, we will also introduce the use of additional databases of protein properties beyond the PDB used so far.

     The proteins in higher eukaryotes are known to have greater interactivity, despite the fact that genomes sizes are not appreciably larger. The **transcriptome** is the set of proteins that are generated from a genome (or set of genomes). In higher eukaryotes, this can be much larger due to multiple splicing of different parts of genes, but even this does not expand the genome size sufficiently to account for the significantly greater interactivity [234, 405, 429]. Thus there must be some factors that lead to greater interactivity in higher eukaryotes, and we will see that the dehydron is a good candidate as an important contributing factor.

     Different species have similar proteins that are similar in sequence, structure, and function. These **homologous** proteins provide a means to study variations in protein sequence and function,

Protein	Species (common name)	PDB code	N	n_{HB}	ρ	n_{DH}
Cytochrome c	Chlamydomonas reinhardtii (algae)	1cyi	89	52	19.74	6
Cytochrome c	Rhodophila globiformis (bacteria)	1hro	105	50	17.52	7
Cytochrome c	Oryza sativa (Asian rice)	1ccr	111	55	14.94	11
Cytochrome c	Katsuwonus pelamis (skipjack tuna)	1cyc	103	41	14.03	12
Cytochrome c	Thunnus alalunga (albacore tuna)	5cyt	103	53	14.25	13
Cytochrome c	Equus caballus (horse)	1giw	104	44	14.01	14
Hemoglobin	Vitreoscilla stercoraria (bacteria)	2vhb	136	102	23.50	0
Hemoglobin	Lupinus luteus (yellow lupin, plant)	1gdj	153	109	23.43	0
Hemoglobin	Paramecium caudatum (protozoa)	1dlw	116	77	22.02	0
Hemoglobin	(Nonsymbiotic) Oryza sativa (rice)	1d8u	165	106	23.58	2
Hemoglobin	Equus caballus (horse)	1gob	146	101	21.45	2
Hemoglobin	Home sapiens (modern human)	1bz0	146	103	21.45	3
Myoglobin	Aplysia limacina (mollusc)	1mba	146	106	23.42	0
Myoglobin	Chironomus thummi thummi (insect)	1eca	136	101	21.31	3
Myoglobin	Thunnus albacares (yellow-fin tuna)	1myt	146	110	21.15	8
Myoglobin	Caretta caretta (sea turtle)	1lht	153	110	21.09	11
Myoglobin	Physeter catodon (sperm whale)	1bz6	153	113	20.98	11
Myoglobin	Sus scrofa (wild boar)	1mwc	153	113	19.95	12
Myoglobin	Equus caballus (horse)	1dwr	152	112	18.90	14
Myoglobin	Elephas maximus (Asian elephant)	1emy	153	115	18.90	15
Myoglobin	Phoca vitulina (seal)	1mbs	153	109	18.84	16
Myoglobin	Homo sapiens (modern human)	2hbc	146	102	18.80	16

Table 11.3: Variations in numbers of dehydrons in homologous proteins in different species. N is the number of residues in the protein, n_{HB} denotes the number of hydrogen bonds, ρ is the average number of wrappers of the hydrogen bonds in the protein, with a desolvation radius of 6.4Å and dehydron criterion of $\rho_G < 12$, and n_{DH} denotes the number of dehydrons in the protein.

and they can be used to understand variations in distributions of dehydrons as well. When two such proteins have similar properties, we say that they are **conserved** with respect to evolution. By examining conserved proteins across species, significant differences in the number and distribution of dehydrons were found [145]. Within a conserved domain fold, the number of dehydrons in higher eukaryotes is consistently greater than in species (e.g., bacteria) of lower complexity, as indicated in Figure 2.2 and Table 11.3.

Species for which particular proteins are known to have more complex interactions have a higher number of dehydrons in their proteins, whereas the homologous proteins in more archaic species have far fewer dehydrons. A clear illustration of this trend is provided in Fig. 2 in [142], where three different versions of myoglobin corresponding to aplysia limacina (gastropode, mollusc), whale and human are displayed and their respective distribution of dehydrons is highlighted. The myoglobin from aplysia limacina is one of the best wrappers of hydrogen bonds in the entire PDB [142].

11.4.1 Interactivity correlation

A study of how dehydrons provide a novel indicator of proteomic interactivity was presented in [145]. The number of dehydrons of a given protein in different organisms can vary substantially [142], as we have depicted in Figure 2.2 for myoglobin. Six other groups of proteins were analyzed in [142], and these variations are evident in Table 11.3 which reproduces the data from [142]. A larger number of dehydrons implies a larger number of possible interactions, due to the fact that they are sticky (Chapter 9). Thus it is reasonable to ask if these observations can be used to relate classes of proteins using their potential interactivity, as measured by the number of dehydrons, as an indicator.

Although relating protein folds by their inclusion in a single protein (cf. Sections 5.5 and 11.4.2) reveals some type of fold interactivity, it is possible to choose more direct measures of protein interaction. Interactivity of proteins can be determined by various means, and this is documented in various databases [50], such as DIP, the Database of Interacting Proteins [352]. The objective of DIP is "to integrate the diverse body of experimental evidence on protein-protein interactions" [352].

In [145], a subset of DIP was selected containing domains from the yeast proteome whose interactions were determined by a single class of experiments. For the proteins in this subset having PDB structures, the average amount of wrapping of all hydrogen bonds was determined using a desolvation-sphere radius of 6.4 Å together with the number of dehydrons using a

Figure 11.6: Reproduction of Fig. 1 in [142]. The vertical axis is the number of dehydrons of a given domain (fold) averaged over all proteins with that domain in the PDB. The horizontal axis is number of interactive partners for a given domain as reported in DIP. For any fixed number of partners, there are different folds represented by the different boxes. The boxes are shaded according to the number of PDB entries with that domain. Black squares correspond to domains with more than 20 PDB entries, dark gray squares correspond to domains containing from 11 to 20 PDB entries, light gray squares correspond to domains from 6 to 10 PDB entries, and empty squares indicate domains with ≤ 5 representatives.

strict requirement of $\rho_G < 12$ (cf. Section 8.4) imposed in the definition of dehydron [145]. A significant correlation was demonstrated in Fig. 1 in [145] between the number of dehydrons and the number of interacting partners for a given domain as reported by DIP. Remarkably, the slope of the regression curve in Fig. 1 in [145], reproduced in Figure 11.6, is very close to one, over a range of interactions and dehyrdons that spans an order of magnitude. The correlation coefficient is 0.88 and and the dispersion is 0.29. Although other factors also lead to protein-protein interaction, this nearly one-for-one relationship between dehydrons and interaction partners suggests that dehydrons

must play a significant role in supporting such interactions. Thus there is a strong indication that the potential interactions related to dehydrons as documented in the examples in Sections 11.1, 11.2 and 11.3 correlate with actual protein interactions.

11.4.2 Structural interactivity

In Section 5.5, we noted that there is a distinctive character to the sturctural relationships among the basic folds of protein tertiary structure, reflected in the probability distribution of these interactions. Nodes with few connections are common, and highly connected nodes are rare, in agreement with a power-law (a.k.a. scale free) distribution as in (5.17) [429]. Since we have seen that dehydrons can play a role in structural composition in Section 11.1.2, it is natural to ask if this would be reflected in the network of sturctural interactivity. Not surprisingly, we will see that connectivity of different domains in this representation is proportional to the average number of dehydrons in the family [145], in Figure 11.7. Furthermore, the dehydron patterns associated with structural domains for a given species define scale-free interactive networks as in (5.17), shown in Figure 11.8. The number of dehydrons in structural domains within a given species determines a distinct characteristic exponent for the power-law distribution that describes the node distribution [9, 35, 145].

Figure 11.7: SCOP families distributed according to their average value r of the quantity $r_{d/HB}$ defined in (11.2) (vertical axis). The horizontal axis represents the interactivity ν of the families.

An analysis similar to Section 5.5 was carried out in [145] based on the Structural Classification of Proteins (or SCOP) superfamilies [269, 304] as vertices in a graph. Interactivity among the superfamilies was defined using a variant of method described in Section 5.5 [429] utilizing protein domains with identified interactions (complexation or intramolecular interaction) in PDB files from different superfamilies [5, 107, 206, 214, 322] to define graph edges.

More precisely, two superfamilies ϕ_1 and ϕ_2 in SCOP are deemed to be interacting (which we write as $\phi_1 \mathcal{I} \phi_2$) if they each contain folds $f_1 \in \phi_1$ and $f_2 \in \phi_2$ that are found in a single PDB file, either within a single protein (in the case of a multidomain protein) or in a protein complex represented in a single PDB file. This provides an equivalence relation $\mathcal{I}$ on fold superfamilies. We can then define an interactivity index ν as the number of fold superfamilies that a given fold superfamily interacts with:

$$\nu(\phi_1) = \text{cardinality}\left(\left\{\phi_2 \in \text{SCOP} \mid \phi_1 \mathcal{I} \phi_2\right\}\right), \tag{11.1}$$

where cardinality means the number of elements of the set. This index is used as the horizontal axis in Figure 11.7.

To account for differences in protein length, an average $r_{d/HB}$ of dehydrons per hundred hydrogen bonds was determined [145], defined as

$$r_{d/HB} = 100 n_{DH}/n_{HB}, \tag{11.2}$$

where n_{HB} is the number of hydrogen bonds in the protein and n_{DH} is the number of dehydrons in the protein, as defined in Table 11.3. For each SCOP superfamily ϕ, define r_ϕ to be the average of $r_{d/HB}$ for all members of ϕ represented in the PDB.

Again, a strong correlation emerged between interactivity and $r_{d/HB}$, as shown in Figure 11.7 (which reproduces Fig. 3a in [145]). In this figure, the horizontal axis is the number ν of interactions, as determined by the number of interactions represented in the PDB when this data is available, and from the database Pfam [151, 111] (which does not require PDB structures) when it is not. (See Section 5.5 for more information on Pfam.) The dark squares in Figure 11.7 represent cases when the PDB-based interactivity definition agreed with Pfam, and the open squares correspond to cases where interactivity in the PDB is under-reported. The numbers such as 1.110.1 in Figure 11.7 indicate the SCOP superfamily classification. The value of the vertical axis in Figure 11.7 used to plot the squares is the average r_ϕ for a given superfamily as indicated by these classifications. The error bars represent the dispersion of $r_{d/HB}$ for all the superfamilies ϕ with that value of ν.

Now let us examine the network of interactions in more detail. For each superfamily ϕ, let f_ϕ denote the percentage of protein (SCOP) superfamilies having the average value of $r_{d/HB}$ approximately equal to r_ϕ. Following Section 5.5, we consider data values (r_ϕ, f_ϕ) separately for different species. More precisely, both the calculation of r_ϕ

Figure 11.8: SCOP families distributed according to their average value r of the quantity $r_{d/HB}$ defined in (11.2) (horizontal axis). The vertical axis represents the percentage f of total number of families with approximatly this value of r. The data represents three different species: H. sapiens (open squares), M. musculus (solid triangles), E. coli (open circles).

and f_ϕ are done by restricting to PDB files specific for the different species. Due to the limitations of data in the PDB, this was possible to do for only three species, E. coli, mouse and human. In Figure 11.8 (reproducing Fig. 3b in [145]), the set of values (r_ϕ, f_ϕ) are plotted on a log-log scale, and a power-law behavior

$$f(r) \approx c r^{-\gamma} \tag{11.3}$$

is seen for each species (E. coli, mouse and human). Not surprisingly, in view of Table 11.3, the values of γ were significantly different for different species (2.1, 1.49, 1.44, respectively).

The results reviewed here present two different views of the correlation of dehydrons and interactivity. On the one hand, interactivity of broad classes of protein with similar folds (including many different species) tends to increase as a function of the average number of dehydrons in that class. Thus for example, we see in Table 11.3 that hemoglobin has far fewer dehydrons than myoglobin, when averaging over all the species represented. However, it is also possible to subdivide the distribution of dehydrons across all proteins according to species, and a differentiation appears in the form (11.3). Although these two statements may seem contradictory at first, they are rather complementary.

11.5 Sheets of dehydrons

It is possible for protein systems to form using only indirect dehydration forces. One example is given by associations of β-sheets [126, 305].

11.6 Exercises

Exercise 11.1 *Examine the PDB file 1AFV in a protein viewer. Identify Tyr 42 in the L chain and the nearby hydrogen bond in the A chain. In addition to wrapping the hydrogen bond, what other interactions could Tyr 42 be having with the A chain? Which way do you think the hydrogen attached to the terminal oxyten in Tyr 42 would be pointing?*

Exercise 11.2 *Scan the PDB for instances of crystal contacts. Plot a distribution of wrapping for hydrogen bonds that have atoms in their desolvation domains from artificial contacts in other chains.*

Exercise 11.3 *Determine the residues most likely to be involved in catalytic activity in the active site of an enzyme.*

Exercise 11.4 *The residue Asp is often involved in catalytic activity in the active site of an enzyme. It is often found to make local sidechain-mainchain bonds (cf. Chapter 15) in an underwrapped environment. Explore the possible correlation of these two observations.*

Chapter 12

Aromatic interactions

Aromatic sidechains are special for many reasons. First of all, they are just big. But they also have a subtle dual role as hydrophobic groups and as acceptors of polar interactions (e.g., as hydrogen bond acceptors [1]). On the one hand, the benzene-like rings in them are largely hydrophobic, and they have very small dipole moments. However, the charge distribution of these rings creates a significant quadrupole moment (Section 10.5), as we will explain.

The center of a surface above each side of the face of the aromatics is negatively charged, and the C-H groups in the ring form positive charge centers [105, 276]. This can be easily visualized since all of the protons lie in the plane of the ring, whereas the electrons are distributed all around them. The fact that the positive charges are in the plane means that there is a net positive charge at points in this plane, with a net negative charge in planes above and below this. More precisely, we should define three thin slabs, as shown in Figure 12.1, one around the plane of the aromatic ring, and two above and below this slab. The upper and lower slabs contain electrons but no protons, thus have a net negative charge. Since the sum of the charges in the three slabs is zero, the middle slab is positively charged. Moreover, the positive charge is twice the charge of each of the other two slabs, so we have a typical arrangement of a quadrupole.
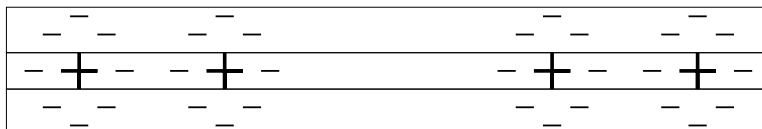
Figure 12.1: Side view of an aromatic ring showing three slabs containing alternating negative and net positive charges, providing a directional quadrupole perpendicular to the face of the ring. The + charges correspond to eight times as much charge as each - charge.
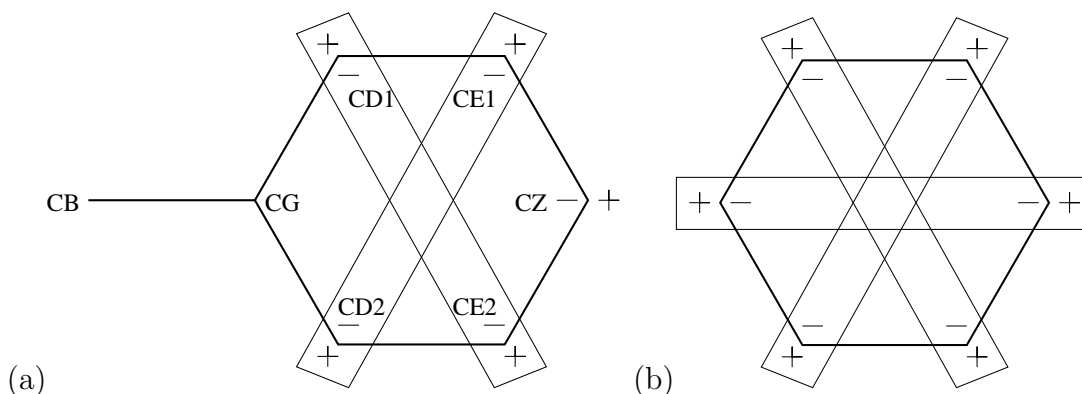
Figure 12.2: Arrangement of partial charges from Table 12.1 for (a) the phenylalanine sidechain (only carbons are shown). Corresponding partial charges are depicted for (b) a benzene ring. The boxes indicate the counterbalancing dipoles which combine to form quadrupoles.

12.1 Partial charge model

Partial charges are frequently used to model aromatics [66, 258] as shown in Table 12.1. However, this model is planar and thus does not directly represent the three-dimensional charge distribution depicted in Figure 12.1. In fact, it replaces this polarity that is orthogonal to the face with one that is within the plane of the aromatic ring.

In Figure 12.5(a), we plot the electrostatic potential corresponding to the partial charges in Table 12.1 in a plane parallel to the plane of the ring at a distance of 1.0Å from the plane of the ring. We do see that the face of the ring has a negative charge, as required.

However, the planar partial charges cause a large polar behavior in the plane of the ring near the locations of the hydrogens. In Figure 12.5(b), we plot the electrostatic potential corresponding to the partial charges in Table 12.1 in a plane near the plane of the ring (at a distance from the plane of the ring of only 0.1Å). In this plane, there is a strong polarity, one that might lead to hydrophilic behavior.

A more sophisticated approximation of the aromatic ring might be to put negative charges at positions near the carbons but in the direction normal to the ring. This improvement would be similar to current models of water, such as Tip5P [279].

12.2 Cation-π interactions

Among pair interactions at interfaces (Section 7.5), the Arg-Trp interaction has the fourth highest log-odds ratio. This pair is an example of what is known as a cation-π pair [45, 88, 164, 290, 329, 337, 351, 420, 431, 435]. It has a strength comparable to that of a hydrogen bond. It is based on an interaction between the negative charge on the face of aromatic residues and positively charged (cation) residues (Lys, Arg, His). The cation-π motifs play a special role in protein interfaces [88, 431]. The cation-π interaction also has a significant role in α-helix stabilization [369].

Residue	atom type	PDB code	charge
PHE	C	CDi, CEi, $i = 1, 2$, CZ	-0.1
	HC	HDi, HEi, $i = 1, 2$, HZ	0.1
TYR	C	CDi, CEi, $i = 1, 2$	-0.1
	HC	HDi, HEi, $i = 1, 2$	0.1
	C	CZ	0.15
	OA	OH	-0.548
	H	HH	0.398
TRP	C	CG	-0.14
	C	CD1, CE3, CZi, $i = 2, 3$, CH2	-0.1
	HC	HD1, HE3, HZi $i = 2, 3$, HH2	0.1
	NR	NE1	-0.05
	H	HE1	0.19

Table 12.1: Partial charges from the Gromos force field for aromatic amino acids; cf. also [258].

Cation-π interactions can take place with other residues, such as a phosphorylated tyrosine, cf. Figure 12.3. The aromatic ring still provides a distributed negative charge, unaffected by the addition of the phosphate group. And the two-sided nature of the polarity of an aromatic ring means that it can interact with two cations at one time, one on each side. This is depicted in Figure 12.3(a) which shows a phosphorylated tyrosine flanked by an arginine and a lysine in the SH2 domain in the PDB file 1JYR. SH2 domains [265] specifically recognize phosphorylated tyrosines and bind proteins containing them, and the cation-π interaction is presumably important in the binding process.

Correspondingly, a cation could be sandwiched between two aromatics, as shown in Figure 12.3(b). The arginine (A67) interacts with both a phosphorylated tyrosine (PTR-I3) and a phenylalanine (Phe-I4). In addition, Arg-A67 is hydrogen bonded with the backbone oxygen on Phe-I4 and one of the terminal oxygens on PTR-I3.

Finally, we show a cation-π grouping involving a complex of two cations and two aromatics, as shown in Figure 12.4. The arginine is interacting with both of the aromatics, whereas the lysine is interacting only with the phosphorylated tyrosine.

12.3   Aromatic-polar interactions

Aromatic rings can in principle interact with any polar moiety [60, 61, 258]. We will consider different classes of polar interactions. The most familiar will be a type of hydrogen bond, in which the face of the aromatic ring forms the acceptor of a hydrogen bond. Cation-π interactions can take place with the NH groups on Asn and Gln [45]. However, there are also other types of interactions that appear to be possible.
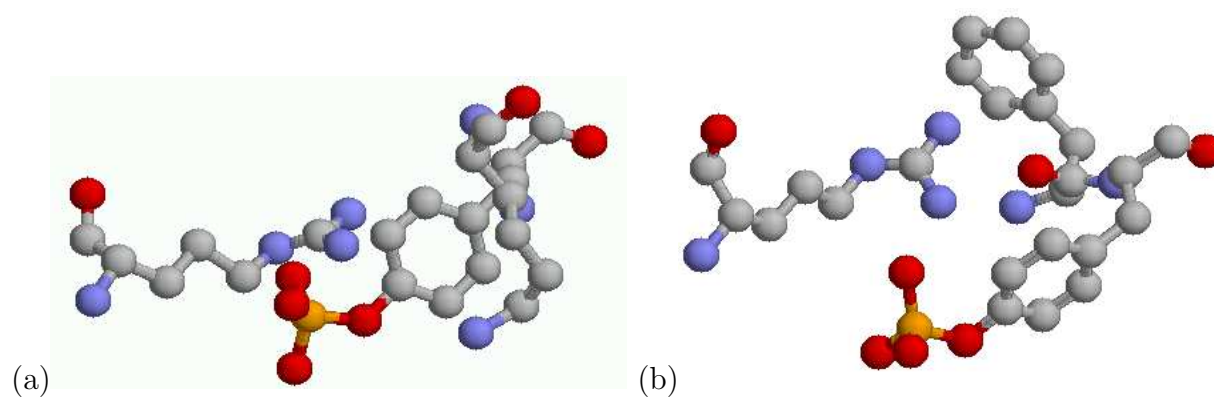
Figure 12.3: Cation-π groups found in the SH2 domains in the PDB files (a) 1JYR and (b) 1BMB. (a) Shown are a phosphorylated tyrosine (PTR1003) flanked by Arg67 (upper left) and Lys109 (lower right). (b) Shown are a phosphorylated tyrosine (PTR-I3, lower right) and a phenylalanine (Phe-I4, upper right) flanking an Arg-A67 (middle left).
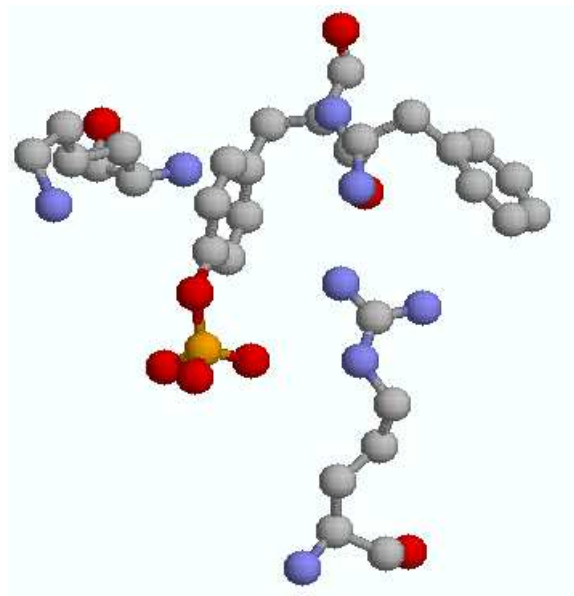
Figure 12.4: Cation-π group found in the SH2 domain in the PDB file 1TZE. Shown are a phosphorylated tyrosine (PTR-I4, middle left) and a phenylalanine (Phe-I3, upper right) together with Arg-E67 (lower right) and Lys-E109 (upper left).

(a)
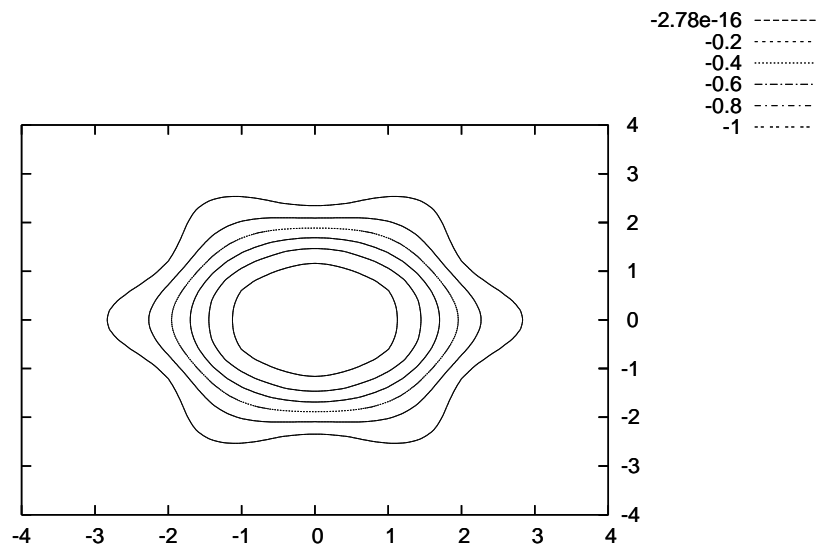
(b)
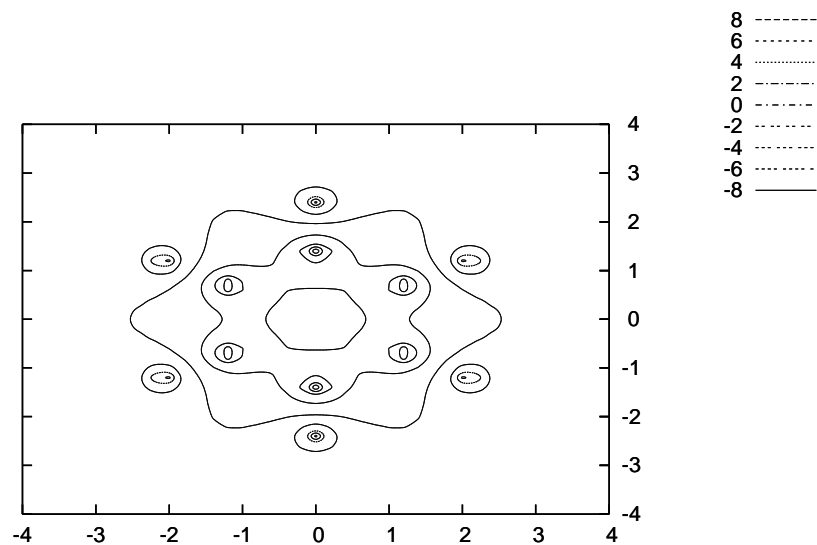
Figure 12.5: Electric potential of the partial-charge model of a benzene ring [258] in parallel planes above the ring, using the partial charges in Table 12.1: (a) 1.0 Ångstrom above and (b) 0.1 Ångstrom above.
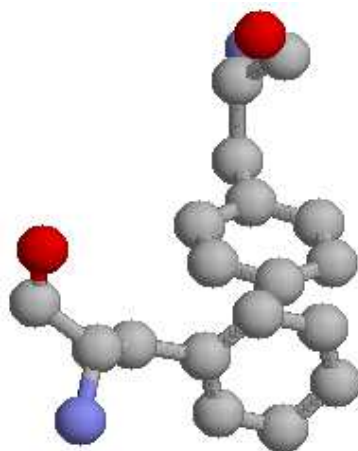
Figure 12.6: Pairwise interactions of two phenylalanines (117 and 153) in a mutant (Ser117 → Phe) of T4 lysozyme detailed in the PDB file 1TLA [6].

12.3.1   Aromatics as acceptors of hydrogen bonds

In [209], a hydrogen bond with the polar face of an aromatic ring is described (in Figure 6). The aromatic face can interact with typical hydrogen bond donors [64, 258, 330, 380, 381, 386]. One criterion for such a hydrogen bond involves the distance from the center of the aromatic face, with the angle formed between the donor and the aromatic face also taken into account (cf. Figure 1 in [258]). See Exercises 12.3 and 12.4 for a further exploration of interactions of this type.

12.4   Aromatic-aromatic interactions

It is possible for two aromatics to interact via more complex interactions, including edge-to-face interactions [68, 253, 432]. In Figure 12.6 the relative orientations of two phenylalanines is shown [6] as found in the PDB file 1TLA. A vector along the line formed by the C_γ—C_ζ carbons in Phe117 (upper) is pointing toward the face of Phe153.

The interaction depicted in Figure 12.6 is consistent with a model of Phe in which there is a small positive partial charge, or positive polarity, at the end of the sidechain, that is near the C_ζ carbon. The partial charge model in Table 12.1 is precisely of that type. There is a dipole formed between C_ζ and H_ζ, with no opposing dipole related to C_γ. The other dipoles in the ring are always counter-balanced, forming quadrupoles. Thus there is a net dipolar behavior to the carbonaceous ring of Phe in the partial charge model in Table 12.1, consistent with the Phe-Phe interaction in Figure 12.6.

See Exercises 12.5 and 12.6 for a further exploration of interactions of this type.
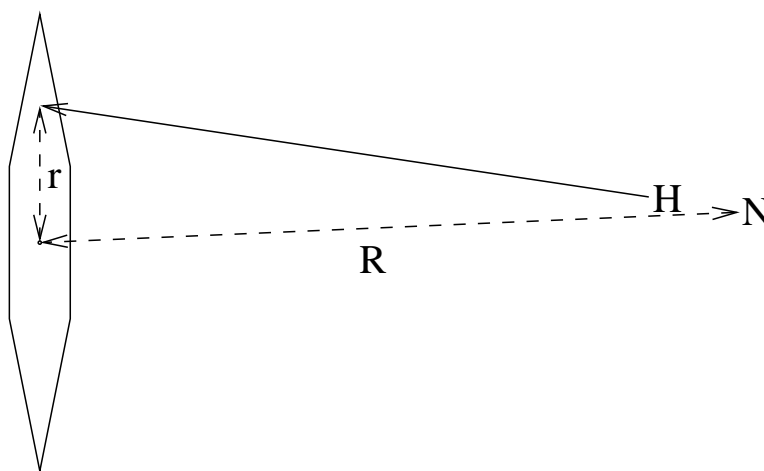
Figure 12.7: Coordinate definitions for Exercise 12.3.

12.5 Exercises

Exercise 12.1 *Are the aromatic rings planar? Three points determine a plane, but there are six points in an aromatic ring. Determine the plane that best fits six points in three dimensions by least squares. Test this on Phe, Tyr and Trp in several PDB files. How big is the deviation of the six points from the plane (compute the squared deviation)?*

Exercise 12.2 *Where is the center of an aromatic ring? The three lines connecting opposite carbons shown in Figure 12.2 should intersect at the center of the plane formed by the aromatic ring. The midpoint of each line segment should be at the center of the ring. But if the carbon positions do not lie in a plane, or if the hexagon is not regular, these lines may not intersect. Determine an algorithm for finding a point that best approximates the near intersection of these three lines. For example, you could just average the three midpoints of each of the above line segments. Does this point lie on the plane determined in Exercise 12.1? Test this on Phe, Tyr and Trp in several PDB files. For each aromatic ring, give the sum of squares of distances from each line-segment midpoint to the center point you determine.*

Exercise 12.3 *Look for instances separately for Phe, Tyr and Trp forming acceptors for hydrogen bonds where the hydrogen bond donors are either mainchain (NH) donors or NH groups on Asn and Gln. For each aromatic, look for the closest candidate hydrogen bond donor. Plot the distance from the nitrogen donor to the center of the aromatic face (Exercise 12.2) and the distance from this center point to the point of intersection of the vector generated by NH to the plane (Exercise 12.1) of the aromatic. See Figure 12.7 for a depiction of these coordinates. Use (6.1) to determine hydrogen placement for the mainchain donors, and use (6.2) and (6.3) for the sidechains. Use a large subset of the PDB, such as the non-redundant PDB, for your experiments. Give the distributions for the mainchain and sidechain donors separately, as well as for each aromatic sidechain. (This means six different distributions.)*

Exercise 12.4 *Catalog all hydrogen bonds having aromatics as an acceptor in the PDB. Determine the distribution of distances from the donor to the face of the other aromatic, as well as the angles formed between the donor, hydrogen and the plane of the aromatic. Compare this with the two distances used in Exercise 12.3.*

Exercise 12.5 *Investigate the interactions of two phenylalanine residues as depicted in Figure 12.6. Determine the distribution of distances from the terminal carbon of one of the residues to the face of the other. Also record the angle between the two faces.*

Exercise 12.6 *Investigate interactions of a phenylalanine residue and another aromatic (Tyr or Trp) similar to the Phe-Phe interaction depicted in Figure 12.6. Determine the distribution of distances from the terminal carbon of one the Phe residues to the face of the other aromatic. Also record the angle between the two faces.*

Exercise 12.7 *Investigate the Trp-Pro 'sandwich' [343] by examining the geometry of Trp-Pro pairs that appear in PDB files in close proximity.*

Exercise 12.8 *Investigate the interaction geometry for all planar groups in sidechains [70].*

Chapter 13

Peptide bond rotation

We now consider an application [122] of data mining that links the quantum scale with the continuum level electrostatic field. In other cases, we have considered the modulation of dielectric properties and the resulting effect on electronic fields. Here we study the effect of the local electronic field on a particular covalent bond. Wrapping by hydrophobic groups plays a role in our analysis, but we are interested primarily in a secondary effect of the interaction.

Typically, changes in the dielectric environment will have no direct effect on a covalent bond. However, such changes can effect the local electric field, and in many cases this can change the covalent bond structure. One example is the effect of the protonation state of a His sidechain, in which two different states are possible as indicated in Figure 4.9. Another has to do with the peptide bond itself. We consider modulations in the local electric field that cause a significant change in the electronic structure of the peptide bond and lead to a structural change in the type of the covalent bond.

We first describe the results and then consider their implications for protein folding.

13.1 Peptide resonance states

The peptide bond is characterized in part by the planarity [324] of the six atoms shown in Figure 13.1. The angle ω (see Section 5.2.1) quantifies the orientation of this bond, with planarity corresponding to $\omega = \pi$ radians in the trans case as depicted in Figure 4.2(a), or $\omega = 0$ in the cis form as depicted in Figure 4.2(b).

It has been known for some time that the variation in ω is much greater than the variation in other parameters describing the peptide structure [72, 114, 207, 277]. It has been known [324] even longer that the planar state is not the preferred vacuum state. What determines the variation in ω is the local electronic environment [122], as we will review here.

The peptide bond is what is known as a **resonance** [324] between two states, shown in Figure 13.1. The "keto" state (A) on the left side of Figure 13.1 is actually the preferred state in the absence of external influences [324]. However, an external polarizing field can shift the preference to the "enol" state (B) on the right side of Figure 13.1, as we illustrate in Figure 13.2 [324].

We have already seen several examples of atomic groups whose electronic structure is a resonance
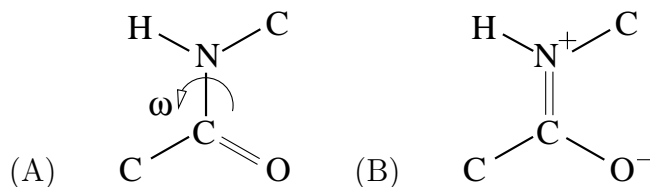
Figure 13.1: Two resonance states of the peptide bond. The double bond between the central carbon and nitrogen keeps the peptide bond planar in the right state (B). In the left state (A), the single bond can rotate, cf. Figure 5.9(b).

between two distinct states. The C-C bonding structure in benzene, which is similar to the aromatic rings in Phe, Tyr and Trp, is a combination of single and double covalent bonds. Similarly, the allocation of charge in the oxygens in the symmetric ends of Asp and Glu represents a resonance, which includes the bond configuration with their terminal carbons. The guanadine-like group at the end of the Arg sidechain represents an even more complex type of resonance, in that the positive charge on the NH_2 groups is shared as well by the NH group in the ϵ position.

Resonance theory assumes that a given (resonant) state ψ has the form

$$\psi = C_A\psi_A + C_B\psi_B, \tag{13.1}$$

where ψ_X is the electronic wavefunction (eigenfunction of the Schrödinger equation) for state X ($=A$ or B). Since these are eigenstates normalized to have L^2 norm equal to one, we must have $C_A^2 + C_B^2 = 1$, assuming that ψ_A and ψ_B are orthogonal, as is the assumption for the ideal case of resonance theory. We will take this as the basis for the following discussions.

The resonant state could be influenced by an external field in different ways, but the primary cause is hydrogen bonding. If there is a hydrogen bond to either the carbonyl or amide group in a peptide bond, this induces a significant dipole which forces the peptide bond into the (B) state shown in Figure 13.1. Such hydrogen bonds could be either with water, with sidechains, or with other backbone donor or acceptor groups, and such a configuration is indicated in Figure 13.2 in the representation (1) on the left side of the figure. An acceptor for the amide (NH) group is indicated by $\cdots$O, and a donor for the carbonyl (CO) group is indicated by H$\cdots$. Other acceptors could also be involved instead.

On the right side of Figure 13.2, in the representation (2), the polar environment is indicated by an arrow from the negative charge of the oxygen partner of the amide group in the peptide bond to the positive charge of the hydrogen partner of the carbonyl group. The strength of this polar environment will be less if only one of the charges (that is, only one of $\cdots$O or H$\cdots$, or equivalent) is available, but either one (or both) can cause the polar field.

However, if neither type of hydrogen bond is available, then the resonant state moves toward the preferred (A) state in Figure 13.1. The latter state involves only a single bond and allows ω rotation. Thus the electronic environment of peptides determines whether they are rigid or flexible. Since any hydrogen bond can enable the (B) state, it will prevail whenever a water molecule or a mainchain or sidechain donor or acceptor is appropriately located. Otherwise said, the (A) state persists only when water is removed and there are no other binding partners.
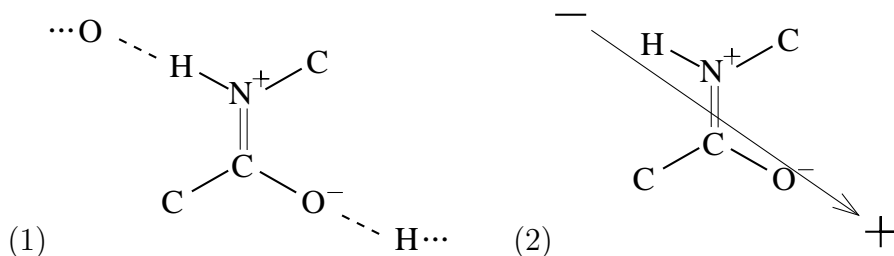
Figure 13.2: The resonance state (B) of the peptide bond shown, on the left (1), with hydrogen bonds (dashed lines) to an acceptor (indicated by the oxygen preceded by dots) and a donor (indicated by the nitrogen followed by dots). These hydrogen bonds induce a polar field that reduces the preference for state (A). On the right (2), the (B) state is depicted with an abstraction of the dipole electrical gradient induced by hydrogen bonding indicated by an arrow.

13.2 Measuring variations in ω

There is no simple way to measure the flexibility of the peptide bond from typical structural data. If flexibility were linked with mobility of the ω bond, then it could potentially be inferred from the 'fuzziness' of the electron densities, e.g., as measured by the B-factors reported in the PDB. But such observations might be attributed to other factors, and there is no reason to believe that flexibility in ω would necessarily mean that the angle adopted would not be fixed and well resolved.

For any given peptide bond, the value of ω could correspond to the rigid (B) state even if it is fully in the (A) state. The particular value of ω depends not only on the flexibility but also on the local forces that are being applied. These could in principle be determined, but it would be complicated to do so. However, by looking at a set of peptide bonds, we would expect to see a range of values of ω corresponding to a range of local forces. It is reasonable to assume that these local forces would be randomly distributed in some way if the set is large enough. Thus, the *dispersion* Δ_ω for a set of peptide bond states would be proportional to the flexibility of the set of peptide bonds.

The assessment of the flexibility of the peptide bond requires a model. We assume that the amount of rotation ω around the C-N bond is proportional to the applied force f, that is, $\omega - \omega_0 = \kappa f$, where we think of κ^{-1} as representing the strength of the bond to resist rotations, and $\omega_0 = \pi$ corresponds to the planar configuration. If $\kappa = 0$, then the bond is infinitely stiff. Thus if we have a set of rotations, due to a set of forces with dispersion Δf, then $\Delta \omega = \kappa \Delta f$.

Suppose that we assume that the bond adopts a configuration that can be approximated as a fraction C_B of the rigid (B) state and corresponding the fraction C_A of the (A) state. Let us assume that the flexibility of the peptide bond is proportional to C_A, that is, the 'spring constant' κ of the bond depends linearly on the value of C_A: $\kappa = \kappa_0 CA$. Thus we are assuming that the (B) state is infinitely rigid. For a set of bond configurations, we thus obtain $\Delta \omega = \kappa_0 CA\Delta f$ relating the dispersion in ω's to the dispersion in applied forces f.

Let us summarize the conclusions of the model. We suppose that the peptide bond is subjected to a set of forces. Some of these forces will leave ω in the same value as for the rigid state. However, others will modify ω, and the extent of the modification will be proportional to the 'spring constant'

κ of the bond and thus proportional to the value of C_A. Thus we can assert that the dispersion Δ_ω in ω is proportional to C_A:

$$\Delta_\omega = \gamma C_A, \tag{13.2}$$

where γ is a constant of proportionality.

The value of γ can be estimated [122] based on the observation that the value $C_B = 0.4$ [324] corresponds to the vacuum state of the peptide bond. In the peptide data, this state is approached with fully desolvated peptide groups that form no hydrogen bonds. Thus the constant will be determined as a by-product of the analysis. We are assuming that the resonant state ψ has the form (13.1), with $C_A^2 + C_B^2 = 1$. Therefore

$$\Delta_\omega = \gamma C_A = \gamma \sqrt{1 - C_B^2}. \tag{13.3}$$

We can now see how to determine γ. Suppose that Δ_ω^∞ is the observed value of dispersion in ω for fully desolvated peptide groups that form no hydrogen bonds. Then we must get $C_B = 0.4$ (and thus $C_A \approx 0.917$), which means that

$$\gamma \approx 1.09 \Delta_\omega^\infty. \tag{13.4}$$

We can thus invert the relationship (13.3) to provide C_B as a function of Δ_ω:

$$C_B = \sqrt{1 - (\Delta_\omega/\gamma)^2}. \tag{13.5}$$

The assumption (13.2) can be viewed as follows. The flexibility of the peptide bond depends on the degree to which the central covalent bond between carbon and nitrogen is a single bond. The (A) state is a pure single bond state and the (B) state is a pure double bond state. Since the resonance state is a linear combination, $\psi = C_A\psi_A + C_B\psi_B$, it follows that there is a linear relationship between C_A and flexibility provided that the single bond state can be quantified as a linear functional. To prove this relationship, we seek a functional L_A such that $L_A\psi_A = 1$ and $L_A\psi_B = 0$. If ψ_A and ψ_B are orthogonal, this is easy to do. We simply let L_A be defined by taking inner-products with ψ_A, provided (as we assume) that ψ_A and ψ_B are orthogonal.

13.3   Predicting the electric field

Since the preference for state (A) or (B) is determined by the local electronic environment, the easiest way to study the flexibility would be to correlate it with the gradient of the external electric field at the center of the peptide bond. However, this field is difficult to compute precisely due to the need to represent the dielectric effect of the solvent (Chapter 16) and to account for the polarizability of all the molecular groups. A dynamic simulation with an explicit water model and full representation of polarizability might be able to correctly estimate this accurately, but this is beyond current technological capability. However, it is possible to estimate the likelihood of a significant dipole moment based on the local environment [122].

The major contributor to a local dipole would be the hydrogen bonding indicated in Figure 13.2. These bonds can arise in two ways, either by backbone hydrogen bonding (or perhaps backbone-sidechain bonding, which is more rare) or by contact with water. The presence of backbone hydrogen

bonds is indicated by the PDB structure, but the presence of water is not consistently represented. However, there is a a proxy for the probability of contact with water: the wrapping of the local environment. So we can approximate the expected local electric field by analyzing the backbone hydrogen bonding and the wrapping of these amide and carbonyl groups.

In [122], sets of peptide bonds were classified in two ways. First of all, they were separated into two groups, as follows. Group I consisted of peptides forming no backbone hydrogen bonds; in particular, this group excludes ones involved in either α-helices or β-sheets. Group II consisted of peptides forming at least one backbone hydrogen bond, thus including ones involved in either α-helices or β-sheets, but also including other residues which can be identified as forming hydrogen bonds in the PDB structure. Groups I and II are further subdivided according to wrapping. We will describe this in two steps, starting with a simple approach and then giving the full approach developed in [122].

A simplistic way of viewing the data is as follows based on dividing groups I and II into the underwrapped subgroups (a) and the well wrapped subgroups (b). We thus consider four groups:

- Group Ia consists of peptides forming no backbone hydrogen bonds and the amide and carbonyl groups are not well wrapped,

- Group Ib consists of well wrapped peptides forming no backbone hydrogen bonds,

- Group IIa consists of peptides capable of forming a backbone hydrogen bond but not well wrapped, and

- Group IIb consists of well wrapped peptides forming a backbone hydrogen bond.

Then we can estimate the electronic environment, and its impact on the resonant state, as depicted in Table 13.1. We make the simple assumption that well wrapped peptides will not have water around, but underwrapped ones will be solvated. With this assumption, we can classify the different groups according to whether or not water will be available to form hydrogen bonds with the peptide donor or acceptor. Thus in Table 13.1 we divide into four possibilities as above and assign the resonant state appropriate for the corresponding electronic environment. We see that only Group Ib can be in the A state.

Group number number of HB's	underwrapped waters (a)	well wrapped no water (b)	symbol used in Figure 13.3
Group I 0 HB's	$\Delta\omega \approx 0$ B state	$\Delta\omega >> 0$ A state	squares
Group II ≥ 1 HB's	$\Delta\omega \approx 0$ B state	$\Delta\omega \approx 0$ B state	x's

Table 13.1: Simple model of local electrostatic environment around a peptide bond and the resulting preferred resonant state.

Since wrapping can be defined using a count of nonpolar carbonaceous groups, it is possible [122] to give a more refined analysis than indicated in Table 13.1. In each major group (I and II), subsets can be defined based on the level ρ of wrapping in the vicinity of backbone [122]. For each subgroup, the dispersion of ω angles can be measured in the corresponding PDB files, and this data is plotted Figure 13.3 as a function of ρ. Figure 13.3 presents an even more refined analysis. It involves groups Iρ and IIρ for different values of ρ.

The left side of Figure 13.3 corresponds to Group Ia and Group IIa, and we see that the behavior is the same for the two groups. That is, the underwrapped peptides have a similar dispersion Δ_ω, corresponding to a dominant (B) state, whether or not they appear to be capable of hydrogen bonding, although the dispersion is decreasing with increasing wrapping. On the other hand, there is a difference between the Ib and IIb groups, as indicated on the right side of Figure 13.3. Group Ib prefers the vacuum state (A), whereas IIb tends to the (B) state.

Figure 13.3 depicts the resulting observations for group I peptide bonds (small squares), using the model (13.5) to convert observed dispersion $\Delta_\omega(\rho)$ to values of C_B, with a constant γ as given in (13.4). This value of γ was determined using the estimated value $C_B = 0.4$ [324] for the vacuum state of the peptide bond which is approached as ρ increases, in the absence of mainchain hydrogen bonds. Well wrapped peptide bonds that do not form hydrogen bonds should closely resemble the vacuum state. However, poorly wrapped peptide amide and carbonyl groups would be strongly solvated, and thus strongly polarized, leading to a larger component of (B) as we expect and as Figure 13.3 shows.

Using the value (13.4) of γ allows an interesting assessment of the group II peptide bonds, as shown in Figure 13.3 (data represented by x's).

Figure 13.3: Portion of the double-bond (planar) state in the resonance for residues in two different classes, adapted from [122]. The horizontal axis is the number of nonpolar carbonaceous groups inside two spheres of radius 6 Ångstroms centered at the midpoint of the peptide bond, cf. Section 8.4. The vertical axis is the quantity C_B indicating the prevalance of the (B) state. Small squares indicate the case when neither amide nor carbonyl group is engaged in a backbone hydrogen bond (group I). The x's indicate the case when at least one of the amide or carbonyl groups is engaged in backbone hydrogen bond (group II). The dotted line represents a hypothetical response due only to the dehydration of the mainchain bonds.

These are bonds that, according to the PDB structures, are capable of participating in backbone hydrogen bonds. We see that these bonds also have a variable resonance structure depending on the amount of wrapping. Poorly wrapped backbone hydrogen bonds will likely be solvated, and thus the group I and group II peptides can be expected to behave similarly for small ρ. As with group I peptide bonds, we expect state (B) to be dominant for small ρ. Indeed, the two curves in Figure 13.3

(the squares and the x's) are quite similar for small ρ. As dehydration by wrapping improves, the polarity of the environment due to water decreases, and the proportion of state (B) decreases. But a limit occurs in this case, unlike with group I, due to the fact that wrapping now enhances the strength of the backbone hydrogen bonds, and thus increases the polarity of the environment.

The interplay between the decreasing strength of polarization due to one kind of hydrogen bonding (with water) and the increasing strength of backbone hydrogen bonding is quite striking. As water is removed, hydrogen bonds strengthen and increase polarization of peptide bond. Figure 13.3 shows that there is a middle ground in which a little wrapping is not such a good thing. That is, small amounts of wrapping appear to remove enough water to decrease the polar environment. Moreover, with minimal wrapping, the backbone hydrogen bonds are screened, and therefore the resulting external polar environment of the peptide is weaker. But the effect of the hydrogen bonds increase as wrapping is increased. The solid line in Figure 13.3 represents a guess of the effect of wrapping of the backbone bonds alone.

It is striking that the group II data has a distinctive minimum. One might guess that there would be constant polarity in the transition from fully solvated peptides to fully desolvated peptides. But apparently there are two distinct behaviors. The fully solvated states appear to provide a strong dipole through the contact with water in a nearly bulk-like state. As wrapping is added, this polarity is disrupted as water becomes both excluded and disordered by the hydrophobic groups. On the other hand, the backbone hydrogen bonds can also be disrupted by just a few water molecules. Thus it takes a large amount of wrapping to be establish a stable polar environment. Thus it is not just a simple exchange of one type of polarity for another. There are two different mechanisms. One has to do with the structure of water, and the resulting ability of water to establish a consistent polar environment around a peptide base. The other has to do with the requirements of backbone donors and acceptors to form stable attachments with other parts of the protein structure, unaffected by the presence of water. It is not surprising that these two disparate mechanisms would operate on their own scales and thus not cancel each other as the amount of wrapping is varied.

13.4   Implications for protein folding

After the "hydrophobic collapse" [436] a protein is compact enough to exclude most water. At this stage, few hydrogen bonds have fully formed. But most amide and carbonyl groups are protected from water. The data in Figure 13.3(a) therefore implies that many peptide bonds are flexible in final stage of protein folding. This effect is not included in current models of protein folding. This effect buffers the entropic cost of hydrophobic collapse in the process of protein folding.

New models need to allow flexible bonds whose strengths depend on the local electronic environment [346]. Typical molecular dynamics (MD) models would either have peptide bonds fixed in the planar configuration or have a large spring constant for rotation in the ω angle. Here we need the spring constant to depend on the gradient of the electric field in the vicinity of the peptide bond.

The gradient of the electric field at a point $\mathbf{r}$ is given by

$$\sum_k q_k \frac{\mathbf{r} - \rho_k}{|\mathbf{r} - \rho_k|^3}, \tag{13.6}$$

cf. (16.81). In particular, the quantity of interest is the strength of the dot product of the electric field gradient and the vector $\overline{O-H}$ pointing from O to H in the peptide group. If $\mathbf{r}_0$ is the centroid of the peptide group, then one would seek a bending strength depending on the quantity

$$\sum_k q_k \frac{(\mathbf{r}_0 - \rho_k) \cdot \overline{O-H}}{|\mathbf{r}_0 - \rho_k|^3}. \tag{13.7}$$

Note that we are invoking a sum over all (charged) atoms in the system, and this type of global term will make the simulation much more costly. Using a cut-off radius to limit the number of charged atoms involved may be practical, but it introduces an approximation into the model whose effect would have to be assessed.

13.5 Exercises

Exercise 13.1 *The dipole vector is twice as strong if both amide and carbonyl groups are involved in hydrogen bonds. Split group II into two groups, group II1 and II2 depending on the number of hydrogen bonds. How does the preference for the (B) state differ between groups II1 and II2?*

Exercise 13.2 *Scan some PDB files and form the groups Ia, Ib, IIa, IIb indicated in Section 13.3. Plot the distributions of ω angles for each group. Do the distributions for Ia and IIa look similar? How are the distributions for Ib and IIb different?*

Exercise 13.3 *Using a model of the dielectric effect (cf. Chapter 16), estimate the dipole vector at each peptide in a set of PDB files. Use this estimate to predict whether the peptide bond is in the (A) state or the (B) state. Compare this with the measured value of ω.*

Exercise 13.4 *Using a molecular dynamics model with explicit water, estimate the dipole vector at each peptide in a set of PDB files. Use this estimate to predict whether the peptide bond is in the (A) state or the (B) state. Compare this with the measured value of ω.*

Exercise 13.5 *Using a quantum chemistry model, calculate the flexibility of the ω bond as a function of an imposed dipole as indicated in Figure 13.2.*

Chapter 14

Dehydron patterns

We saw in Figure 2.2 that the number of dehydrons varies substantially among evolutionarily related proteins. Here we explore this variation in more detail in two important cases. One is dehydron variation in antibody complexes. The other is dehydron patterns in tyrosine kinases.

14.1 Comparing F10.6.6 and D44.1 antibodies

In the paper [69], the structures of an antibody in two different maturation states were compared. Here we review their analysis. We find some differences in interpretation of the changes in bonding structure. Our main conclusion that differs substantially from [69] is that the change in strength of the individual bonds can account for the difference in binding affinity of the two complexes.

The structural comparisons are based on the PDB structures 1P2C and 1MLC. They pertain to antibodies denoted by F10.6.6 and D44.1, respectively. Each PDB file contains two, asymmetric copies of the antigen-antibody complexes. These provide some redundancy in the analysis, but as we will see they tend to be somewhat contradictory in certain details. Thus one is not sure which one of the asymmetric copies is native, and perhaps neither represents the state achieved in a context different from the process of crystallization, cf. Section 11.1.3.

The chains in these files are denoted by different descriptors. The light chains in D44.1 are denoted by A,C in 1MLC, whereas in F10.6.6 they are denoted by A,D in 1P2C. The heavy chains in D44.1 are B,D and in F10.6.6 are B,E. The HEL chains (C,F) are numbered 6xy in the 1P2C PDB file, and this corresponds to xy in the HEL chain (E,F) in the 1MLC PDB file. This is summarized in Table 14.1 for reference.

PDB file/antibody	Light	Heavy	HEL
1P2C/F10.6.6	A,D	B,E	C,F
1MLC/D44.1	A,C	B,D	E,F

Table 14.1: Naming conventions for different PDB files for closely related antigen-antibody structures.

PDB	Type	Donor	Acceptor	Distance
1P2C	M-M	F-Arg 1521 N	A-Asp 151 O	2.85
1P2C	S-M	A-Ser 153 OG	F-Asn 1519 O	2.55
1P2C	M-S	F-GLY1522 N	A-Ser 153 OG	2.60
1P2C	S-S	A-Arg 188 NH2	F-Asp 1518 OD1	2.78
1MLC	S-S	F-Arg 128 NE	A-Asn 76 OD1	2.66
1MLC	S-S	F-Arg 128 NH2	A-Asn 76 OD1	2.74

Table 14.2: Nonphysical crystal artifacts in PDB files 1P2C and 1MLC. Hydrogen bonds between the A and F chains, which are in different copies of the antibody-antigen systems.

type	donor : acceptor	F10.6.6(1)	F10.6.6(2)	D44.1(1)	D44.1(2)
M-M	hel-Gly649-N : L-Gly/Asn92-O	2.88	2.90	2.83	not found
S-M	hel-Ser681-OG : H-Thr330-O	3.02	2.77	not found	2.97
S-S	H-Trp333-NE1 : hel-Tyr653-OH	2.84	2.77	2.87	2.91
S-S	hel-Thr643-OG1 : H-Ser357-OG	2 WB	2 Ws	3.48	not found
S-M	H-Ser357-OG : hel-Gln641-O	1 WB	3.31	not found	not found

Table 14.3: Intermolecular hydrogen bonds between antibody and antigen. S=sidechain, M=mainchain. W indicates a water molecule hydrogen bonded instead of the S or M hydrogen bond; B indicates that an apparent water bridge exists. The numbers given are the distance in Ångstroms between the donor and acceptor (heavy) atoms in the hydrogen bond. The notation L-Gly/Asn92 in the mainchain-mainchain bond with hel-Gly649 refers to the alignment D44.1/Asn92 → F10.6.6/Gly92.

Thus the interfaces of interest in 1MLC are AE,BE,CF,DF; the interfaces of interest in 1P2C are AC,BC,DF,EF. However, one also finds significant other interactions that cross between the two copies of the antigen-antibody structures. This accounts in part for the asymmetry of the two copies.

The definition of hydrogen bonds is subjective to a certain extent. Thus it is natural that one definition would provide a different set from another. For example, in [69], the mainchain-sidechain hydrogen bond in 1MLC, E-Thr 47-O — A-Asn 92-ND2, is claimed between HEL and the light chain. However, the angle between the N-H vector and the C-O vector is quite large, about 116 degrees. On the other hand, there is a water (HOH1601) in 1MLC which provides a plausible link between E-Thr 47 and A-Asn 92, although more likely to A-Asn 92-OD1, as shown in Figure 14.1(a).

The number of hydrogen bonds that pass a rigorous test is actually quite small. These are enumeriated in Table 14.3. For the most part, the number of intramolecular hydrogen bonds for F10.6.6 and D44.1 are the same, except that some hyrogen bonds in one structure appear as water-mediated bonds in the other. One can argue that F10.6.6 has more, or better, hydrogen bonds to HEL that D44.1 does, although the advantage is not overwhelming. But the two S-M bonds in F10.6.6 do not have comparable analogues in D44.1.
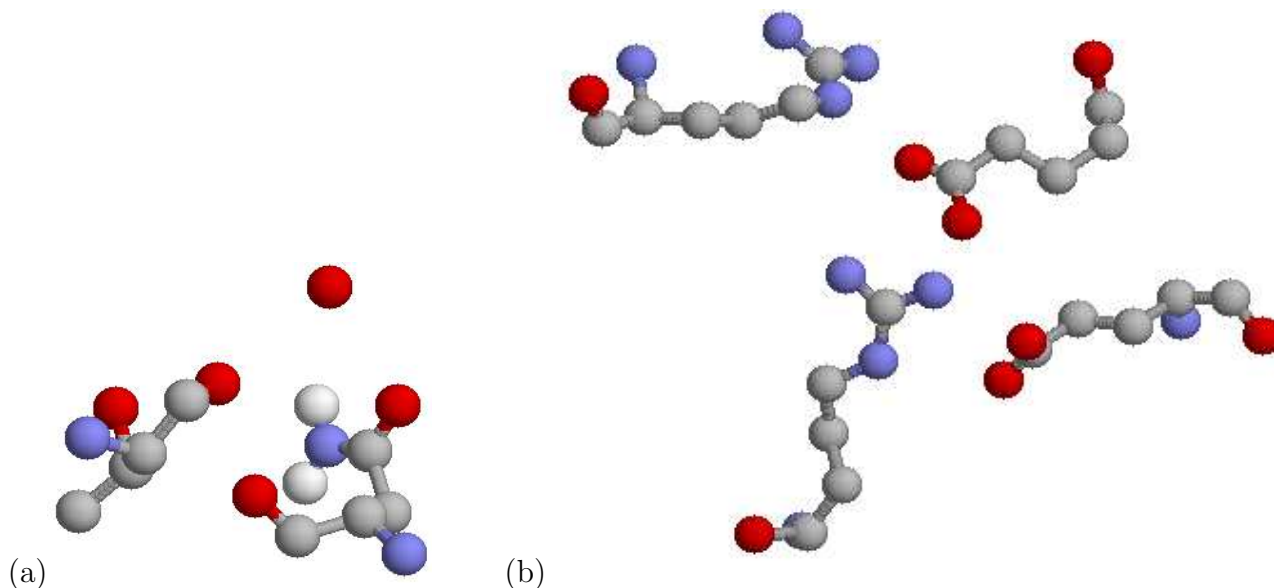
Figure 14.1: (a) Configuration of E-Thr 47 (left), A-Asn 92 (right), and HOH1601 (single red dot at top) in 1MLC. The hydrogen positions on the terminal N in Asn have been determined synthetically. Configuration of the double salt bridge in the PDB file 1P2C. Note that the positive and negative charges are nearly coplanar, and closely approximate opposing corners of a rhombus, a configuration consistent with a quadrupole.

The mutation of D44.1/Asn92 to F10.6.6/Gly92 has only a minor effect, since the mainchain-mainchain bond hel:Gly49-light:Asn92 in D44 becomes the mainchain-mainchain bond hel:Gly49-light:Gly92 in F10. The angular arrangement in F10.6.6 is slightly better than D44.1, whereas the bond distance is comparable.

In the heavy chains, there is a mutation of D44.1/Ser330 to F10.6.6/Thr330. Correspondingly, the strong sidechain-mainchain bond between hel-Ser81 and H-Thr330 in the F10.6.6 complex replaces a rather weak bond in the D44.1 complex involving H-Ser330, and it appears in only one copy. Suprisingly, B-Ser30 (in the first copy of the complex in 1MLC) does not form any sort of bond with other atoms.

There is a hydrogen bond that appears in the D44.1 complex that does not have a direct analogue in the F10.6.6 complex. It is between hel-Thr643 and H-Ser357, a residue that is preserved in both complexes. However, there is a neighboring mutation D44.1/H-Gly356 to F10.6.6/H-Asp356 which presumably affects the local environment. The (hel-Thr643, H-Ser357) bond in the D44.1 complex, which is not noted in [69], is replaced by a water-mediated hydrogen bond (hel-Thr643, HOH2545, H-Ser357).

In the analysis in [69], cation-π interactions were omitted. However, there are significant cation-π interactions as detailed in Table 14.4 between the light and heavy chains and HEL. Intermolecular salt bridges are also detailed in Table 14.4. Like the cation-π interactions, the salt bridges are similar for both F10.6.6 and D44.1. It is significant that the salt bridges form a complex, and the positive and negative charges lie very nearly in a plane (in fact, they are almost co-linear). Moreover, their

cation	aromatic	F10.6.6(1)	F10.6.6(2)	D44.1(1)	D44.1(2)
hel-Arg 668	L-Trp 94	12.3	11.9	12.2	9.93
hel-Arg 668	H-Trp 333	18.7	18.4	17.2	19.0
hel-Arg 645	H-Trp 333	8.84	8.66	8.41	8.92
hel-Arg 645	H-Tyr 359	20.1	18.7	18.8	18.0
L-Arg 96	L-Trp 94	none	15.3	16.6	16.7

Table 14.4: Cation-π interactions in D44.1 and F10.6.6. The letter 'H' refers to the 'heavy' chain, and the letter 'L' refers to the 'light' chain. The numbers given are the quality estimate of the bonds modeled as multipole interactions.

cation	anion	F10.6.6(1)	F10.6.6(2)	D44.1(1)	D44.1(2)
hel-Arg 668	H-Glu 335	71.5	75.5	73.4	70.0
hel-Arg 645	H-Glu 350	69.6	71.5	68.7	80.1
hel-Arg 668	H-Glu 350	113.	110.	102.	94.2

Table 14.5: Salt bridges in D44.1 and F10.6.6. The letter 'H' refers to the 'heavy' chain, and the letter 'L' refers to the 'light' chain. The numbers given are the quality estimate of the bonds modeled as multipole interactions.

arrangment, at the vertices of a rhombus, is consistent with a quadrupole structure.

What is even more striking is the fact that the cations are involved in both salt bridges and interactions with aromatics. The aromatics form a cage around the salt bridges, but it is not clear that there are strong cation-pi interactions in all cases. However, the unique combintation of residues in both F10.6.6 and D44.1 that form bonds with the two arginines in HEL is presumably a critical part of the recognition and binding strategy for this antigen-antibody pair.

To assess the dehydons at the interface between the antibody light (L) and heavy (H) chains and the antigen, HEL, we have created a hierarchy of bond types, based on their proximity to the interface. First we define the residues *at the interface* of two chains by the property that they lie within a sphere of distance x Å from a residue on the other chain. Then we define a set of *neighbors* of these residues (a.k.a. *nearby* residues) which fall within a ball of radius y Å of the interface residues. In Table 14.6, we classify dehydrons by using this classification for each of the residues involed in the bond.

A more significant differentiation between 1P2C/F10.6.6 and 1MLC/D44.1 emerges when the underwrapped intramolecular hydrogen bonds are considered (no such underwrapped intramolecular salt bridges or cation-π interactions exist in these structures). The dehydrons are listed in Table 14.7. The quality of the first three of the five bonds is significantly better in the F10.6.6 complex, and only marginally worse in the fourth.

The underwrapped mainchain-sidechain bond (H-Ser 357, H-Ser 355) is affected by proximity to the mutation D44.1/H-Gly356 to F10.6.6/H-Asp356 which provides additional wrapping not available in D44.1. In addition, the change in local environment reduces the number of available

DI	indicator description
1	both donor and acceptor at interface
2	either donor or acceptor at interface and the other nearby
3	both donor and acceptor are nearby
4	neither donor or acceptor at interface but wrappers from other chain are in desolvation shell
5	minimally wrapped, not at interface but wrappers from other chain are in desolvation shell
6	underwrapped, not at interface with no wrappers from other chain in desolvation shell

Table 14.6: Table of dehydron indicators DI. Residues are at the interface of two chains if that they lie within a sphere of distance x Å from a residue on the other chain. Nearby residues are those within a ball of radius y Å of the interface residues.

	bond details	PDB/antibody	distance	DI	raW	erW
donor	hel-Leu 684 N	1P2C/F10(1)	2.96	1	11	3
acceptor	hel-Ser 681 O	1P2C/F10(2)	2.95	5	13	3
interface	H/hel	1MLC/D44(1)	3.02	5	14	3
bond type	M-M	1MLC/D44(2)	none			
donor	H-Thr 331 N	1P2C/F10(1)	3.13	2	14	0
acceptor	H-Thr 328 O	1P2C/F10(2)	2.99	2	14	0
interface	H/hel	1MLC/D44(1)	3.05	3	15	0
bond type	M-M	1MLC/D44(2)	3.04	3	15	0

Table 14.7: Underwrapped mainchain (M-M) intramolecular hydrogen bonds in two pairs of antibody-antigen complexes. Distance is between the donor and acceptor (heavy) atoms in the hydrogen bond, in Ångstroms. DI: dehydron indicator, as described in Table 14.6. Following DI are the numbers of nonpolar carbons in the desolvation shell from the chain containing the bond (erW, for int*er*molecular wrappers) and from the chain forming the interface (raW, for int*ra*molecular wrappers). In the second complex in 1P2C/F10.6.6, the donor is hel-Arg 661-NH1 instead of hel-Ser 650-OG, and the acceptor is hel-Asp 648-O instead of hel-Asp 648-OD1. In the 1MLC/D44.1 structure, an alternative (M-S) dehydron appears in the second complex: hel-Asp 648-N — hel-Asp 648-OD1, with distance 2.48 and quality 4.03. It is wrapped by L-Asn92-CB. In the first copy of D44.1, the donor for hel-Arg 645 is NH2 instead of NH1.

	bond details	PDB/antibody	distance	quality	DI	raW	erW
donor	H-Ser 357 N	1P2C/F10(1)	3.19	5.41	1	9	5
acceptor	H-Ser 355 OG	1P2C/F10(2)	3.26	4.99	1	9	6
interface	H/hel	1MLC/D44(1)	none				
bond type	M-S	1MLC/D44(2)	3.20	4.19	1	8	3
donor	hel-GLY 667 N	1P2C/F10(1)	2.99	2.22	1	8	7
acceptor	hel-Asn 665 OD1	1P2C/F10(2)	2.98	1.96	1	8	6
interface	H/hel	1MLC/D44(1)	3.22	1.53	4	7	6
bond type	M-S	1MLC/D44(2)	none				
donor	hel-Ser 650 N	1P2C/F10(1)	3.16	5.65	4	5	3
acceptor	hel-Asp 648 OD1	1P2C/F10(2)	3.08	5.80	1	6	3
interface	L/hel	1MLC/D44(1)	3.09	5.72	4	6	4
bond type	M-S	1MLC/D44(2)	3.23	5.96	4	5	4
donor	hel-Asp 648 N	1P2C/F10(1)	none				
acceptor	hel-Asp 648 OD1	1P2C/F10(2)	none				
interface	L/hel	1MLC/D44(1)	none				
bond type	M-S	1MLC/D44(2)	2.48	4.03	3	2	0
donor	hel-Ser 650 OG	1P2C/F10(1)	2.65	8.68	4	5	3
acceptor	hel-Asp 648 OD1	1P2C/F10(2)*	3.24	1.62	6	9	0
interface	L/hel	1MLC/D44(1)	2.33	15.7	4	6	4
bond type	S-S	1MLC/D44(2)*	2.61	4.86	4	5	4
donor	hel-Arg 645 NH1(2)	1P2C/F10(1)	2.80	6.48	1	12	0
acceptor	hel-Thr 643 OD1	1P2C/F10(2)	3.03	4.76	6	12	0
interface	H/hel	1MLC/D44(1)	3.25	3.24	1	12	0
bond type	S-S	1MLC/D44(2)	2.75	4.14	6	13	0
donor	L-Arg 96 NH1	1P2C/F10(1)	2.87	2.29	2	13	0
acceptor	L-Ser 91 O	1P2C/F10(2)	3.08	1.74	2	13	0
interface	L/hel	1MLC/D44(1)	3.00	1.94	2	14	0
bond type	S-M	1MLC/D44(2)	none				

Table 14.8: Underwrapped sidechain hydrogen bonds in two pairs of antibody-antigen complexes. S=sidechain, M=mainchain. The numbers given are (1) the distance between the donor and acceptor (heavy) atoms in the hydrogen bond and (2) the quality estimate of the hydrogen bond modelled as a dipole-dipole interaction. DI: dehydron indicator, as described in Table 14.6. Following DI are the numbers of nonpolar carbons in the desolvation shell from the chain containing the bond (erW, for int*er*molecular wrappers) and from the chain forming the interface (raW, for int*ra*molecular wrappers). In the second complex in 1P2C/F10.6.6, the donor is hel-Arg 661-NH1 instead of hel-Ser650-OG, and the acceptor is hel-Asp648-O instead of hel-Asp648-OD1. In the 1MLC/D44.1 structure, an alternative (M-S) dehydron appears in the second complex: hel-Asp 648-N — hel-Asp 648-OD1, with distance 2.48 and quality 4.03. It is wrapped by L-Asn92-CB. In the first copy of D44.1, the donor for hel-Arg 645 is NH2 instead of NH1.

intermolecular wrappers upon binding from HEL from 5 or 6 to three.

14.2 Four antibody complexes

In another paper [261], the structures of four HEL antibodies were studied. These antibodies are very close in affinity, but they can be ranked in order of affinity, and we have followed this ranking in our presentation of the data.

The antibody complexes are found in four PDB files: 1NDM, 1DQJ, 1C08, and 1NDG. In all of these, the light chain is chain A, the heavy chain is chain B, and the antigen (HEL) is chain C. Unlike the antibody complexes in the previous section, these appear as single complexes in each PDB. If there is any artifact due to crystalization in these complexes, we do not have the same direct evidence as in the previous section.

Again we see a wide range of interactions represented between the antibody and the antibody light and heavy chains. Both cation-π interactions and salt bridges appear, as indicated in Table 14.9. One of the cations (Lys 697) in the antigen interacts with two aromatics and two negatively charged residues in the antibody. This suggests a complex recognition pattern for this residue. Another cation (Arg 621) interacts with four aromatics in the antibody, again suggesting a complex signal.

Intermolecular hydrogen bonds are listed in Table 14.10. We see a diverse set of mainchain and sidechain intermolecular hydrogen bonds, connecting both the light and heavy chains to the antigen. The pattern of bonds among the four antibodies provides a signature for them. Overall, there are fourteen different bonds represented. But in each antibody-antigen system, a range of 7 to 10 of these appears. Only four of the bonds are preserved in all antibody-antigen complexes, including the only mainchain-mainchain hydrogen bond. Curiously, the antibody with the highest affinity has the lowest number of intermolecular hydrogen bonds.

Finally, underwrapped intramolecular mainchain hydrogen bonds are listed in Table 14.11, and underwrapped intramolecular sidechain hydrogen bonds are listed in Table 14.12. Again, there is a signature to the antibodies found in the dehydron patterns, and the antibody with the highest affinity has the lowest number of underwrapped intramolecular hydrogen bonds.

14.3 Drug specificity controlled via wrapping technology

The dehydron provides a marker for drug-target interaction, and it has guided the redesign of drugs to improve specificity and affinity. Wrapping technology analyzes interface dehydrons in a target-ligand complex to predict their change in strength and stability induced by changes in non-polar microenvironment upon target-ligand binding. Differences in target dehydron patterns are exploited to redesign drugs to control selectivity and improve affinity. Modifications to the cancer drug Gleevec and to the natural compound staurosporine, tested by affinity assays and in cell lines, have been carried out.

Differences between targets can be formalized by mathematical metrics that are based on different data. Distance matrices defined by dehydron patterns in targets correlate strongly with

cation	aromatic	1NDM	1DQJ	1C08	1NDG
hel-ARG 621	L-TRP 94	11.4		11.4	11.2
hel-ARG 621	L-TYR 96	38.4	39.5	41.5	39.4
hel-ARG 621	H-TYR 358	27.3	27.2	27.2	1.82*
hel-ARG 621	H-TYR 350	55.9	52.1	56.1	51.9
hel-LYS 696	L-TYR 50	25.3	24.6	26.0	25.1
hel-LYS 697	H-TRP 398		9.50	9.27	9.71
hel-LYS 697	H-TYR 333	21.7	19.7	20.4	20.5
cation	anion	1NDM	1DQJ	1C08	1NDG
hel-LYS 697	H-Asp 332	118	109	112	110
hel-LYS 697	H-GLU/Asp 399	97.4		122	108

Table 14.9: Quality estimates for cation-π interactions and salt bridges in four antibody-lysozyme complexes. In the cation-π interactions, note the change in position 358 in 1NDG from Tyr to Phe and the corresponding change in quality estimate. In the salt bridges, note that H-GLU 399 in 1NDM aligns with H-Asp 399 in the other three structures. Note that position 399 in 1DQJ is a glycine which explains the absence of this salt bridge in 1DQJ. In all of these, the light chain is chain A, the heavy chain is chain B, and the antigen (HEL) is chain C.

pharmacological distances based on affinity assays against a set of drugs, indicating applicability of wrapping technology to broad classes of drugs and targets. Distance matrices defined by dehydron patterns in targets also correlate closely with standard sequence metrics.

14.3.1   Rational drug discovery

Rational drug discovery is rapidly evolving [13, 196, 230, 310, 403], and wrapping technology is one technique that has been used to redesign existing drugs to improve selectivity and affinity [125, 136, 149]. It is based on the observation that the strength and stability of dehydrons can be modulated by an external agent of appropriate design. More precisely, dehydrons can be further desolvated by the attachment of a drug that contributes to dehydration, as depicted in Figure 14.2. Through retrospective studies [137], it appears that dehydrons have played a significant role in drug function due to their prominence in protein-ligand interactions [98, 275]. Wrapping technology exploits differences in wrapping among proteins, especially those with similar structures, to enhance selectivity and affinity of drug ligands.

Although specificity may not be required for clinical activity, controlling drug selectivity for clinically relevant targets can help reduce toxicity and enhance activity. Drug cross reactivity may be displayed as a pharmacological-distance matrix [136] or its derived pharmacological dendrogram [403, 116, 153] quantifying comparisons among the affinities of proteins to a background set of drugs.

type	donor	acceptor	1NDM	1DQJ	1C08	1NDG
M-M	hel-ARG 621 N	L-Asn 92 O	2.68 4.13	2.81 3.32	2.89 3.31	2.80 3.47
M-S	H-Ser 354 N	hel-Asp 701 OD1	3.06 7.24	3.15 6.43	3.20 6.16	(3.59 4.59)
M-S	hel-GLY 702 N	H-Ser 356 OG	2.88 6.46	2.82 6.33	3.00 6.21	
S-M	H-Ser 331 OG	hel-ARG 673 O			3.42 1.28	(4.04 0.483)
S-M	H-TYR 350 OH	hel-Ser 700 O	2.55 5.35	2.60 5.02	2.64 4.63	2.66 4.46
S-M	H-TYR 358 OH	hel-Asp 701 O	3.17 0.32	3.38 0.52		
S-M	hel-ARG 673 NH2	H-Thr 330 O		2.47 0.81		
S-S	hel-ARG 673 NH1	H-Thr 330 OG1	3.16 3.91			
S-S	H-Ser 352 OG	hel-Asp 701 OD1	2.71 8.07	2.70 7.98	2.50 10.2	2.88 2.60
S-S	H-Ser 354 OG	hel-Asp 701 OD1	2.85 5.56	2.49 11.5	2.77 5.93	2.80 6.64
S-S	L-GLN 53 NE2	hel-Asn 693 OD1	2.85 0.56	2.84 0.85		2.81 0.82
S-S	hel-Asn 693 ND2	L-GLN 53 OE1		3.30 0.228	2.83 0.375	3.30 0.296
S-S	hel-LYS 696 NZ	L-Asn 31 OD1			2.95 4.76	2.83 4.64
S-S	hel-LYS 696 NZ	L-Asn 32 OD1			2.82 7.13	
		total bonds	9	10	10	7

Table 14.10: Intermolecular hydrogen bonds in four antibody-lysozyme complexes. S=sidechain, M=mainchain. The numbers given are (1) the distance between the donor and acceptor (heavy) atoms in the hydrogen bond and (2) the quality estimate of the hydrogen bond modelled as a dipole-dipole interaction. Note that the two bonds involving hel-LYS 696 NZ in 1C08 are in conflict, in the sense that one would not normally think of the N-H group represented by NZ as capable of forming two hydrogen bonds. However, this ambiguity reflects the geometry involving this group and the two 'acceptor' atoms (OD1 of L-Asn 31 and L-Asn 32). In 1NDG(H8), the donor for the S-M bond with hel-Arg673-O changes from H-Ser331-OG to H-Arg331-NE. Data in parentheses are for reference only. By relaxing the definition of hydrogen bond, we can determine the data not called a hydrogen bond. In all of these, the light chain is chain A, the heavy chain is chain B, and the antigen (HEL) is chain C.

IC	type	donor	acceptor	1NDM	1DQJ	1C08	1NDG
AC	M-M	A-GLY 68 N	A-Ser 30 O	2.89	2.80	2.82	2.85
BC	M-M	B-Ser 331 N	B-Ser 328 O	3.20		3.15	
CB	M-M	C-GLY 704 N	C-Asp 701 O		3.21	2.86	3.36

Table 14.11: Mainchain dehydrons in four antibody-lysozyme complexes. Key: IC=interface chains; type: S=sidechain, M=mainchain. Distances in Ångstroms between acceptor and donor (heavy) atoms for each dehydron are listed. The two dehydrons in 1C08 involving B-Thr320-OG1 and B-Ser328-OG are counted as only one bond, since only one of the two possibilities can be adopted at any one moment. Both are equally compatible with the mainchain-sidechain bond between B-Thr320-N and B-Ser328-OG.

IC	type	donor	acceptor	1NDM	1DQJ	1C08	1NDG
AC	M-S	A-Ser 30 N	A-Ser 28 OG		3.47 4.96		
AC	S-S	A-Ser 93 OG	A-Asn 92 OD1	2.98 0.62	3.10 0.36		
BC	S-S	B-Ser 328 OG	B-Thr 330 OG1			2.57 5.57	
BC	S-S	B-Thr 330 OG1	B-Ser 328 OG			2.57 9.13	
BC	M-S	B-Thr 330 N	B-Ser 328 OG			2.86 7.47	
BC	S-M	B-Ser 331 OG	B-Ser 331 O	2.59 5.16			
BC	M-S	B-Ser 356 N	B-Ser 354 OG	3.26 5.07	3.41 4.73	3.25 5.51	
BC	S-S	B-Ser 356 OG	B-Ser 354 OG	2.88 3.58	2.94 3.97	3.14 2.84	
CB	M-S	C-Asn 703 N	C-Asp 701 OD2	3.15 6.68	3.09 6.99	2.86 3.08	

Table 14.12: Summary of dehydrons. Key: IC=interface chains; type: S=sidechain, M=mainchain. Distance and quality measures for each dehydron are listed. The two dehydrons in 1C08 involving B-Thr320-OG1 and B-Ser328-OG are counted as only one bond, since only one of the two possibilities can be adopted at any one moment. Both are equally compatible with the mainchain-sidechain bond between B-Thr320-N and B-Ser328-OG.

	1DQJ	1C08	1NDG	total
1NDM	4	5	7	7
1DQJ		5	5	7
1C08			6	8*
1NDG				2

Table 14.13: Distance matrix for dehydrons. The distances and total for 1C08 counts the B-Ser328-OG,B-Thr330-OG1 bonds as only one entity.
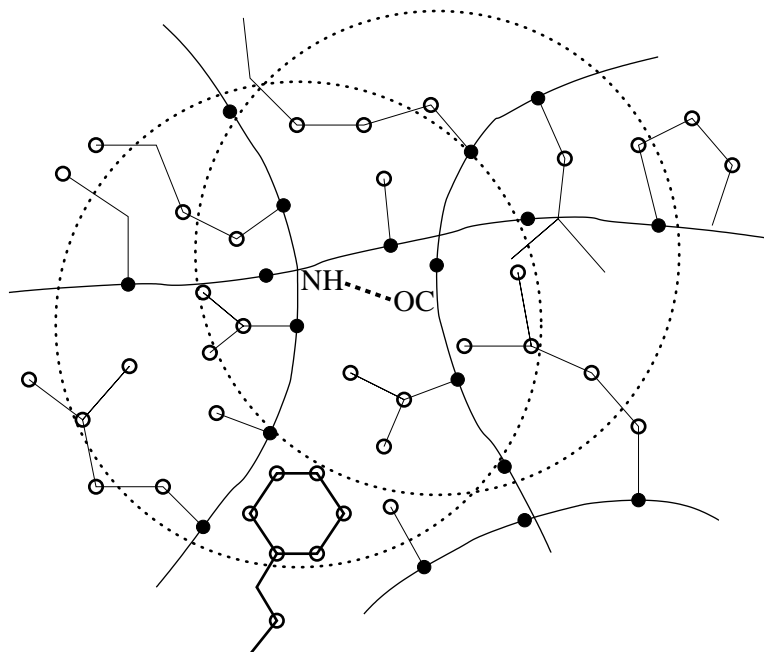
Figure 14.2: Depiction of a drug enhancing the wrapping in the desolvation domain of a hydrogen bond, cf. Figure 8.2. The desolvation domain is depicted as the union of two spheres. The solid black dots represent non-polar carbonaceous groups on the sidechains indicated by broken lines. The drug ligand provides an additional non-polar carbonaceous group (big black hexagon with a tail) in the desolvation domain.

donor	acceptor	dNO	dHO	ang1	ang2	ang3	tw	aw	bw	W	N	A
ARG A 87	ALA A 28	2.75	1.83	152.05	169.96	162.04	19	15	4	3	4	33
ARG B 87	ALA B 28	2.76	1.80	160.92	167.76	166.29	17	4	13	3	11	38
ASN A 88	ASP A 29	3.43	2.45	166.16	151.77	150.02	13	13	0	6	1	9
ASN B 88	ASP B 29	3.30	2.42	147.25	148.48	142.12	14	1	13	5	7	18
GLY A 49	GLY A 52	3.03	2.13	148.84	124.10	117.21	14	10	4	4	5	35
GLY B 49	GLY B 52	2.90	2.00	148.39	164.25	174.50	16	5	11	5	12	31
GLY A 51	ILE B 50	2.93	2.08	141.22	171.07	162.14	18	9	9	3	5	15
GLY A 52	ILE B 50	3.40	2.50	150.39	130.99	123.38	15	11	4	4	5	15
PHE B 99	PRO A 1	3.12	2.20	152.57	164.21	158.65	15	7	8	3	0	0

Table 14.14: Dehydrons in the PDB file 2BPX which depicts the HIV-1 protease complexed with Indinavir (designated MK1 in the PDB file). Key: dXO denotes the distance from X(=N or H) to O in the hydrogen bond, tw is the total number of wrappers, xw is the number of wrappers from chain x (=A or B), W is the number of waters in the desolvation shell (radius 6.5Å), N is the number of nonpolar carbons in MK1 in the desolvation shell (radius 6.5Å), A is the number of all atoms in MK1 in the desolvation shell.
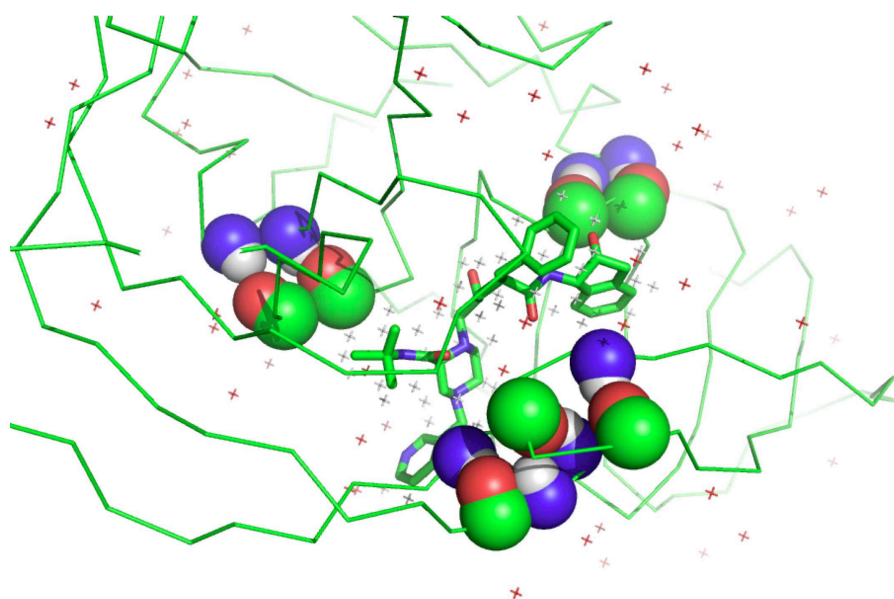
Figure 14.3: HIV-1 protease in complex with inhibitor Indinavir that acts as dehydron wrapper, cf. Table 14.14. The protease is shown as lines between C_α atoms, Indinavir is depicted as sticks, and the spheres denote the dehydron constituents (donor and acceptor atoms).

14.3.2 Ligands as dehydron wrappers in protein-inhibitor complexes

Inhibitor design is often guided by structural descriptors of protein binding sites, such as accessibility [80, 252], curvature [34], and hydrophobicity [73, 308]. However, out of 814 protein-inhibitor PDB complexes studied in [125], at least 488 of them had binding cavity hydrophobicity not significantly higher than the rest of the surface [125]. While intermolecular electrostatic bonds are obvious candidates to promote protein-ligand association and specificity [430], we have seen that dehydrons often play an equally significant role, as is the case for antibody binding (Section 14.1).

Specificity of association is an essential aspect of biological proteins, and a critical goal of drug discovery. But many drug targets, such as protein kinases [81], share similar properties which foster promiscuity unless differentiating features are targeted. For example, dehydrons are often not conserved in families of paralog proteins [130]. By contrast, it appears that surface nonpolar moieties are a highly conserved feature of protein interfaces [275]. Thus, side effects resulting from off-target ligand binding may be minimized by selectively targeting dehydrons. Standard phylogenetic analysis based on differences in dehydron patterns provide a dendrogram which displays relationships crucial to selectivity as depicted in two examples in Figure 14.6.

Over three-fourths of the protein-inhibitor PDB complexes studied in [125] have nonpolar groups in the inhibitors within the desolvation domain of protein dehydrons. That is, inhibitors are typically dehydron wrappers. This situation is illustrated in Figure 14.3 which shows the inhibitor Indinavir (Crixivan) bound to the functionally dimeric HIV-1 protease (PDB file 2BPX) (cf. Fig. 1 of [125]). Six intramolecular dehydrons in the protease, the backbone hydrogen bonds Ala28-Arg87, Asp29-Asn88, and Gly49-Gly52 in both chains A and B (marked in green in Fig. 1 in [125]), frame the cavity associated with substrate binding. These dehydrons help to explain the mechanism of substrate binding and clarify the gating mechanism of the flap [125].

The dehydrons shown in Figure 14.3 appear in three clusters at distributed sites which form a distinctive pattern around the nonpolar regions of the ligand. Clusters of dehydrons are common, since decreased wrapping around one hydrogen bond is likely to contribute to decreased wrapping nearby. Indinavir contributes several desolvating groups (Table 14.14) to wrap the dehydrons in the enzymatic cavity. The association of Indinavir deactivates the dehydrons that promote water removal and charge de-screening required to facilitate the enzymatic nucleophilic attack [137]. Additional intermolecular dehydrons in the protease are also significantly desolvated by nonpolar groups in Indinavir, contributing to its binding affinity. In total, 8 dehydrons are identified in Table 14.14 that benefit from the water removal induced by Indinavir binding.

We see in Table 14.14 the extensive intrusion of Indinavir (designated MK1 in the PDB file) in the desolvation spheres around the most underwrapped hydrogen bonds in the PDB file 2BPX. Shown in Table 14.14 are all hydrogen bonds whose total number of wrappers is less than 20 using a desolvation shell radius of 6.5Å around the CA carbons. Also indicated in Table 14.14 is the breakdown of the wrappers by chain; the maximum number of wrappers for the intramolecular hydrogen bonds from one chain is 15, so these are extremely underwrapped before the dimer is formed. The intermolecular hydrogen bonds between the two chains draw wrappers from the two chains even more strongly, and two of these have significant contributions to wrapping from MK1. Interestingly, there are no other hydrogen bonds in 2BPX with more than 8 atoms from MK1 in the corresponding desolvation domain, or more than 4 nonpolar carbons from MK1.

bond	MK1 atom	distance	heavy atom
none	N1		
water	N2(H)	3.06	HOH B 332
water?	N3	3.59	HOH A 308
MCHB	N4(H)	3.02	O GLY B 27
SCHB	N5	3.63	NH1 ARG B 8
water	O1	2.90	HOH A 308
SCHB	O2(H)	2.71	OD2 ASP A 25
water	O3	2.97	HOH A 308
MCHB	O4(H)	2.90	N ASP B 29

Table 14.15: Polar contacts between Indinavir (MK1) and the HIV-1 protease in PDB file 2BPX. Key: bond indicates the type of hydrogen bond (HB) made (MC is mainchain, SC is sidechain). The last column gives the heavy atom participating in the hydrogen bond. The hydrogen bond between N3 and HOH A 308 is unlikely due since it competes with the more likely partners O1 and O2.

In addition to the extensive dehydron network involved in the binding of Indinavir to the HIV-1 protease, the PDB file 2BPX reveals conventional intermolecular polar interactions between Indinavir and the HIV-1 protease. In Table 14.15, all of the polar atoms (nitrogens and oxygens) in MK1 are listed, together with the hydrogen bonds they make. Of the 9 polar atoms, only 4 make hydrogen bonds with the HIV-1 protease; two with mainchain acceptors and two with sidechains. Three others are hydrogen bonded to waters, and the remaining two do not appear to be making strong polar bonds. Thus the inventory of enthalpic contributions by intermolecular polar interactions between Indinavir and the HIV-1 protease is smaller than the enthalpic contribution of the dehydron network depicted in Table 14.14.

14.4   Reviewing Imatinib

The cancer drug Imatinib (a.k.a., Gleevec, STI-571) was the first cancer drug used to target specific proteins involved in the cancer process, tyrosine kinases, and the success of Imatinib has had a substantial impact on drug design []. Janet Rowley, Brian Druker, and Nicholas Lydon received the 2012 Japan Prize "for their contribution to the development of a new therapeutic drug targeting cancer-specific molecules" according to the prize announcement, the drug in question being Imatinib.

Imatinib, shown in Figure 14.4, was originally developed to target the Bcr-Abl protein in the treatment of chronic myeloid leukemia. However, its success led to attempts to use it to target other tyrosine kinases. The C-KIT kinase is a therapeutic target for treating gastrointestinal stromal tumors, and it is inhibited by imatinib at the nanomolar level.

The PDB file 1T46 details the binding of Imatinib to the tyrosine kinase C-Kit. As listed in Table 14.16, Imatinib only makes four polar contacts with C-Kit in 1T46, and two of the other polar sites on Imatinib make apparant hydrogen bonds with water. Of the former, the potential
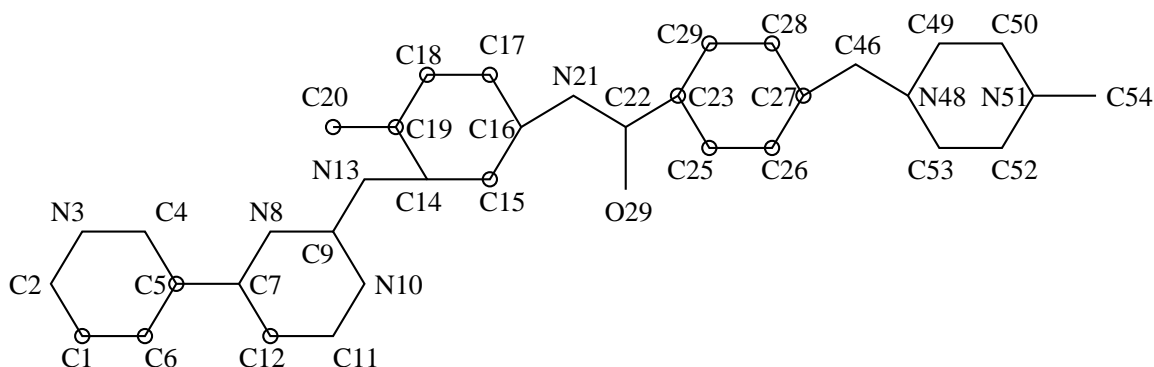
Figure 14.4: PDB atom names for Imatinib (STI in PDB file 1T46).

bond	MK1 atom	distance	heavy atom
MCHB	N3	2.85	N CYS 673
SCHB	N13(H)	2.95	OG1 THR 670
SCHB	N21(H)	2.86	OE1 GLU 640
MCHB	O29	3.08	N ASP 810
water	N10	2.82	O HOH 1105
water	N48	3.17	O HOH 1108

Table 14.16: Polar contacts between Imatinib (STI) and C-Kit in PDB file 1T46. Key: bond indicates the type of hydrogen bond (HB) made (MC is mainchain, SC is sidechain). The last column gives the heavy atom participating in the hydrogen bond. The hydrogen bond between N3 and HOH A 308 is unlikely due since it competes with the more likely partners O1 and O2.

hydrogen bond between N21 and OE1 GLU 640 does not have an ideal geometry, and GLU 640 makes a salt bridge with LYS 623. Similarly, the geometry of the hydrogen bond between O29 and the mainchain donor on Asp 810 is not ideal. The sites N8 and N51 do not make obvious polar contacts with either.

Curiously, Imatinib has minimal interactions with dehydrons in C-Kit. There are 16 hydrogen bonds in 1T46 wrapped by fewer than 21 nonpolar carbons within dehydration domains consisting of spheres of radius 6.5Å around the two CA atoms for the donor and acceptor. None of these desolvation domains contain any atoms from Imatinib (MK1 in 1T46). Thus, unlike the impact of Indinavir on the dehydrons in the HIV-1 protease, Imatinib appears to bind without wrapping its target in an obvious way, if at all.

Unfortunately, there are several deficiencies of imatinib, and some attempts have been made to improve the ligand. We review two of these, one which attempts to respond to a type of Imatinib resistance, and the other trys to minimize side effects.

14.4.1 Imatinib resistance

In [118], the authors identified the possible dehydron formed by Phe 811 and Ala 814 as a possible target for modification to Imatinib. The dehydration domain consisting of the spheres of radius 6.5Å around the two CA atoms of Phe 811 (acceptor) and Ala 814 (donor) contains 21 nonpolar carbons, 10 water molecules, and 9 atoms from Imatinib (code STI in 1T46). Of the latter, the polar carbon C11, which is covalently bonded to N10 in Imatinib, was identified for enhancement by methylation, which would further enhance wrapping of this bond.

The Imatinib atoms in the desolvation domain of Ala 814–Phe 811 are all within the 6.5Å sphere around Phe 811 CA. The distance from Phe 811 CA to C11 is 4.92Å, and the closest nonpolar carbon is C12 (6.08Å). If we had chosen the desolvation radius to be 6Å, then there would have been no nonpolar Imatinib atoms in the desolvation domain. The carbon C15 (distance 6.42Å) is also nonpolar, but the carbon C22 (distance 6.40Å) is polar. The polar atom O29 (6.15Å) is also nearby, and the distance from Phe 811 CA to N10 is 5.22Å. Moreover N10 is 2.82Å from the water HOH 1105, indicating a hydrogen bond, and this water appears to be making a polar bond with the end of the sidechain Lys 623. This water is also only 3.45Å from Phe 811 CA.

Thus the Imatinib nonpolar atoms are peripheral to the desolvation domain of Ala 814–Phe 811, and the methylation at C11, the closest heavy atom to Phe 811 CA, could thus be expected to enhance this hydrogen bond. According to MolProbity, the distance from Phe 811 CA to the hydrogen attached to C11 is 3.97Å, so the location of the new methyl group would be in a very favorable position for desolvation. Thus the addition of a methyl group at C11 could be anticipated to add a new type of binding mechanism (wrapping) to Imatinib to C-Kit.

The authors of [118] tested a modification, called WBZ_7, to Imatinib created by methylation at the C11 position. WBZ_7 was tested and shown to have a slightly higher affinity than Imatinib for native C-Kit. Thus the strategy of adding a wrapping mechanism to Imatinib was successful, despite the danger that enlarging the ligand might cause a steric hindrance that would reduce affinity. In fact, affinity was improved, and since a novel mechanism of binding was added, it seems plausible that the binding of WBZ_7 would be more robust to perturbations in the target, C-Kit.

It was known that just such a modification to C-Kit, the mutation D816V (Asp 816 is converted to Val 816), confers resistance to Imatinib. The affinity of Imatinib to the mutatant D816V is reduced by three orders of magnitude compared with the affinity to the wild-type C-Kit kinase. Fortunately, WBZ_7 was found to bind strongly to the D816 mutant of C-Kit. It is interesting to speculate why this might occur.

We know that WBZ_7 was designed to bind more robustly to C-Kit, by adding a binding mode based on desolvating a dehydron, a binding mode missing from Imatinib binding to C-Kit. If a mutation in a target causes a significant structural change in the target, this may move polar contacts enough to cause a significant decrease in the binding enthalpy between the drug and target. The mutation D816V of C-Kit has a significant impact on wrapping of the hydrogen bond formed by Phe 811 and Ala 814, as well as others. The residue Asp 816 (that is, its CB atom) is found in the desolvation domain of several hydrogen bonds in 1T46, as indicated in Table 14.17. Thus the mutation to Val 816 will likely have a significant effect on the strength and stability of some of the hydrogen bonds in mutant C-Kit, possibly leading to substantial structure changes. There are no Imatinib atoms in the desolvation domains of any of these hydrogen bonds.

donor	acceptor	dNO	dHO	ang1	ang2	ang3	tw	W
GLY 598	GLY 601	2.72	1.74	166.2	164.9	159.9	26	5
ARG 815	GLY 812	3.25	2.28	163.1	126.6	129.6	17	16
ASN 819	ASP 816	3.09	2.23	142.5	124.4	117.0	11	6
TYR 823	ASP 820	2.92	1.95	164.8	125.3	120.2	20	5
ALA 814	PHE 811	2.81	1.84	163.5	132.9	134.3	21	10 waters

Table 14.17: Hydrogen bonds in the PDB file 1T46 which contain the CB atom from Asp 816 in their desolvation domains consisting of spheres of radius 6.5Å around the donor and acceptor CA atoms. The PDB file 1T46 which depicts the tyrosine kinase C-Kit complexed with Imatinib (designated STI in the PDB file). Key: dXO denotes the distance from X(=N or H) to O in the hydrogen bond, tw is the total number of wrappers in a desolvation domain, and W is the number of waters in the desolvation domain. None of these desolvation domains contain any of the Imatinib (STI) atoms. For comparison, the same data is given for the Ala 814–Phe 811 hydrogen bond, whose desolvation domain contains 9 STI atoms, of which 2 are nonpolar.

The mutation D816V also could have significant effects on interactions in addition to the main-chain hydrogen bonds listed in Table 14.17. The oxygen OD1 in ASP 816 in 1T46 is 3.23Å from the water HOH 1060, indicating a likely hydrogen bond. The mutation D816V would likely remove this water from the environment. The oxygen OD2 in ASP 816 is 2.96Å from ND2 in ASN A 819 (and according to MolProbity, 1.99 Å from HD22 in ASN 819), indicating a likely sidechain hydrogen bond. This bond could not be supported by Val 816, further contributing to possible structural change.

Thus it is reasonable to assume that the mutation D816V changes structural features of C-Kit that make binding by Imatinib harder. Since Imatinib appears to bind solely due to polar contacts, small structural changes can easily disrupt them. By contrast, wrapping a dehydron is much more robust to small structural changes, since the desolvation effects of nonpolar groups are less directional than polar contacts. The binding of WBZ_7 to the D816V mutant may continue to be driven in part by the F811–A814 dehydron.

14.4.2   Avoiding side effects

Another modification to Imatinib, called WBZ_4, was studied [139] in an attempt to retain binding to C-Kit but avoid binding to other tyrosine kinases. WBZ_4 is derived from Imatinib by methylation at the C2 position in STI (Figure 14.4). The carbon atom C2 in STI is found in a 6.5Å desolvation domain around three hydrogen bonds listed in Table 14.18. Although these hydrogen bonds are not extremely underwrapped, it does appear that there are several waters in their vicinity. Molecular dynamics studies [139] showed that the mainchain hydrogen bond between C673 and G676 is substantially enhanced by the methylation at the C2 position in STI, including removal of intevening waters.

From Table 14.18, we see that the hydrogen bond between C673 and G676 has reduced wrapping

donor	acceptor	dNO	dHO	ang1	ang2	ang3	tw	W	N	A
GLY A 676	CYS A 673	2.89	1.93	162.22	132.28	126.23	23	13	3	10
LEU A 800	GLY A 676	2.83	1.93	147.44	166.36	155.79	25	9	3	9
PHE A 681	ASP A 677	3.02	2.06	159.71	157.60	151.02	31	9	2	6

Table 14.18: Hydrogen bonds in the PDB file 1T46 which contain the C2 atom from Imatinib (STI) in their desolvation domains consisting of spheres of radius 6.5Å around the donor and acceptor CA atoms. Key: dXO denotes the distance from X(=N or H) to O in the hydrogen bond, tw is the number of wrappers excluding ones from Imatinib, W is the number of waters in the desolvation domain, N is the number of nonpolar carbons in MK1 in the desolvation domain, A is the number of all atoms in MK1 in the desolvation shell.

and a substantial number of neighboring waters. The position (according to MolProbity) of the hydrogen H21 attached to C2 in STI is favorable for placing a methyl group. H21 is about 3Å from the center of the hydrogen bond between C673 and G676, and it is about 3Å from HOH 1187 in 1T46. The nearest atom to H21 that could cause a steric clash is CE2 in Tyr 672, at a distance of 3.05Å. The distance from C2 in STI to CE2 in Tyr 672 is 3.54Å.

Unlike the WBZ_7 variant of Imatinib, WBZ_4 was designed not to bind to the Abl kinases, due to their impact on heart disease [139]. Through a series of tests, WBZ_4 was shown [139] to bind selectively to C-Kit in preference to the Abl kinases.

14.5   Scope of wrapping technology

The possibility of multiple binding partners for a given protein inhibitor can arise in paralog proteins because they share common domain structures [302]. Toxic side effects can result unless nonconserved features are specifically targeted. Remarkably, dehydrons are differentiating features among proteins with nearly identical structures. Figure 14.5 illustrates the differences in dehydrons in the closely aligned structures of two paralogs, Chk1 and Pchk1. The backbone structures of the proteins align almost perfectly, but the dehydrons are surprisingly different, due to the differences in packing determined by the composition of the individual sidechains.

Different modifications of Gleevec have been designed using the wrapping technology reviewed here, and they have been tested experimentally with remarkable success [118, 125, 136, 149, 139]. A common view is that "identification of structure features that allow differentiation between effect and side-effect profiles of medicinal agents is currently rate limiting in drug discovery" [153]. Wrapping technology has had significant success in improving specificity and affinity [125, 136, 149] precisely because it exploits such features, namely, dehydrons.

Experimentally based approaches to rational drug design [153, 196, 310, 403] seek to reveal 'what' is significant for drug discovery. However, wrapping technology exploits 'why' ligands interact with proteins [125, 136, 149]. Thus we counter the view that "computational solutions that precisely link molecular structure to broad biological response are currently not possible" [153].

Typical analysis of protein-ligand binding involves pairwise, intermolecular interactions [182] combined with steric considerations [80]. Surprisingly, the modulation of the strength and stability of intramolecular hydrogen bonds by hydrophobic groups on ligands (drugs) plays a significant role in many target-ligand complexes. Such ligands do not utilize pairwise intermolecular interactions but rather promote water exclusion near dehydrons upon association.

We have focused here on an analysis of static information related to protein-ligand binding, yet "the need to account for the dynamic behavior of a receptor has long been recognized as a complicating factor in computational drug design" [67]. Not only does induced fit play a role [182], but structural flexibility can also lead to unwanted promiscuity [77]. Kinase targets pose a major challenge to structure-based drug design because their flexible loopy regions framing the ATP pocket and the transphosphoesterification environment are likely to adopt an induced fit, so far unpredictable, upon association with ligands. The relationship between disorder and defective wrapping [130] helps to explain the order-upon-binding scenario and may be turned into a design concept whereby the putative ligand becomes a wrapper of the packing defects that the loopy region forms as it visits specific conformations. Thus, the activation loop of C-Kit kinase was stabilized in the activated, induced-fit conformation by redesigning the drug inhibitor Gleevec to wrap better a dehydron in the activation loop [149].

Figure 14.5: Aligned backbones for two paralog kinases, Chk1 and Pdk1, represented by virtual bonds joining α-carbons. The dehydrons (SAHBs) are marked as segments joining the α-carbons of the residues contributing the amino or carbonyl group to the backbone hydrogen bond. Dehydrons for Chk1 are marked in green and those for Pdk1 are in red. While the kinases are structurally similar (RMSD= 4.1 Å), their pattern of packing defects (dehydrons) is different and wrapping technology takes advantage of this fact.

Wrapping technology represents a new paradigm for rational design of drugs that exclude water near solvent accessible hydrogen bonds on the surface of the protein target. This represents a clear departure from the standard approach which focuses on the possibility of promoting pairwise interactions between ligand and target. Not only does wrapping technology utilize a novel modality crucial to the functioning of large classes of ligands, it also guides modifications to existing ligands as opposed to a combinatorial search paradigm. A major obstacle to targeting kinases is their high level of structural similarity, but it is overcome by wrapping technology which exploits packing differences. Wrapping technology identifies potential ligands with high selectivity, utilizing packing defects as selectivity filters and switches that can
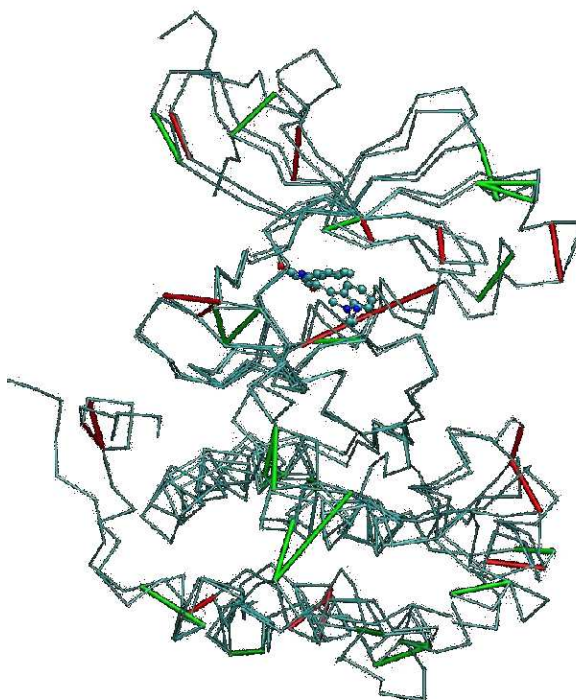
allow the targeting of specific proteins.

14.6   Distance metrics, similarity trees and dendrograms

Different features can indicate a propensity for protein-protein interaction, such as surface curvature, wrapping (dehydrons), hot spots, etc. These features can be compared on different proteins to see if they are similar or not. Different features may not be conserved across paralogs. A numerical measure can help to quantify relationships. This can be done in a variety of ways, but one way is to first form a standard structural alignment (`http://www.ncbi.nlm.nih.gov/structure/CN3D/cn3d.shtml`) and then to trim the protein sequences by restricting to those residues that have been structurally aligned in a one-to-one correspondence.

One metric that can be useful is to compare hydrogen bonds in the different structures. An indicator matrix $A(i, j)$ for each protein is constructed that is indexed by the residues involved in hydrogen bonds. That is, $A(i, j) = 1$ if i and j are linked by a hydrogen bond, and zero otherwise. Then, a Hamming-type distance [136] can be defined based on the number of disagreements between two such indicator matrices:

$$d(A, B) = \sum_{i,j} |A(i,j) - B(i,j)|. \qquad (14.1)$$

Similarly, for dehydron analysis, an indicator matrix for each protein is constructed that is indexed by the residues involved in dehydrons. It consists only of zeros unless two residues are paired by a dehydron, in which case the corresponding entry is one. Again, a Hamming-type distance (14.1) can be defined based on the number of disagreements between two such indicator matrices. We refer to this as **dehydron distance** or **packing distance**.

Figure 14.6: Packing similarity tree (PST, bottom in black) for the seven structurally aligned paralogs of Bcr-Abl. The PST restricted to the alignments of the Gleevec wrapped region in Bcr-Abl is shown (top) with blue dashed lines. The paralogs in red have the most similar packing in the region that aligns with the Gleevec wrapped region in Bcr-Abl and are also primary targets of this inhibitor [116].

A pharmacological distance matrix can also be constructed based on affinity profiling against specified drugs [136, 116]. The dehydron distance for a set of kinases displays remarkable similarity to pharmacological distance defined using these kinases and a set of available drugs [136, 116]. Tree representations familiar in phylogenetic analysis [302, 178] provide a useful, albeit not unique, representation of relationships among drug targets. The direct correlation of distance matrices provides a unique comparison and confirms the efficacy of wrapping technology in predicting selectivity [136].

For geometric [34] features, a distance could be constructed by comparing the curvature of the protein-ligand interface at the appropriate positions of two proteins.

Such distance matrices can be used to construct dendrograms (or trees) and are implicitly behind the dendrograms commonly seen [116, 153]. Such trees are easily constructed by using standard

algorithms for phylogenetic analysis [178]. Distance in the tree reflects distance as encoded in the distance matrix, with the proviso that certain restrictions must apply to the distance matrix to get an exact representation [178]. For example, with dehydrons, we refer to this tree as a **packing similarity tree** (PST). The resulting tree gives a visual indication of closeness for the various proteins, cf. Figure 14.6 which depicts two such trees. The PST allows the assessment of possible effects of targeting given features in drug design to determine potential specificity. Nearby proteins (Abl and Lck, for example) have similar features in our simple example.

Although visual inspection and comparison of trees is useful, the lack of uniqueness of trees is a cause for concern. A direct comparison of distance matrices provides a more rigorous comparison of measured properties, such as a comparison of pharmacological distance and packing distance [136]. In general, if a distance matrix is to be represented faithfully by a tree, it must satisfy a **four-point condition** [178] that includes the familiar triangle inequality but is substantially more restrictive. This requirement implies that typical biological trees will not uniquely represent a given biological distance matrix. Thus direct comparison of distance matrices is a more reliable technique [136].

14.7 Exercises

Exercise 14.1 *Examine the hapten-antibody complex in the PDB file 2RCS [76, 414]. Determine which bonds are involved and how they change with antibody maturation.*

Exercise 14.2 *In Chapter 7, different approaches to defining interfaces were sketched. In some cases, attempts have been made to differentiate different types of interfaces based on function [71, 215]. Determine the numbers of intramolecular dehydrons in various interfaces and compare this with the corresponding density of intermolecular hydrogen bonds in Section 7.1.*

Exercise 14.3 *Examine the antibody-antigen complexes in the PDB files 3hfm, 1c08, and 1bwh [374]. Determine if there are underwrapped hydrogen bonds are involved and, if so, how they change with antibody maturation.*

Chapter 15

Sidechain-mainchain hydrogen bonds

It is remarkable that there are more mainchain-mainchain hydrogen bonds than ones involving sidechains. Data taken from a subset of the PDB Select database consisting of 1547 proteins complexes (PDB files) has only 68,917 hydrogen bonds between sidechains and mainchains, contrasted with 233,879 mainchain-mainchain hydrogen bonds. This data has been restricted to intramolecular hydrogen bonds within a single chain for simplicity. In addition, there are 33,021 sidechain-sidechain hydrogen bonds, 152 of which involve terminal oxygens. If we classify bonds according to whether the sidechain is the donor (S-M) or acceptor (M-S), then we find that 30,640 of the total hydrogen bonds between sidechains and mainchains have the sidechain as acceptor (M-S), not including another 75 which involve terminal oxygens. Correspondingly, the remainder (38,277) are S-M bonds.

The rotameric flexibility of the sidechains (Section 5.2.4) is considerable greater than the rotational degrees of freedom (cf. Figure 5.9) of the peptide backbone for many sidechains. Thus one might expect that sidechain hydrogen bonds would play a dominant role in protein structure. However, the above data shows that the opposite is true. To be sure we are counting things correctly, we now consider carefully how to compare the opportunities for hydrogen bonding of different types.

15.1   Counting the bonds

In thinking about the likelihood of finding one type of bond versus another in proteins, there are two ways of looking at the question. Above, we have taken the view that is suitable for the following question: when looking at a protein, are we more likely to see mainchain-only hydrogen bonds than ones involving sidechains? This is a useful question to ask, since it says something about the contributors to the energy of binding among the various types of hydrogen bonds. But there is a different point of view we could take. We might instead be interested in the likelihood of a particular donor or acceptor being involved in a hydrogen bond. This is a different question because the numbers of donors and acceptors are different for the mainchain versus the sidechains. Exploring this question reveals typical issues that have to be dealt with when datasets are examined from different perspectives.

Some of the differences in numbers of mainchain-mainchain hydrogen versus bonds involving sidechains can be explained by the differences in the number of hydrogen bond donors and acceptors.

Every mainchain unit can form hydrogen bonds, but not all sidechains can. In Table 6.2, we see that there are about one-third fewer (thirteen out of twenty) donors and acceptors for sidechains than mainchains. Note that the counts are simplifed because there are both thirteen donors and thirteen acceptors for sidechains. If the numbers were different, it would be more difficult to compare them with the mainchain donors and acceptors.

Making a correction for the differences in numbers of donors/acceptors narrows the gap somewhat, but the observed difference is still greater. To account for the deficit in donors and acceptors, we can multiply the number of M-S and S-M bonds by $20/13 \approx 1.54$, the ratio of potential donors/acceptors for M-S or S-M versus M-M bonds. The resulting number corresponds to 'virtual' bonds that would exist if the numbers of donors and acceptors were the same, and these numbers can be directly compared with the number of M-M bonds. Combining the number of M-S and S-M bonds (68,917) and multiplying by the factor $20/13$, we get about 106K 'virtual' bonds. For the sidechain-sidechain bonds, we need to multiply by $(20/13)^2 \approx 2.37$, yielding about 78K 'virtual' bonds, for a total of 186K 'virtual' bonds. Thus the likelihood of forming mainchain-mainchain bonds could be viewed as about a quarter more frequent than formation of a sidechain bond. Nevertheless, in terms of energy budget, the mainchain-mainchain bonds remain dominant.

The likelihood of finding a sidechain involved in a hydrogen bond depends on the likelihood of finding that sidechain in a protein. Thus a more careful analysis would involve the frequencies of individual residues in proteins (cf. Table 7.2). We leave this task as Exercise 15.2.

15.2   Proline-like configurations

Proline is the unique residue that turns back and forms a second covalent bond to the backbone, as shown in Figure 4.6. This forms a very rigid configuration which thus has special properties. However, other residue configurations can form ring-like structures with somewhat the same character, based on hydrogen bonds formed by the sidechains with the nearby amide or carbonyls on the backbone. We review this interesting behavior here and note the relationship to wrapping.

One feature of hydrogen bonds between mainchains and sidechains is that a large portion of them involve a bond between the sidechain and the amide or carbonyl of their own peptide, or of a neighboring peptide in the sequence. In this way, these sidechains form a structure that is similar to that of proline, but with the covalent bond in proline replaced by a weaker hydrogen bond.

One thing that characterizes such local attachments is that they tend to be underwrapped compared to hydrogen bonds between sidechains and mainchains that are more distant in sequence. The distribution of wrapping of all hydrogen bonds in this set of protein structures is depicted in Figure 15.1. First of all, we need to say how the wrapping of bonds involving sidechains was computed. To keep within the framework used for estimating wrapping of mainchain-mainchain bonds (Chapter 8), we used again a desolvation domain consisting of spheres centered at the two C_α carbons of the corresponding peptides. However, it should be noted that hydrogen bonds involving sidechains are quite different, and it is likely that a different metric would be more appropriate (and accurate).

We see that the wrapping of all hydrogen bonds involving sidechains is less than that for mainchain-mainchain bonds, with the deviation greatest for the M-S bonds for which the sidechain
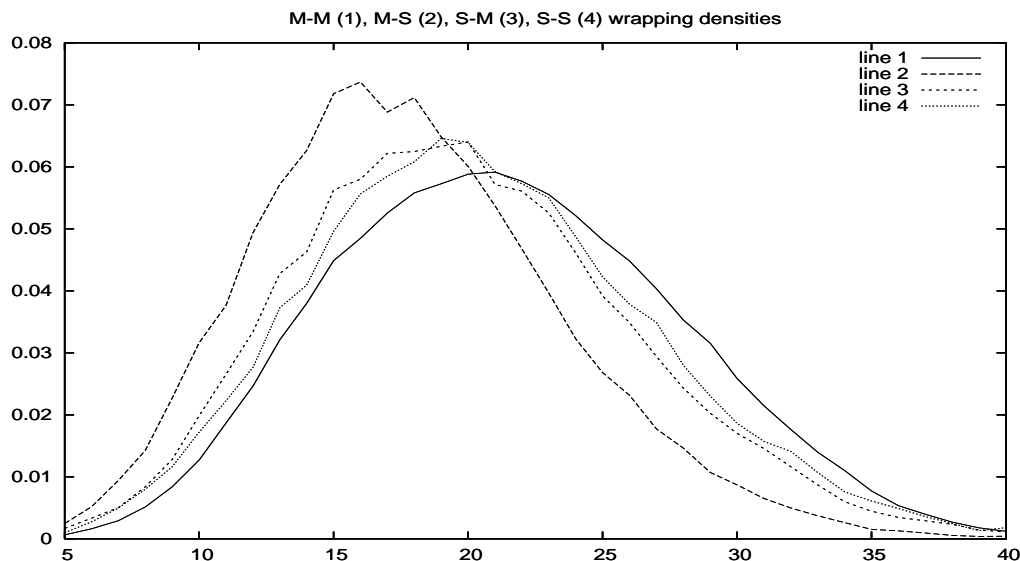
Figure 15.1: Distribution of wrapping for hydrogen bonds in a subset of the PDB Select structures. Bonds involving self-attachment, which have a significantly smaller desolvation domain, have been excluded. Solid line: M-M; large-dash line: M-S; small-dash line: S-M; dotted line: S-S. Desolvation radius 6 Å.

forms the acceptor of the bond, the donor being a mainchain amide group. An underwrapped mainchain amide or carbonyl would be a likely target for water attachment. Thus the structural defect associated with underwrapping appears to be corrected by certain sidechains making hydrogen bonds with the exposed backbone amides or carbonyls. Moreover, the formation of the hydrogen bond also removes the sidechain from water exposure as well. The major contributors to this motif are Thr, Ser and Asp, which have relatively little sidechain flexibility, so only a limited amount of entropy is lost in these associations.

We should note that we are only considering sidechain bonds that are formed *within the same chain*. That is, the intermolecular bonds formed between different chains are not counted here.

15.2.1 Nearest neighbor connections

Some sidechains can form a hydrogen bond with their own mainchain peptide group. For example, Asp and Glu can form hydrogen bonds between their terminal oxygens and the NH group on the backbone. An example of this is found in the PDB file 1NDM in the bond between the NH and OE1 of B-GLU306. In this bond, the N-O distance is only 2.78 Å(cf. Table 6.1, second row), and the angle between the NH and the CO is quite favorable. This is depicted in Figure 15.2(a). A similar bond is formed by A-GLU81 in 1NDM, and the N-O distance is only 2.37 Å. In the homologous structure in 1NDG, we find the second of these motifs (A-GLU81) repeated, as well as two more: A-GLU123 and B-GLU301. But the simple motif involving B-GLU306 becomes a complex of two Glu's (B-GLU 306 and B-GLU 405) with symmetric mainchain-sidechain bonds (N-H bonded to OE's) between the two as well as a self-bond (B-GLU 405), as shown in Figure 15.2(b).

(a)   (b)
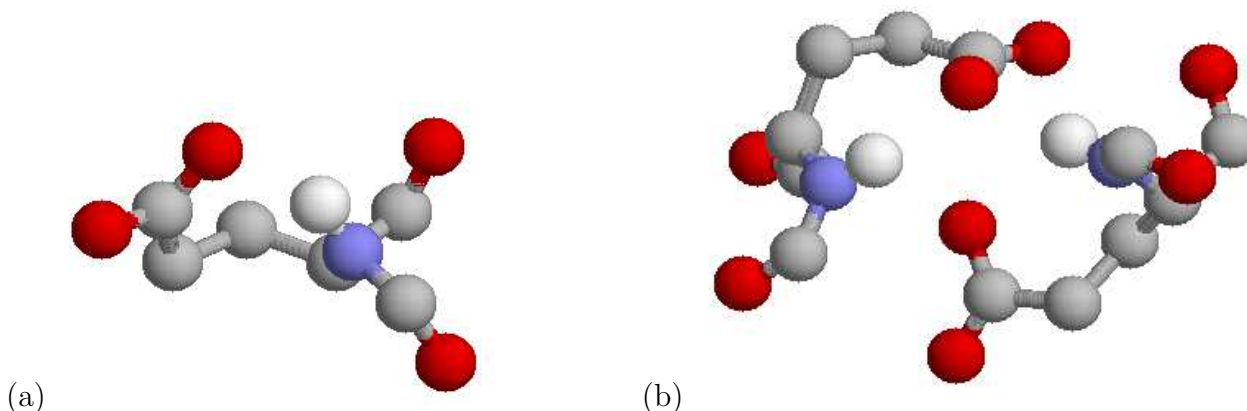
Figure 15.2: Ball and stick representation of the main atoms. Oxygens in red, nitrogen in blue, carbons in light grey, and hydrogen in white. Hydrogen positions have been estimated. (a) B-GLU306 sidechain in the PDB file 1NDM, including the C-O group from peptide 305. (b) B-GLU306 (lower right) and B-GLU 405 (upper left) complex in the PDB file 1NDG, including the C-O groups from sidechains 305 and 404.

The analogous type of bond can be formed with Gln, as in 1MPA (H-GLN113) and 1WEJ (H-GLN109). Although the turn is tighter, this motif also occurs with Asp: in 2H1P (H-ASP480) and in 2BSR (A-ASP106). And similarly, the motif occurs with Asn: in 1CU4 (L-ASN138), 1E4W (P-ASN4) and 1JRH (I-ASN53).

A similar type of motif can occur with Asn, in which the terminal amide group bonds with the backbone oxygen of the same residue, such as in 1IC4 (Y-ASN37), 1JRH (I-ASN62) and 1MPA (H-ASN3). With Thr, the terminal OH group can bond with the backbone oxygen of the same residue as well, as in 1NDM (B-THR431) and 1DQM (H-THR132). With Ser, the terminal OH group can bond with the backbone oxygen of the same residue as well, as in 2H1P (L-SER32) and 1IGC (L-SER202).

The terminal NH_3 group on lysine can bond with its own backbone oxygen, as occurs in 1WEJ (H-LYS136). A related type of motif can occur with Gln, in which the terminal amide group bonds with the oxygen of the preceding residue, such as (A-GLN89 NE2 — A-GLN90 O) in both PDB files 1NDG and 1NDM. This can also happen with Ser, with the terminal oxygen bonding with the next backbone amide group, as in 1DQJ (C-SER86 N—C-SER85 OG).

In Table 15.1, we tabulate the occurrences all of the observed local bonds where the sidechain bonds to its own backbone. This data is taken from a subset of the PDB Select database consisting of 1547 proteins complexes (PDB files) having 68,994 hydrogen bonds between sidechains and mainchains. (This is to be contrasted with a total of 233,879 mainchain-mainchain hydrogen bonds, and 33,217 sidechain-sidechain hydrogen bonds, 152 of which involve terminal oxygens.) If we classify bonds according to whether the sidechain is the donor (S-M) or acceptor (M-S), then we find that 30,642 of the total hydrogen bonds between sidechains and mainchains have the sidechain as acceptor (M-S), not including another 75 which involve terminal oxygens. Correspondingly, the remainder (38,277) are S-M bonds. Thus the special bonds tabulated in Table 15.1 represent

offset	residue	type	donor	acceptor	frequency	wrapping	corrected
0	ASN	M-S	N	OD1	116	8.5	14.4
0	HIS	M-S	N	ND1	236	9.0	15.3
0	GLN	M-S	N	OE1	267	10.2	17.3
0	ASP	M-S	N	OD1	478	8.7	14.7
0	GLU	M-S	N	OE1/2	981	10.5	17.8
0	CYS	S-M	SG	O	24	8.0	13.6
0	LYS	S-M	NZ	O	55	9.6	16.3
0	SER	S-M	OG	O	67	6.8	11.5
0	THR	S-M	OG1	O	79	8.3	14.1
0	GLN	S-M	NE2	O	83	11.0	18.7
0	ASN	S-M	ND2	O	119	7.7	13.0
0	ARG	S-M	Nx	O	224	10.0	17.0
0	ARG	S-M	NH2	O	19	9.7	16.4
0	ARG	S-M	NE	O	65	9.7	16.4
0	ARG	S-M	NH1	O	140	10.1	17.1

Table 15.1: Observed local hydrogen bonds between a residue and its own mainchain amide or carbonyl groups. The bonds made by arginine are further subdivided according to the specific hydrogen bond donor group. The corrected wrapping values in the final column are simply the wrapping value times 1.7, to account for the difference in size of desolvation domain (desolvation radius 6 Å). See Figure 15.3 parts (a) and (b).

2,729 (2,078 M-S and 651 S-M) hydrogen bonds, or about 4% of the total hydrogen bonds between sidechains and mainchains (6.7% of M-S and 1.7% of S-M).

The hydrogen bonds in Table 15.1 appear at first to be extremely underwrapped, but it must be remembered that in the case of a self-bond, the definition of the desolvation domain would involve only one sphere. Thus the desolvation domain is about 40% smaller than a desolvation domain consisting of two spheres centered on C_α's separated by 6 Å (for a sphere radius of 6 Å). Thus we might increase the wrapping numbers by about 70% in order to get a realistic comparison. With this correction (listed in the last column in Table 15.1), the bonds still appear slightly underwrapped.

The average amount of wrapping found in the same set of protein complexes was 17.8 for M-S hydrogen bonds and 19.9 for S-M hydrogen bonds (for a desolvation sphere radius of 6 Å), with the bonds removed from the calculation that involve the same residue, due to the smaller desolvation domain in that case. The distribution is shown in Figure 15.1. Adjusting these for a 40% decrease in volume of the desolvation domain would suggest expected values of 10.7 (M-S) and 11.9 (S-M), respectively. For reference, the mean wrapping of mainchain-mainchain hydrogen bonds for this data set is 21.4 and 20.5 for sidechain-sidechain hydrogen bonds. Thus we see that, in general, sidechain hydrogen bonds are less well wrapped than mainchain-mainchain bonds, with M-S hydrogen bonds significantly less well wrapped.

In particular, we see that both Glu and Gln are about as well wrapped when they make M-S
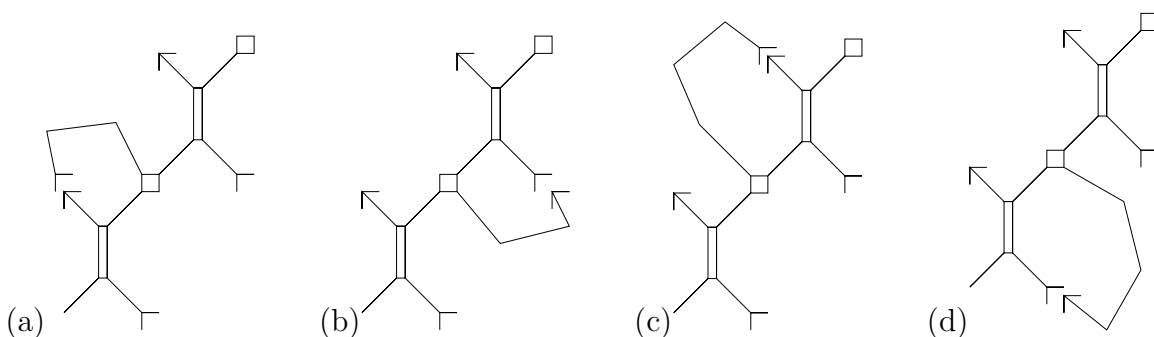
Figure 15.3: Hydrogen bond-forming sidechain configurations in the case of (a) zero-offset, mainchain-sidechain (M-S) bonds; (b) zero-offset, sidechain-mainchain (S-M) bonds; (c) +1 off-set, mainchain-sidechain (M-S) bonds; and (d) -1 offset, sidechain-mainchain (S-M) bonds. See Table 15.1 for (a) and (b) and Table 15.2 for (c) and (d).

hydrogen bonds to their own backbone amide groups as M-S hydrogen bonds are in general. On the other hand, Asp and Asn tend to be significantly less well wrapped when they make M-S hydrogen bonds to their own backbone amide groups compared with M-S hydrogen bonds are in general. The residues of Asp and Asn themselves contribute one less wrapper to the desolvation domain, but even adding one to the corrected wrapping values (to account for the additional intrinsic wrapper in the sidechains of Glu/Gln versus Asp/Asn) leaves their mean wrapping values more than two lower than the average of M-S hydrogen bonds in general.

On the other hand, all of the sidechains that form S-M bonds to their own carbonyl groups are, on average, significantly underwrapped compared to the average wrapping (19.9) of S-M bonds. The only exception to this is Gln, whose average wrapping in this configuration is only one less than the average.

The configuration of the sidechains relating to the data in Table 15.1 is depicted in Figure 15.3(a-b). We can see that these are the two closest possible mainchain locations for hydrogen bonding by the sidechain. But we also realize that other locations are also quite close, involving nearest sequence neighbors. These possible hydrogen bonds are depicted in Figure 15.3(c-d). In Table 15.2, we list all of the observed local bonds that can occur where the sidechain bonds to the nearest position on the backbone of its sequence neighbor. The special bonds tabulated in Table 15.2 represent 1,321 M-S and 920 S-M hydrogen bonds, or about 4.3% of M-S and 2.4% of S-M hydrogen bonds. Thus the combined M-S bonds involving sidechain hydrogen bonds with the amide group on either the same peptide or the subsequent peptide constitute 11% of all M-S hydrogen bonds.

The desolvation domains for sequence-neighbor residues will also be slightly smaller in size. The mean separation of C_α's in sequence neighbors is about 3.84 Å, and thus the desolvation domain (with radius 6 Å) is about 13% smaller in volume that a desolvation domain where the C_α's are a typical 6 Å apart. However, even with this correction, the amount of wrapping depicted is still significantly depressed from the expected averages (17.8 for M-S and 19.9 for S-M hydrogen bonds). Indeed, the volume of the desolvation domain depends on the distance between C_α atoms used in its definition, but taking a 6 Å separation as typical, we see that when the separation varies from 5

offset	residue	type	donor	acceptor	frequency	wrapping
+1	HIS	M-S	N	ND1	38	12.0
+1	GLN	M-S	N	OE1	69	14.1
+1	THR	M-S	N	OG1	95	13.5
+1	SER	M-S	N	OG	105	13.3
+1	GLU	M-S	N	OE1/2	138	14.9
+1	ASN	M-S	N	OD1	249	14.1
+1	ASP	M-S	N	OD1/2	627	15.0
-1	ASN	S-M	ND2	O	19	16.2
-1	CYS	S-M	SG	O	38	14.8
-1	GLN	S-M	NE2	O	42	15.7
-1	LYS	S-M	NZ	O	120	15.6
-1	ARG	S-M	Nx	O	179	15.0
-1	ARG	S-M	NH2	O	42	15.3
-1	ARG	S-M	NE	O	54	15.2
-1	ARG	S-M	NH1	O	83	14.7
-1	THR	S-M	OG1	O	236	14.7
-1	SER	S-M	OG	O	286	14.7

Table 15.2: Observed local hydrogen bonds between the residue and the nearest mainchain amide or carbonyl groups of its sequence neighbors. The bonds made by arginine are further subdivided according to the specific hydrogen bond donor group. Nx refers to the collection of the three NH groups; the frequency is the sum of the frequencies and the wrapping is the average. See Figure 15.3 parts (c) and (d).

Å to 7 Å, the corresponding volume variation is no more that 6% (for a desolvation radius of 6 Å), cf. Exercise 15.1.

The next closest positions for a residue to make hydrogen bonds with the mainchain of its sequence neighbor are depicted in Figure 15.4. In Table 15.3, we list all of the observed local bonds that can occur where the sidechain bonds to the nearest position on the backbone of its sequence neighbor. The special bonds tabulated in Table 15.2 represent 64 M-S and 842 S-M hydrogen bonds, or about 0.2% of M-S and 2.2% of S-M hydrogen bonds.

There are a few examples of bonds that occur only rarely in the dataset considered here. These are collected in Table 15.4.

15.2.2   Further neighbors

By contrast, once we look beyond nearest sequence neighbors, the picture changes dramatically. There are 7,969 M-S hydrogen bonds between the sidechain of residue k and the amide group on peptide $k + 2$. This represents over a quarter of all M-S bonds in this dataset. Three fifths of them involved either Asp or Asn. However, the mean wrapping for these hydrogen bonds is still low:
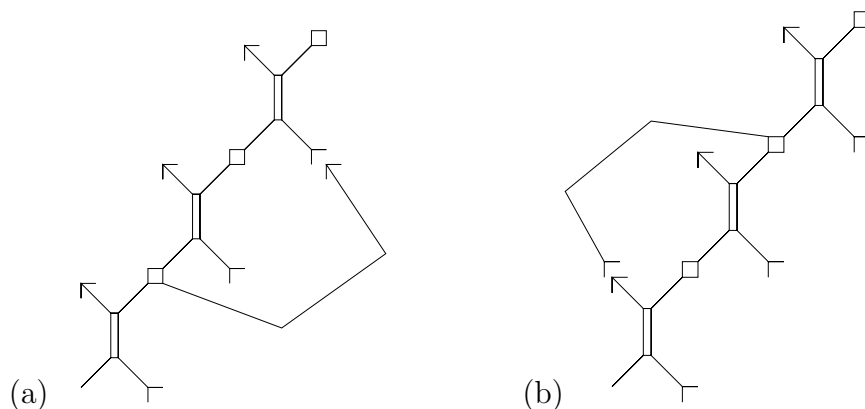
Figure 15.4: Configuration of the (a) +1 offset sidechains for the sidechain-mainchain (S-M) configuration, and the (b) -1 offset sidechains for the mainchain-sidechain (M-S) configuration. See Table 15.3.

offset	residue	type	donor	acceptor	frequency	wrapping
+1	CYS	S-M	SG	O	21	14.0
+1	GLN	S-M	NE2	O	22	13.8
+1	ASN	S-M	ND2	O	48	13.6
+1	LYS	S-M	NZ	O	62	13.7
+1	SER	S-M	OG	O	177	11.1
+1	THR	S-M	OG1	O	217	13.2
+1	ARG	S-M	Nx	O	295	16.3
+1	ARG	S-M	NH2	O	35	15.9
+1	ARG	S-M	NE	O	45	15.9
+1	ARG	S-M	NH1	O	215	16.5
-1	ASN	M-S	N	OD1	6	11.0
-1	GLN	M-S	N	OE1	8	10.1
-1	GLU	M-S	N	OE1/2	50	11.6

Table 15.3: Observed local hydrogen bonds between the residue and the next nearest mainchain amide or carbonyl groups of its sequence neighbors. The bonds made by arginine are further subdivided according to the specific hydrogen bond donor group. Nx refers to the collection of the three NH groups; the frequency is the sum of the frequencies and the wrapping is the average. See Figure 15.4.

offset	residue	type	donor	acceptor	frequency	wrapping
0	THR	M-S	N	OG1	1	13.0
-1	HIS	S-M	NE2	O	1	8.0
-1	TRP	S-M	NE1	O	1	12.0
-1	HIS	M-S	N	ND1	2	9.5
-1	THR	M-S	N	OG1	2	9.0
-1	ASP	M-S	N	OD1	3	15.7

Table 15.4: Rarely observed local hydrogen bonds between the residue and various nearby mainchain amide or carbonyl groups of its sequence neighbors.

offset	residue	type	donor	acceptor	frequency	wrapping
0	ASN	M-S	N	OD1	116	(14.4)
0	ASP	M-S	N	OD1	478	(14.7)
+1	ASN	M-S	N	OD1	249	14.1
+1	ASP	M-S	N	OD1/2	627	15.0
+2	ASN	M-S	N	OD1	1613	14.9
+2	ASP	M-S	N	OD1	3175	14.2

Table 15.5: Comparison of wrapping for Asp and Asn M-S bonds in different contexts. Wrapping values for bonds to their own peptide are the corrected values, shown in parentheses, from Table 15.1.

14.2 for Asp and 14.9 for Asn. In Table 15.5, we compare the data on these configurations with the data on Asp and Asn that bond with closer neighbors. In these configurations, Glu and Gln are relatively less frequent, occurring only 153 and 48 times, respectively. Ser (1530) and Thr (1052) are more strongly represented in this group, while His (279) and Cys (117) are less other commonly occurring sidechains found in $k + 2$ M-S bonds. The number of S-M bonds in this configuration is much smaller, having only 1180 occurrences. Similarly there are only 128 M-S, and 708 S-M, hydrogen bonds between the sidechain of residue k and peptide $k - 2$.

15.3   All sidechain hydrogen bonds

To amplify the assessment of attachments of sidechains to mainchain amides and carbonyls on peptides nearby in sequence, we now analyze sidechain hydrogen bonds with mainchains in general. In Table 15.6, we collect some pertinent statistics, again drawn from the same subset of 1547 proteins complexes from the PDB Select database.

Some explanations are required for the data in Table 15.6. The column 'type' indicates whether the sidechain forms a bond with an amide (M-S) or a carbonyl (S-M). That is, the first letter indicates the hydrogen bond donor and the second letter indicates the hydrogen bond acceptor. The column 'count' gives the total number of residues of this type in the PDB Select subset studied. The

res.	type	count	bonds	M	%-loc	NNW	in-W	out-W	%-out
ARG	S-M	63565	8913	5	14.49	15.87	17.61	20.00	64.31
ASN	M-S	52997	4571	2	67.11	14.66	15.78	18.63	25.86
ASN	S-M		1688	4	19.98	14.37	17.90	19.08	58.41
ASP	M-S	68906	10185	2	61.04	14.91	15.44	18.09	25.51
CYS	M-S	29731	285	2	66.20	NA	14.83	15.97	24.91
CYS	S-M		1052	4	69.36	14.43	21.26	20.24	16.83
GLN	M-S	49222	1745	-3	28.27	13.53	17.69	20.26	42.92
GLN	S-M		1365	2	31.16	15.05	17.47	20.12	47.91
GLU	M-S	84047	4555	0	31.40	14.07	16.70	19.89	36.55
HIS	M-S	27377	784	2	77.14	11.90	16.88	19.23	16.20
HIS	S-M		763	3	8.14	8.00	19.24	19.77	79.03
LYS	S-M	85292	4191	3	21.26	14.96	18.13	20.17	61.66
SER	M-S	71354	4496	2	69.12	13.42	15.45	17.94	21.57
SER	S-M		8298	4	59.46	13.31	16.60	18.45	24.57
THR	M-S	65455	3272	2	69.58	13.48	15.68	19.08	22.68
THR	S-M		9127	4	64.09	14.14	17.61	19.11	21.96
TRP	S-M	16203	1119	5	15.30	12.00	25.08	26.98	70.78
TYR	M-S	39743	747	-5	13.14	NA	24.72	26.78	74.97
TYR	S-M		1761	4	10.34	NA	25.19	27.38	80.01

Table 15.6: Key: 'type' of bond (see text); 'count' is the number of residues of this type in the PDB Select subset; 'bonds' is the number of hydrogen bonds of the specified type involving this residue; M is the mode of the distribution of sequence distances between a sidechain and mainchain making a hydrogen bond (the largest number of sidechains i are bonded to the mainchain amide or carbonyl of peptide $i + M$); %-loc= percentage of hydrogen bonds with sequence distances of the form $M \pm i$ for $|i| \leq 2$; NNW=average wrapping for the nearest neighbors (sidechain i bonded to the mainchain amide or carbonyl of peptide $i \pm 1$; NA indicates that there are no such bonds); in-W=average wrapping for sidechain i bonded to the mainchain amide or carbonyl of peptide $i + j$ for $1 < |j| \leq 10$; out-W=average wrapping for sidechain i bonded to the mainchain amide or carbonyl of peptide $i + j$ for $|j| > 10$; %-out=percent of sidechains i bonded to the mainchain amide or carbonyl of peptide $i + j$ for $|j| > 10$;
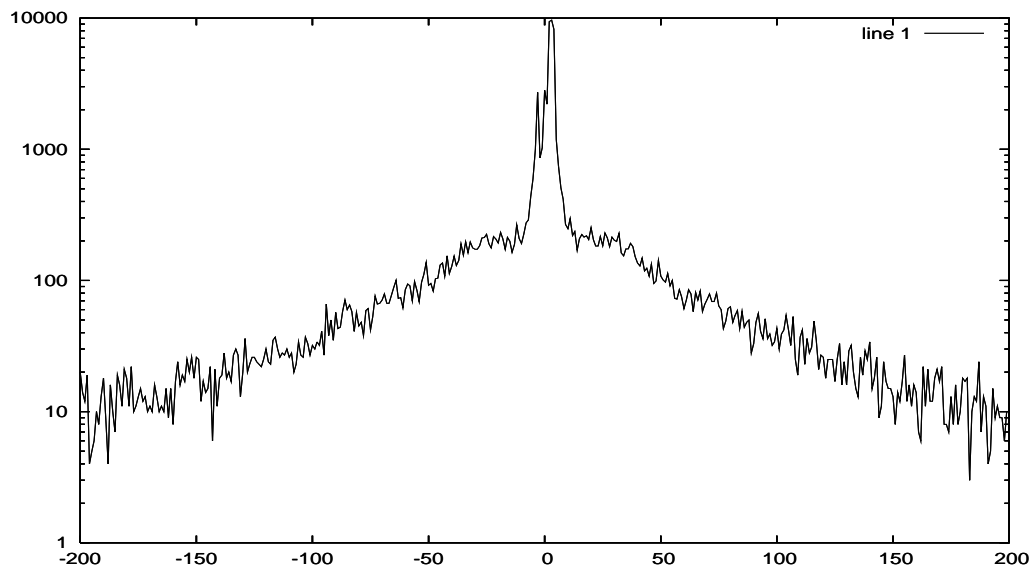
Figure 15.5: Distribution of sequence distances between donors and acceptors in hydrogen bonds between mainchains and sidechains. Both M-S and S-M bonds are included.

column 'bonds' indicates the number of hydrogen bonds of the specified within the total dataset.

The mode 'M' refers to the distribution of sequence distances between a sidechain and mainchain making a hydrogen bond. More precisely, we form the frequency distribution of sidechains with sequence number i that are bonded to the mainchain amide or carbonyl of peptide $i+j$ as a function of j. These distributions are highly peaked (cf. Figure 15.7), and the mode provides a useful statistic to characterize them. The mode M of this distribution of sequence distances is the number such that the largest number of sidechains i makes bonds with peptide $i + M$. For example, $M = 0$ means that the majority of the sidechains are bonded to their own mainchain.

The distribution of all sequence distances for all types of sidechains is shown in Figure 15.5. We see that it is heavily concentrated on small distances, and Figure 15.6 provides a view of the distribution in this region. From this, we conclude that a substantial fraction of the distribution is concentrated for distances of magnitude ten or less, and that the character of the distribution changes outside this region. There is a significant difference in the distributions for M-S versus S-M bonds, and these differences are contrasted in Figure 15.6 as well.

The subsequent column (%-loc) helps to characterize further the distribution of sequence distances between a sidechain and mainchain making a hydrogen bond. It gives the percentage of sidechains with sequence number i that are bonded to the mainchain amide or carbonyl of peptide $i + M + j$ for $|j| \leq 2$. When this percentage is large, it indicates how peaked the distribution is around its mode M. The smaller percentages indicate distributions that are more spread out. Typical distributions are highly peaked around the mode, as indicated in Figure 15.7 for the M-S bonds for Thr, Ser and Asp. The distributions for the S-M bonds for Thr and Ser are similar but just shifted to the right by two units, corresponding to having $M = 4$.

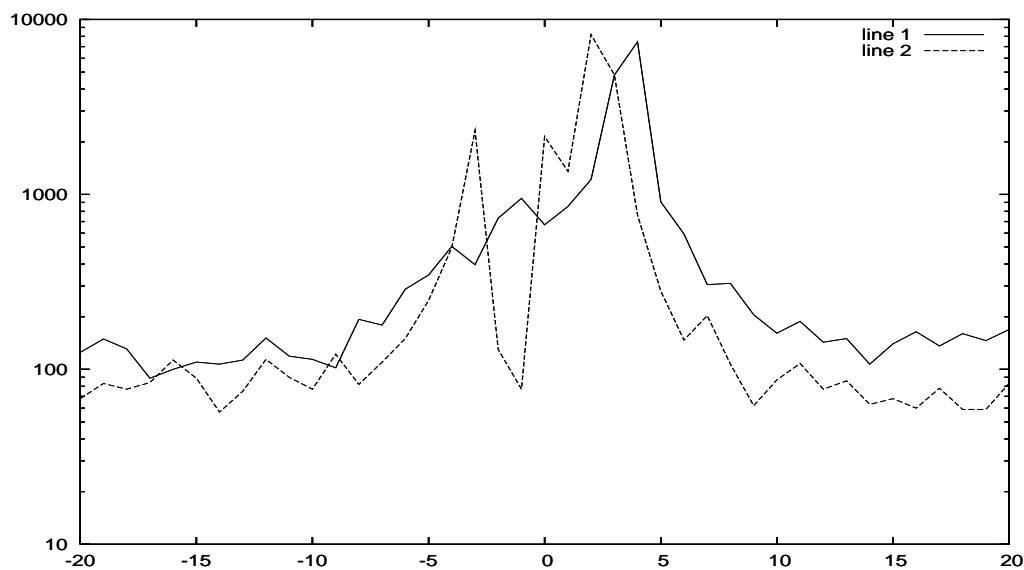The column NNW gives the average wrapping for sidechains i bonded to mainchains $i \pm 1$. If

Figure 15.6: Distribution of sequence distances between donors and acceptors in hydrogen bonds between mainchains and sidechains. Solid line: S-M bonds; dashed line: M-S bonds.
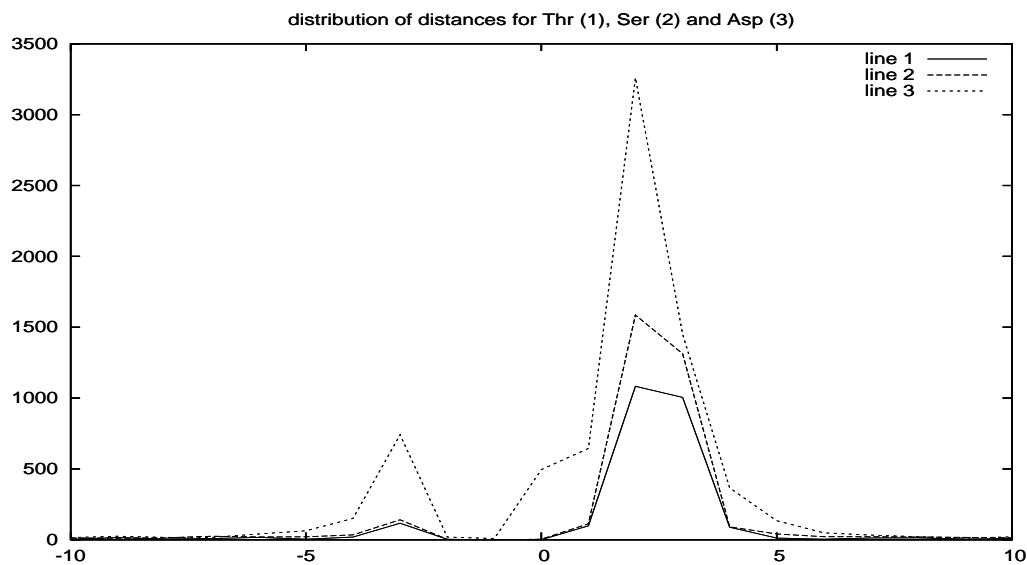

Figure 15.7: Distribution of sequence distances for M-S hydrogen bonds for Thr (solid line), Ser (dashed line) and Asp (dotted line).

there are no such hydrogen bonds, then an average cannot be formed, and this is indicated in the table by NA. This special group is singled out due to the fact that the desolvation domain is about 13% smaller than is typical. This was done to avoid contamination of the assessment 'in-W' of the wrapping of hydrogen bonds between sidechains with sequence number i that are bonded to the mainchain amide or carbonyl of peptide $i + j$ for $1 < |j| \leq 10$. Self-bonding ($i = j$) was also eliminated in computing the 'in-W' statistics (these data are found in Table 15.1).

In most cases, the ratio of in-W to NNW is about what would be expected due to the slight difference in sizes of the desolvation domains. In the cases where there is a significant difference, the set of sidechains i bonded to mainchains $i \pm 1$ is quite small. The column 'out-W' lists the average wrapping of hydrogen bonds between sidechains with sequence number i that are bonded to the mainchain amide or carbonyl of peptide $i + j$ for $|j| > 10$. These are the hydrogen bonds without any bias due to locality in sequence. The distribution of sequence distances extends into the hundreds in each direction. The percentage of such hydrogen bonds is indicated by %-out. In some cases, this is the minority of bonds, but the percentages are still large enough in the important cases to give a good estimate of the average wrapping for non-local sidechain-mainchain bonds of the particular types.

What is most striking about the data in Table 15.6 is that the local average wrapping, as indicated by NNW and in-W, is significantly less than the non-local average wrapping, as indicated by out-W. That is, the indicated sidechain-mainchain hydrogen bonds are far more likely to occur with nearby sidechain-mainchain pairs when there is a local wrapping deficit. These sidechains tend to correct the wrapping defect by forming hydrogen bonds with the mainchain.

A significant fraction of certain residues are devoted to these local bonds formed in underwrapped environments. For example, 19% of threonine residues form sidechain bonds with the mainchain, as do 18% of the serines and 15% of aspartates. The majority of these bonds are made with near sequence neighbors and are deficiently wrapped.

The distribution of distances for arginine is somewhat of an exception from the others. It is still quite localized, as indicated in Figure 15.8. However, it is spread more broadly over -10 to +10, rather than being concentrated in a region ± 2 around its mode ($M = 5$).

15.4 Unusual hydrogen bonds

15.4.1 Hydrophobic pairs

When Val is paired with Val in a backbone-backbone hydrogen bond it is frequently the case that there is a pair of bonds. For example in 1CU4, there are both H-VAL133N—H-VAL180O and H-VAL180N—H-VAL133O bonds. This type of pair is also found in 1CU4 and 1E4W (133,180 and 194,203), 1F90 (138,193 and 210,219), 1IGC (143,190 and 204,213), 1JRH (210,219), 1MPA (144,191 and 205,214), 1WEJ (140,187 and 201,210), 2CII (199,249), and 2H1P (443,490 and 504,513).

A similar Met pair is in 1MPA (20,81), and a His pair is found in 1G6V (94,119). The Ile pair (34,51) appears in 1CU4. Tyr pairs appear frequently: the pair (142,172) appears in both 1CU4 and 1E4W, the pair (7,99) appears in 2BSR, 2BSS, 2BVO, and 2BVQ, and the pair (36,87) is in
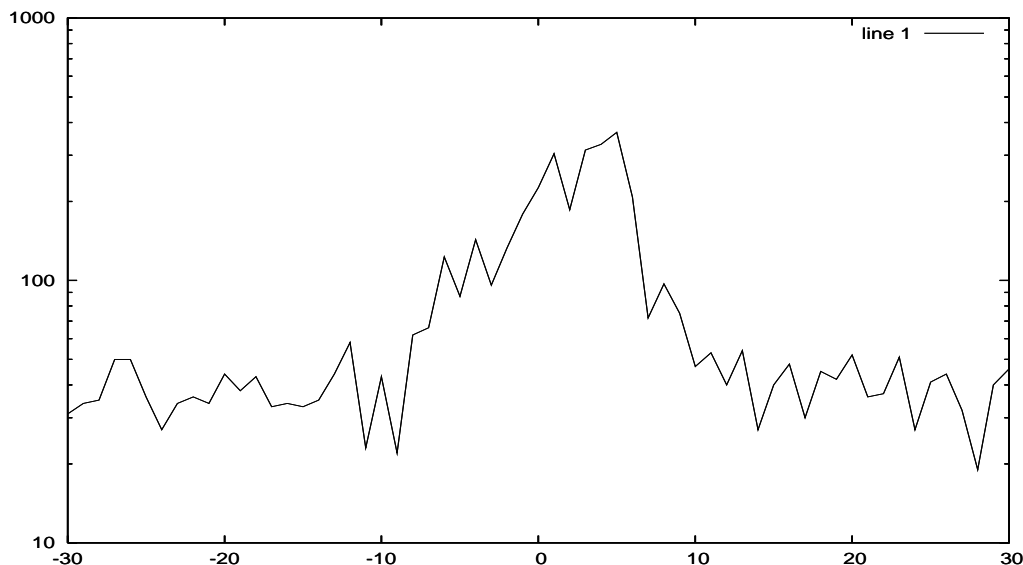
Figure 15.8: Distribution of sequence distances between donors and acceptors in arginine S-M bonds.

1JRH. Phe pairs are also common: (208,241) in 2BSR, 2BSS, 2BVO, and 2BVQ; (30,62) in 2BSR and 2CII; and (36,87) in 1E4W.

15.4.2   Unusual trios

The trio of arginines (75,79,83) appears in 2BSR, 2BSS, and 2BVO. These are involved in a sequence of mainchain hydrogen bonds: A-ARG 83 N — A-ARG 79 O, A-ARG 79 N — A-ARG 75 O. A similar trio of lysines (9,55,103) is in 1IGC. Such a trio of residues forms an unusual charged structure. A similar type of trio also occurs with Phe: H-PHE32 N — H-PHE29 O , H-PHE29 N — H-PHE27 O is found in 1IGC.

15.5   Exercises

Exercise 15.1 *Let $V_{\rho,R}$ be the volume of the desolvation domain when the desolvation radius is ρ and the separation between the C_α is R. When $R \geq 2\rho$, $V_{\rho,R} = 2V_{\rho,0}$. Compute the change in volume in the desolvation domain as a function of the separation of the C_α atoms used in its definition. Define a separation parameter $r = R/2\rho$, so that $r = 0$ is the case when the desolvation spheres coincide, and $r = 1$ is the point at which they become completely separated. Verify that the ratio of volumes $\phi(r) = V_{\rho,R}/V_{\rho,2\rho}$ may be written as*

$$\phi(r) = \tfrac{1}{2} + \tfrac{3}{4}(r - \tfrac{1}{3}r^3) \tag{15.1}$$

and is in particular independent of ρ.

Exercise 15.2 *Repeat the analysis in Section 15.1 but count the potential frequency of bond formation by including the relative frequencies of each amino acid based on the data in Table 7.2.*

Exercise 15.3 *Determine if wrapping correlates (or not) with the formation of mainchain-mainchain hydrogen bonds. Consider the average wrapping of carbonyls and amides that do not form hydrogen bonds and compare this with what is found in Figure 15.1. Note that there is a significant number of amide and carbonyl groups that are well wrapped and yet form no hydrogen bonds; these are the peptides in group Ib in Section 13.3.*

Exercise 15.4 *Sidechains that form local attachments to the mainchain presumably alter the local structure. Determine if the ϕ and ψ angles, or other mechanical properties (Section 5.2), are different when these local attachments are formed.*

Chapter 16

Continuum equations for electrostatics

The dielectric properties of materials are important in many contexts [46, 285, 375]. A **dielectric medium** [54] is often characterized by having charges organized in local groups with net charge zero that are free to rotate to align with the ambient field. The properties of a dielectric can be dervived from first principals that take into account the polar properties of the dielectric constituents.

In Section 3.2, we describe the basic relationship between charge and electric potential and force. Using (3.2) and assuming linearity, the basic equations of electrostatics for a collection of charges of strength q_i at positions $\mathbf{r}_i$ in a vacuum can be derived from the simple expression

$$\nabla \cdot (\epsilon_0 \mathbf{e}) = \sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i), \tag{16.1}$$

where ϵ_0 is the permittivity of the vacuum (see Section 17.2.3). Here $\mathbf{e}$ is the induced electric field.

Using the ideas in Section 3.2, the equation (16.1) can be solved: $\mathbf{e} = -\nabla \phi$ where

$$\phi(\mathbf{r}) = \sum_i \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}, \tag{16.2}$$

where we have chosen physical units so that $\epsilon_0 = 1/4\pi$. But the ability to solve this so precisely is misleading. In interesting systems, the positions and orientations of many charge groups is unknown, and they must be deduced as part of the solution process.

A dielectric medium is a collection of polar molecules which react to the electric field in such a way as to moderate the overall charge. Our case of interest is primarily water, and we know water is a polar molecule. We can model the dielectric effect by representing these polar molecules as a collection of charges (cf. Section 10.5.4). Although this is just an approximation, we can imagine making better and better approximations by larger collections of charges at appropriate points to represent the polar nature of water. All such collections would of course have net charge zero for each water molecule.

16.1   Dielectric materials

We think of our charged system as consisting of two parts: fixed charges (ρ) and dielectric charges (γ). We will split the total charge distribution $\sum_i q_i \delta(\mathbf{r} - \mathbf{r}_i)$ into two parts, $\gamma + \rho$, where γ is the

part of the charge density corresponding to the dielectric, e.g., charge groups with net charge zero, and ρ denotes the remainder of the charge density (fixed charges). More precisely, we assume that the set of all charges is decomposed into two sets, I_ρ and I_γ, and

$$\rho = \sum_{i \in I_\rho} q_i \delta(\mathbf{r} - \mathbf{r}_i), \qquad \gamma = \sum_{i \in I_\gamma} q_i \delta(\mathbf{r} - \mathbf{r}_i). \tag{16.3}$$

Correspondingly, we define

$$\phi_\rho(\mathbf{r}) = \sum_{i \in I_\rho} \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}, \qquad \phi_\gamma(\mathbf{r}) = \sum_{i \in I_\gamma} \frac{q_i}{|\mathbf{r} - \mathbf{r}_i|}, \tag{16.4}$$

and $\phi = \phi_\rho + \phi_\gamma$. In particular,

$$\nabla \cdot (\epsilon_0 \nabla \phi_\rho)(\mathbf{r}) = \rho(\mathbf{r}), \qquad \nabla \cdot (\epsilon_0 \nabla \phi_\gamma)(\mathbf{r}) = \gamma(\mathbf{r}). \tag{16.5}$$

The key assumption is that the dielectric charges are free to self-organize in response to the overall electric field. For example, we imagine that water molecules can rotate so that their dipoles align with the electric field.

Define the **polarization** electric field $\mathbf{p} = -\nabla \phi_\gamma$, so that

$$\nabla \cdot (\epsilon_0 \mathbf{p}) = -\gamma. \tag{16.6}$$

Set $\mathbf{d} = \mathbf{e} + \mathbf{p}$. Then

$$\nabla \cdot (\epsilon_0 \mathbf{d}) = \nabla \cdot (\epsilon_0 \mathbf{e}) + \nabla \cdot (\epsilon_0 \mathbf{p}) = \rho + \gamma + \nabla \cdot (\epsilon_0 \mathbf{p}) = \rho. \tag{16.7}$$

The electric field $\mathbf{e}$ thus satisfies $\mathbf{e} = \mathbf{d} - \mathbf{p}$, where $\mathbf{d}$ can be interpreted as the field generated by the fixed charges ρ in the absence of the charges associated with the dielectric effect.

Debye [94] postulated that the mobile charge groups would orient to oppose the resulting electric field. In a thermalized system, there will be fluctuations in the orientations of these charge groups, and we can only talk about mean orientations. Debye's postulate can thus be written

$$\widetilde{\mathbf{p}} = s\,\widetilde{\mathbf{e}}, \tag{16.8}$$

where s can be thought of as a constant of proportionality in the simplest case, but more generally we will think of s as an operator that maps $\widetilde{\mathbf{e}}$ to $\widetilde{\mathbf{p}}$. Here, $\widetilde{\mathbf{p}}$ (resp., $\widetilde{\mathbf{e}}$) can be interpreted as a temporal average over a timescale that is long with respect to the basic thermal motions. (More precisely, it is a thermodynamic quantity that represents an average over many realizations of the system.) The thermal motions occur on the order of fractions of picoseconds, so we could imagine a time average of the order of picoseconds.

It is useful to define $\epsilon = \epsilon_0 s + \epsilon_0$. If we are thinking of s as an operator, then the second term is just the identity operator scaled by the permittivity coefficent. Debye's postulate can now be written as

$$\epsilon_0 \widetilde{\mathbf{p}} = (\epsilon - \epsilon_0)\widetilde{\mathbf{e}}, \tag{16.9}$$

where we will see that ϵ can be thought of as an effective permittivity (which may be a constant in the simplest case, or more general as an operator).

If one is uncomfortable with the Debye ansatz, we can define ϵ as follows. We decompose $\mathbf{p}$ into one part in the direction $\mathbf{e}$ and the other perpendicular to $\mathbf{e}$. That is, we write

$$\epsilon_0 \mathbf{p} = (\epsilon - \epsilon_0)\mathbf{e} + \zeta \mathbf{e}^{\perp}, \tag{16.10}$$

where ϵ is defined by

$$\epsilon = \epsilon_0 \left(1 + \frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}} \right), \tag{16.11}$$

with the appropriate optimism that $\mathbf{p} = 0$ when $\mathbf{e} = 0$. Here, we may assume that $\mathbf{e}^{\perp}$ is a unit vector. That is, $\epsilon - \epsilon_0$ reflects the correlation between $\mathbf{p}$ and $\mathbf{e}$. As defined, ϵ is a function of $\mathbf{r}$ and t, and potentially singular. However, Debye postulated that a suitable average

$$\widetilde{\epsilon} = \epsilon_0 \left(1 + \left\langle \frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}} \right\rangle \right), \tag{16.12}$$

should be well behaved. Implicitly, it is assumed that averages of ζ defined in (16.10) may be ignored.

However we interpret the Debye ansatz, we arrive at (16.8) as our model equation for the relationship between the electric field and the polarization field. For simplicity, we will drop the tildes and think from now on that everything represents temporal, or spatial, averages. Combining (16.9) (without the tildes) with the definition of $\mathbf{d}$, we conclude that

$$\epsilon_0 \mathbf{d} = \epsilon_0 \mathbf{e} + \epsilon_0 \mathbf{p} = \epsilon \mathbf{e}. \tag{16.13}$$

Thus (16.7) implies

$$\boxed{\nabla \cdot (\epsilon \nabla \phi) = \nabla \cdot (\epsilon \mathbf{e}) = \rho.} \tag{16.14}$$

This gives us an equation for $\mathbf{e}$ (or ϕ) in terms of ρ, provided we have some way of estimating ϵ. Looking carefully at the derivation, we see that ϵ could be any operator, even a nonlinear one. We will see that it is appropriate to think of ϵ as just a simple constant in many cases, but in others it will be more complicated.

The expression (16.12) provides an operational definition for a computationally determined dielectric constant. That is, in a molecular dynamics computation, one can define a local dielectric constant by averaging the correlation coefficient

$$\widetilde{\epsilon} - \epsilon_0 = \epsilon_0 \frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}}, \tag{16.15}$$

over space, over certain molecules, and/or time. This correlation coefficient need not be positive, so it is conceivable that $\widetilde{\epsilon} < \epsilon_0$, and we could even have $\widetilde{\epsilon} < 0$.

Monitoring the other term $\zeta \mathbf{e}^{\perp}$ in (16.10) also allows us to quantify the extent to which the dielectric model holds. Using the orthogonality we have

$$\begin{aligned} \zeta^2 &= \zeta \mathbf{e}^{\perp} \cdot \zeta \mathbf{e}^{\perp} = (\epsilon_0 \mathbf{p} - (\epsilon - \epsilon_0)\mathbf{e}) \cdot \zeta \mathbf{e}^{\perp} = \epsilon_0 \mathbf{p} \cdot \zeta \mathbf{e}^{\perp} \\ &= \epsilon_0 \mathbf{p} \cdot (\epsilon_0 \mathbf{p} - (\epsilon - \epsilon_0)\mathbf{e}) = \epsilon_0 \mathbf{p} \cdot \left(\epsilon_0 \mathbf{p} - \epsilon_0 \frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}} \mathbf{e} \right) \\ &= \epsilon_0^2 \mathbf{p} \cdot (\mathbf{p} - (\mathbf{p} \cdot \hat{\mathbf{e}})\hat{\mathbf{e}}) = \epsilon_0^2 \mathbf{p} \cdot (\mathbf{p} - \mathbf{p}_e) = \epsilon_0^2 \mathbf{p} \cdot \mathbf{p}_e^{\perp} \\ &= \epsilon_0^2 \mathbf{p}_e^{\perp} \cdot \mathbf{p}_e^{\perp}, = \epsilon_0^2 |\mathbf{p}_e^{\perp}|^2, \end{aligned} \tag{16.16}$$

where we denote by $\hat{\mathbf{e}}$ the unit vector in the direction of $\mathbf{e}$, $\hat{\mathbf{e}} = |\mathbf{e}|^{-1}\mathbf{e}$,

$$\mathbf{p}_e = (\mathbf{p} \cdot \hat{\mathbf{e}})\hat{\mathbf{e}} \tag{16.17}$$

is the projection of $\mathbf{p}$ in the direction of $\mathbf{e}$, and

$$\mathbf{p}_e^\perp := \mathbf{p} - \mathbf{p}_e. \tag{16.18}$$

The quantity of interest would be the ratio of $|\zeta|$ and $|\epsilon - \epsilon_0|\,|\mathbf{e}|$, since these are the two terms in (16.10). The first one can be computed via (16.16). The second is also easily calculated. Taking the dot product of (16.10) with $\mathbf{e}$, we find

$$|(\epsilon - \epsilon_0)|\,|\mathbf{e}|^2 = |(\epsilon - \epsilon_0)\mathbf{e} \cdot \mathbf{e}| = \epsilon_0|\mathbf{p} \cdot \mathbf{e}| = \epsilon_0|\mathbf{p} \cdot \hat{\mathbf{e}}||\mathbf{e}| = \epsilon_0|\mathbf{p}_e|\,|\mathbf{e}|. \tag{16.19}$$

Therefore

$$\frac{|\zeta|}{|\epsilon - \epsilon_0|\,|\mathbf{e}|} = \frac{|\mathbf{p}_e^\perp|}{|\mathbf{p}_e|}. \tag{16.20}$$

Thus the ratio of the sizes of the terms in the Debye ansazt is easily monitored by the expression (16.20). Recall that $\mathbf{e}$ is just the electric vector resulting from all of the charge groups, both fixed and free to rotate, and $\mathbf{p}$ is the polarization vector resulting from the free charges. The terms on the right-hand side of (16.20) are defined in (16.17) and (16.18).

16.1.1 Bulk water dielectric

A very good approximation is obtained in bulk water by the assumption that ϵ is just a constant factor. The model

$$\epsilon \approx 87.74 - 40.00\,\tau + 9.398\,\tau^2 - 1.410\,\tau^3, \tag{16.21}$$

where $\tau = T/100$ and T is the temperature in Centigrade (for $T > 0$), is supported by extensive experiments for liquid water [183]. Note that the formula (16.21) is only intended to apply for water, hence $\tau \in [0, 1]$.

This result is remarkable for many reasons. The fact that ϵ is much greater than one means that the strength of the opposing polarization field $\mathbf{p} = -\nabla\phi_\gamma$ is much greater than the field that induces it. The fact that ϵ increases with decreasing temperature implies that order is more important than mobility in the dielectric effect. When water freezes, it increases further: for ice at zero degrees Centigrade, $\epsilon \approx 92$.

This simple model of the dielectric in bulk cannot hold when the spatial frequencies of the electric field $\nabla\phi$ are commensurate with the size of a water molecule, since the water molecules cannot orient appropriately to align with the field. Thus frequency-dependent versions of ϵ have been proposed, and these are often called "nonlocal" models since the operator ϵ must be represented either as a Fourier integral (in frequency space), or as an integral in physical space with a nonlocal kernel [49, 365].

16.1.2   Temporal frequency dependence

The relationship proposed by Debye between $\mathbf{p}$ and $\mathbf{e}$ depends on temporal frequency. In particular, many measurements have shown [219]

$$\epsilon(\nu) = \epsilon_0 + \frac{\epsilon_1 - \epsilon_0}{1 + \nu^2 \tau^2} \tag{16.22}$$

where τ is a characteristic time associated with the dielectric material.

This relationship has been verified extensively by experimental data, including recent work [219]. For plane waves, there is a simple relationship between spatial wave number ξ and temporal wave number ν given by $c\xi = \nu$ where c is the speed of light. Thus plane waves with temporal frequency ν have spatial frequency $\xi = \nu/c$. This suggests using a form similar to (16.22) for the behavior of $\epsilon(\xi)$, although this form does not follow rigorously from the temporal experiments.

16.1.3   Spatial frequency dependence

We are interested in electric fields which are *not* time varying (i.e., $\nu = 0$) but rather spatially varying. It is easy to see that the driving electrical fields ρ generated by proteins have high frequencies. Salt bridges involve charge alternations on the order of a few Ångstroms. And polar sidechains such as Asn and Gln correspond to even higher frequencies, although at a smaller amplitude. Thus the fixed charge density ρ will have significant frequency content on the order of the size of a water molecule. These high frequencies in ρ may be primarily in directions along the surface of the protein, but there also must be high frequencies normal to the surface as well resulting from the jump in dielectric, as we now explain.

The dielectric coefficient varies by a factor of nearly one hundred from inside the protein to the bulk dielectric value away from the surface. This forces a kink in the electric field in the vicinity of the boundary in the direction normal to the surface of the protein. Kinks (discontinuities in the derivative) imply high frequency components in the electric field. Let us explain how such kinks can arise.

Suppose that the fixed charges are all within a domain Ω containing the protein, and that water is excluded from this domain. Then we expect $\epsilon = \epsilon_0$ inside Ω, and $\epsilon = \epsilon_1$ outside Ω. But the equation $\nabla \cdot (\epsilon \mathbf{e}) = \rho$ forces the quantity $\epsilon \mathbf{e}$ to have a continuous normal component across $\partial\Omega$, assuming there are no charges on $\partial\Omega$. Therefore $\mathbf{n} \cdot \mathbf{e}$ has to be discontinuous across $\partial\Omega$ to compensate for the discontinuity of ϵ (see Exercise 16.7).

Thus there are high-frequency components in the electric field both in the direction normal to the surface of the protein as well as in directions along the surface.

On the other hand, it is clear that the dielectric response has to go to zero for high frequencies. If the electric field varies at a spatial frequency whose wavelength is smaller than the size of a water molecule, the water molecule feels a diminished effect of that field component. Therefore, the dielectric coefficient must be a function of spatial wave number and go to zero for high frequencies.

Thus we adopt the ansatz that the dielectric properties depend on spatial wave number ξ proportional to a factor κ given by

$$\kappa(\xi) = \epsilon_0 + \frac{\epsilon_1 - \epsilon_0}{1 + |\lambda\xi|^2}, \tag{16.23}$$

where λ is a length scale determined by the size of the dielectric molecules. In general, λ could be a matrix, allowing for anisotropy. But for the time being we will think of it as a scalar.

We have anticipated that the dielecric coefficient may depend on space (and time), so we will be interested in cases where λ depends on the spatial variable $\mathbf{r}$, and in some cases the length scale will tend to infinity. For this reason, we introduce $\mu = 1/\lambda$, and write

$$\kappa(\mathbf{r}, \xi) = \epsilon_0 + \mu(\mathbf{r})^2 \frac{\epsilon_1 - \epsilon_0}{\mu(\mathbf{r})^2 + |\xi|^2} \tag{16.24}$$

where $\mu(\mathbf{r})$ is a spatial frequency scale. However, for simplicity we assume that the model (16.23) is sufficient for the moment.

16.1.4   Poisson-Debye equation: bulk case

We can expand $\mathbf{e}$ and $\mathbf{p}$ in a Fourier series and use the Debye-like relationship (16.23) to relate the resulting coefficients in the series. That is, we have (using the inverse Fourier transform)

$$\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi)\widehat{\mathbf{e}}(\xi) \, d\xi = \mathbf{p}(\mathbf{r}), \tag{16.25}$$

where here and subsequently we use the notation $\widehat{u}$ to denote the Fourier transform of a function u:

$$\widehat{u}(\xi) := \int_{\mathbb{R}^3} e^{i\xi\cdot\mathbf{r}} u(\mathbf{r}) \, d\mathbf{r} . \tag{16.26}$$

Therefore the basic equation is

$$\nabla \cdot \left(\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi)\widehat{\mathbf{e}}(\xi) \, d\xi \right) = 4\pi\rho(\mathbf{r}). \tag{16.27}$$

We can write $\mathbf{e} = \nabla\phi$ using Maxwell's equations. Therefore $\widehat{\mathbf{e}}(\xi) = i\xi\widehat{\phi}(\xi)$. Inserting this into (16.27) yields

$$\nabla \cdot \left(\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi) i\xi\widehat{\phi}(\xi) \, d\xi \right) = 4\pi\rho(\mathbf{r}). \tag{16.28}$$

Taking the Fourier transform (16.26) of (16.28) provides the simple relation

$$\widehat{\phi}(\xi) = \frac{4\pi\widehat{\rho}(\xi)}{|\xi|^2 \kappa(\xi)} \tag{16.29}$$

which can be used to compute ϕ (and thus $\mathbf{e}$) from ρ.

The expression (16.27) can be simplified in certain limits. We have

$$\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi)\widehat{\mathbf{e}}(\xi) \, d\xi \approx \epsilon_j \mathbf{e}(\mathbf{r}) \tag{16.30}$$

where $j = 1$ when $\mathbf{e}$ is very smooth and $j = 0$ when $\mathbf{e}$ consists of only high frequencies. However, for general fields $\mathbf{e}$ it is not possible to approximate the Fourier integral in this way. Thus we cannot think of (16.27) as a partial differential equation, except approximately in special cases.

When a non-dielectric material (or one with a much smaller dielectric effect) is immersed in a dielectric (e.g., a protein in water), it might be plausible to approximate (16.30) with $j = 1$ in the dielectric, switching to $j = 0$ at the interface of the non-dielectric material (which introduces high frequencies due to the abrupt change in material). This leads to the standard Poisson equation with a spatially varying permittivity ϵ_j that jumps from $j = 1$ in the dielectric to $j = 0$ in the non-dielectric material; this is often used to model macro-molecular systems in solvent [278].

However, it is not clear what to do when very small non-dielectric objects, such as nanotubes [365], are introduced into a dielectric. The scale of a nanotube is so small that there would be almost no ϵ_0 region in such models, so that any predictions of electrostatics would be essentially the same as if there were pure dielectric. It is possible to introduce a spatially varying permittivity that changes more smoothly between the two extremes [365], but this does not capture accurately the behavior of the wave-number dependence.

The characteristic scale λ represents a correlation length relating the way changes in the dielectric influence each other spatially. When the dielectric molecules are constrained, for example, at a material boundary, the characteristic scale λ increases. This is because the dielectric molecules become ordered near a wall, and thus changes propagate further than in bulk. It is also clear that these changes may be anisotropic, with changes parallel to the wall more affected than perpendicular to the wall. Such changes near an interface could cause λ to increase effectively to infinity at the surface of the bounding material. Thus it might be reasonable to view the kernel κ in this case as continuous across material boundaries.

The equation (16.27) involves a psuedo-differential operator [109]. Due to the special form of $\kappa(\nu)$, it is possible to write (16.27) as a fourth-order elliptic partial differential operator for the potential ϕ:

$$\nabla \cdot \left(\left(\epsilon_1 - \epsilon_0 \lambda^2 \Delta \right) \nabla \phi \right) = \left(1 - \lambda^2 \Delta \right) \rho \tag{16.31}$$

provided that λ is constant [415, 190, 189]. However, if λ is a function of $\mathbf{r}$ this is no longer valid. Also, the limit $\lambda \to \infty$ is harder to interpret in this setting.

16.1.5   Nonlocal relationship between p and e

The model (16.24) has been used by different people [74, 121, 236, 255] to account for the frequency dependence of the (zero temporal frequency) dielectric relationship. It is often expressed as a non-local dependence of the polarization on the electric field and written in the form

$$\mathbf{p}(\mathbf{r}) = \int K(\mathbf{r}, \mathbf{r}') \mathbf{e}(\mathbf{r}') \, d\mathbf{r}' \tag{16.32}$$

where the averaging kernel K satisfies

$$K(\mathbf{r}, \mathbf{r}') = K(\mathbf{r} - \mathbf{r}') = \int e^{i\mathbf{k} \cdot (\mathbf{r} - \mathbf{r}')} \kappa(\mathbf{k}) \, d\mathbf{k} \tag{16.33}$$

with the expression κ representing the Debye-like frequency dependence (16.23). Taking Fourier transforms, we see that the "non-local" model (16.23) is the same as (16.25). However, it is not possible to represent the mollifier K as an ordinary function. Clearly,

$$K = \delta + \frac{\epsilon_1 - \epsilon_0}{\epsilon_0} \widetilde{K}, \tag{16.34}$$

where

$$\widetilde{K}(\mathbf{r}) = \int e^{i\mathbf{k}\cdot\mathbf{r}} \tilde{\kappa}(\mathbf{k}) \, d\mathbf{k} \tag{16.35}$$

and $\tilde{\kappa}$ is defined by

$$\tilde{\kappa}(\mathbf{k}) = \frac{1}{1 + |\mathbf{k}|^2 \lambda^2}. \tag{16.36}$$

We easily identify $\tilde{\kappa}$ as the Fourier transform of the fundamental solution of the Laplace operator $1 - \lambda^2 \Delta$, so that

$$\widetilde{K}(\mathbf{r}) = \frac{\lambda e^{-|\mathbf{r}|/\lambda}}{4\pi|\mathbf{r}|}. \tag{16.37}$$

Although this expression appears singular, we realize it is less singular than the Dirac δ-function, which simply evaluates a function at a point instead of averaging. The exponential decay insures that the averaging is fairly local in nature. The kernel for the non-local expression for the polarization can be written formally (in the sense of distributions [363]) as

$$K(\mathbf{r}) = \delta + \frac{\lambda(\epsilon_1 - \epsilon_0) e^{-|\mathbf{r}|/\lambda}}{4\pi\epsilon_0|\mathbf{r}|}. \tag{16.38}$$

More precisely, we have

$$\epsilon_0 \mathbf{p}(\mathbf{r}) = \epsilon_0 \mathbf{e}(\mathbf{r}) + \frac{\lambda(\epsilon_1 - \epsilon_0)}{4\pi} \int \frac{e^{-|\mathbf{r}-\mathbf{r}'|/\lambda}}{|\mathbf{r} - \mathbf{r}'|} \mathbf{e}(\mathbf{r}') \, d\mathbf{r}'. \tag{16.39}$$

16.2 Computing the energy

Estimating the electrostatic energy of a protein-water-ion system is not easy. At a quantum-mechanical level of description, the energy is an eigenvalue that emerges as a global quantity of the system. In molecular mechanics, the partial charges of individual atom sites is estimated, accounting for polarization effects (most individual atoms would have net charge zero in their native state). In this setting, the electrostatic (enthalpic) energy is given by

$$\sum_{i \in I_\rho} \sum_{j \neq i \in I_\rho} q_i q_j |\mathbf{r}_j - \mathbf{r}_i|^{-1}, \tag{16.40}$$

where the q_i's denote the partial charges assigned at locations $\mathbf{r}_i$. Note that care is taken in the double sum to insure that $j \neq i$ to avoid what is known as **self energy**. There is an energetic contribution for an individual atom at a quantum-mechanical level of description, but putting $i - j$

in the expression in (16.40) would give an infinite contribution. Likewise, the other terms for $j \neq i$ represent only approximations to the quantum-mechanics assessment of the energy.

To approximate the quantum-level description better, the electron distributions could be represented as continuous distributions. Ultimately, even the positions of the atom nuclei are only known in a probabilistic sense. So we can imagine a fully continuous charge description of this type, but the scales of representation would be vastly different. The nuclei are more localized than the electrons by several orders of magnitude. Thus it is problematic to use this approach.

Going in the other direction, the pattern of charges in the molecular mechanics description are often modeled as being continuous. This is useful computationally, but it has some pitfalls as we now discuss.

When either the charge distribution ρ or the potential ϕ is continuous, the energy associated with the system modeled by (16.14) is given by the expression

$$\int \phi \rho \, dx = \sum_{i \in I_\rho} q_i \langle \delta_{\mathbf{r}_i}, \phi \rangle , \tag{16.41}$$

where we have used distribution theory (the angle brackets) to allow interpretation of the point charges rigorously in the case of point masses (delta functions):

$$\langle \delta_{\mathbf{r}}, v \rangle = v(\mathbf{r}) \tag{16.42}$$

for any test function v. Unfortunately, we know that, with any of the models considered (nonlocal or nonlinear), the potential ϕ will have a singularity like $|\mathbf{r} - \mathbf{r}_i|^{-1}$ for $\mathbf{r}$ near $\mathbf{r}_i$ when ρ involves true delta functions. Thus the expression (16.41) includes a sum of infinities.

16.2.1   Analytical approximations

We can imagine approximating the singularities by a number of techniques. For simplicity, let us assume we are in a vacuum so that the exact potential ϕ is given by (16.2). For example, we might replace the δ functions by an intergal over a the surface of the sphere of radius ϵ, suitably scaled That is, we define

$$\langle \delta_{\epsilon, \mathbf{r}}, v \rangle = \frac{1}{4\pi\epsilon^2} \oint_{S_{\epsilon, \mathbf{r}}} v \, ds, \tag{16.43}$$

where $S_{\epsilon, \mathbf{r}}$ denotes the sphere of radius ϵ around $\mathbf{r}$. Although this is still a singular distribution, in the sense that it cannot be represented by an integrable function on $\mathbb{R}^3$, the corresponding potential is now continuous. The potential is the same away from the singularity, but it is constant (equal to $1/\epsilon$) inside a ball of radius ϵ:

$$\phi_{\epsilon, \mathbf{r}}(\mathbf{x}) = \begin{cases} |\mathbf{x} - \mathbf{r}|^{-1} & |\mathbf{x} - \mathbf{r}| \geq \epsilon \\ \epsilon^{-1} & |\mathbf{x} - \mathbf{r}| \leq \epsilon. \end{cases} \tag{16.44}$$

Now at least the pairing

$$\langle \delta_{\epsilon, \mathbf{r}}, \phi_{\epsilon, \mathbf{r}} \rangle = \epsilon^{-1} \tag{16.45}$$

is finite. This is often called the Born (electrostatic) approximation [186].

Making the systematic change of $\delta_{\epsilon,\mathbf{r}}$ for δ_ϵ in (16.41) changes the singular terms to

$$\sum_{i \in I_\rho} q_i \langle \delta_{\epsilon,\mathbf{r}_i}, \phi_\epsilon \rangle = \sum_{i \in I_\rho} \sum_{j \neq i \in I_\rho} q_i^2 \epsilon^{-1} + \sum_{i \in I_\rho} \sum_{j \neq i \in I_\rho} \frac{q_i q_j}{4\pi\epsilon^2} \oint_{S_{\epsilon,\mathbf{r}_i}} |\mathbf{r}_j - \mathbf{r}|^{-1}\, ds, \qquad (16.46)$$

where we recall $S_{\epsilon,\mathbf{r}_i}$ denotes the sphere of radius ϵ around $\mathbf{r}_i$, assuming that the minimum distance between $\mathbf{r}_i$ and $\mathbf{r}_j$ is bigger than ϵ. The latter sum converges when $\epsilon \to 0$:

$$\lim_{\epsilon \to 0} \sum_{i \in I_\rho} \sum_{j \neq i \in I_\rho} \frac{q_i q_j}{4\pi\epsilon^2} \oint_{S_\epsilon(\mathbf{r}_i)} |\mathbf{r}_j - \mathbf{r}|^{-1}\, ds = \sum_{i \in I_\rho} \sum_{j \neq i \in I_\rho} q_i q_j |\mathbf{r}_j - \mathbf{r}_i|^{-1}, \qquad (16.47)$$

and this limiting expression is the correct electrostatic energy in this case. The other term

$$\epsilon^{-1} \sum_{i \in I_\rho} q_i^2 \qquad (16.48)$$

is clearly a spurious self-energy term, and it diverges as $\epsilon \to 0$. Note that there is no chance of cancellation due to having charges of opposite signs. All the contributions in (16.48) are of one sign.

On the positive side, we see that the expression

$$\int \phi_\epsilon \rho_\epsilon\, dx - \epsilon^{-1} \sum_{i \in I_\rho} q_i^2 \qquad (16.49)$$

tends to the right limit as $\epsilon \to 0$.

16.2.2   Smoothing the delta function

It might be a concern that the approximation of the Dirac delta function in Section 16.2.1 is still too singular. Another approach to interpret the energy is to approximate the Dirac delta functions by smooth functions. This has some physical basis, since the electrons in an atom are distributed over a region on the order of an Ångstrom in size, a scale we would actually resolve in these models. On the other hand, as discussed earlier, the nuclear charge is concentrated on a much smaller scale. Thus let us imagine that we replace δ systematically by δ_ϵ for some fixed $\epsilon > 0$. To be precise, we take

$$\delta_\epsilon(\mathbf{r}) = \epsilon^{-3} \chi(\epsilon^{-1}\mathbf{r}) \qquad (16.50)$$

for all points $\mathbf{r}$, where χ is a smooth function that vanishes for $|\mathbf{r}| \geq 1$ and satisfies $\int \chi(\mathbf{r})\, d\mathbf{r} = 1$. Then for sufficiently smooth functions ψ we have

$$\lim_{\epsilon \to 0} \int \psi(\mathbf{r}) \delta_\epsilon(\mathbf{r} - \mathbf{r}_i)\, d\mathbf{r} = \psi(\mathbf{r}_i). \qquad (16.51)$$

Define ϕ_ϵ to be the solution of (16.14) with $\rho = \rho_\epsilon$ defined by

$$\rho_\epsilon(\mathbf{r}) = \sum_{i \in I_\rho} q_i \delta_\epsilon(\mathbf{r} - \mathbf{r}_i). \tag{16.52}$$

Thus the approximate energy expression

$$\int \phi_\epsilon \rho_\epsilon \, dx = \sum_{i \in I_\rho} q_i \int \phi_\epsilon(\mathbf{r}) \delta_\epsilon(\mathbf{r} - \mathbf{r}_i) \, d\mathbf{r} = \sum_{i \in I_\rho} q_i \int \phi_\epsilon(\mathbf{r} + \mathbf{r}_i) \delta_\epsilon(\mathbf{r}) \, d\mathbf{r} \tag{16.53}$$

appears to be a reasonable replacement for (16.41).

For any constant M,

$$\int \phi_\epsilon \rho_\epsilon \, dx - M \sum_{i \in I_\rho} q_i^2 = \sum_{i \in I_\rho} q_i \int \left(\phi_\epsilon(\mathbf{r} + \mathbf{r}_i) - M q_i \right) \delta_\epsilon(\mathbf{r}) \, d\mathbf{r}. \tag{16.54}$$

It is plausible that

$$\phi_\epsilon(\mathbf{r} + \mathbf{r}_i) \approx \frac{c q_i}{\epsilon} \quad \forall |\mathbf{r}| \leq \epsilon \tag{16.55}$$

for some constant c that would only depend on the choice of ψ. Choosing $M = c/\epsilon$, we conclude that

$$\int \phi_\epsilon \rho_\epsilon \, dx - \frac{c}{\epsilon} \sum_{i \in I_\rho} q_i^2 \tag{16.56}$$

should remain bounded as $\epsilon \to 0$.

16.2.3   Numerical computations

When equations like (16.14) are solved numerically, we obtain an approximation ϕ^h, where $h > 0$ is the spatial discretization parameter. In the case of finite element approximation [51], the expression

$$\int \phi^h \rho \, dx = \sum_{i \in I_\rho} q_i \left\langle \delta(\mathbf{r} - \mathbf{r}_i), \phi^h \right\rangle = \sum_{i \in I_\rho} q_i \, \phi^h(\mathbf{r}_i) \tag{16.57}$$

is perfectly well defined [366]. However, as h tends to zero, the values $\phi^h(\mathbf{r}_i)$ will diverge [367]. In particular, we expect [367] that

$$\phi^h(\mathbf{r}) \approx \frac{c_i q_i}{h_i}, \quad \forall |\mathbf{r} - \mathbf{r}_i| \leq h_i, \tag{16.58}$$

where h_i is the local mesh size near $\mathbf{r}_i$ and c_i is a constant depending on the local mesh behavior and where the singularity falls within a mesh element. Thus

$$\int \phi^h \rho \, dx = \sum_{i \in I_\rho} q_i \, \phi^h(\mathbf{r}_i) \approx \sum_{i \in I_\rho} \frac{c_i q_i^2}{h_i} \tag{16.59}$$

which we recognize as some sort of self-energy term. We can deal with this self-energy term in a rigorous way and show how to remove it as follows.

The system (16.14) is linear, so we can write

$$\phi^h = \sum_{i \in I_\rho} q_i \, \phi_i^h, \tag{16.60}$$

where ϕ_i^h is the discrete (finite difference or finite element) solution of (16.14) with $\rho = \delta(\mathbf{r} - \mathbf{r}_i)$. The self-energy term that we want to eliminate is precisely

$$\sum_{i \in I_\rho} \langle q_i \, \phi_i^h, q_i \delta(\mathbf{r} - \mathbf{r}_i) \rangle = \sum_{i \in I_\rho} q_i^2 \phi_i^h(\mathbf{r}_i). \tag{16.61}$$

Thus the desired energy is (16.57) minus (16.61):

$$\begin{aligned}
E^h = \sum_{i \in I_\rho} q_i \, \phi^h(\mathbf{r}_i) - \sum_{i \in I_\rho} q_i^2 \phi_i^h(\mathbf{r}_i) &= \sum_{i \in I_\rho} \sum_{j \in I_\rho} q_i \, q_j \phi_j^h(\mathbf{r}_i) - \sum_{i \in I_\rho} q_i^2 \phi_i^h(\mathbf{r}_i) \\
&= \sum_{i \in I_\rho} \sum_{i \neq j \in I_\rho} q_i \, q_j \phi_j^h(\mathbf{r}_i).
\end{aligned} \tag{16.62}$$

Computation of (16.62) has exactly the right terms, avoiding the self-interaction terms. However, (16.62) appears to require the solution of lots of separate problems, so in general this would be very expensive. But, in one common case, computing energy differences, it simplifies.

16.2.4   Computing energy differences

Suppose that we want to compare the energy difference $E_1 - E_2$ for two different systems of charges indexed by I_1 and I_2, respectively. That is, we have fixed charge distributions ρ_i given by

$$\rho_k = \sum_{i \in I_k} q_i \delta(\mathbf{r} - \mathbf{r}_i) \tag{16.63}$$

for $k = 1, 2$. Let us assume for simplicity that $I^1 = I^2 \cup \Delta I$ for some set of indices ΔI. In particular, computing pKa's (Section 4.5) often involves sets that differ only in one element.

Let ϕ^k denote the solution of the system (16.14) with $\rho = \rho_k$ for $k = 1, 2$. Here, we have temporarily dropped the mesh parameter h since there are too many indices. Then using the energy definition (16.62), we have

$$\begin{aligned}
\Delta E = E^1 - E^2 &= \sum_{i \in I_1} q_i \, \phi^1(\mathbf{r}_i) - \sum_{i \in I_1} q_i^2 \, \phi_i^h(\mathbf{r}_i) - \sum_{i \in I_2} q_i \, \phi^2(\mathbf{r}_i) + \sum_{i \in I_2} q_i^2 \, \phi_i^h(\mathbf{r}_i) \\
&= \sum_{i \in I_1} q_i \, \phi^1(\mathbf{r}_i) - \sum_{i \in I_2} q_i \, \phi^2(\mathbf{r}_i) - \sum_{i \in \Delta I} q_i^2 \, \phi_i^h(\mathbf{r}_i) \\
&= \sum_{i \in \Delta I} q_i \, \phi^1(\mathbf{r}_i) + \sum_{i \in I_2} q_i \left(\phi^1(\mathbf{r}_i) - \phi^2(\mathbf{r}_i) \right) - \sum_{i \in \Delta I} q_i^2 \, \phi_i^h(\mathbf{r}_i).
\end{aligned} \tag{16.64}$$

We can write

$$\Delta\rho = \rho_1 - \rho_2 = \sum_{i \in \Delta I} q_i\, \delta(\mathbf{r} - \mathbf{r}_i), \tag{16.65}$$

and by linearity of the system (16.14), we can think of $\Delta\phi = \phi^1 - \phi^2$ as the solution of (16.14) with $\rho = \Delta\rho$:

$$\Delta\phi = \phi^1 - \phi^2 = \sum_{j \in \Delta I} q_j\, \phi_j^h. \tag{16.66}$$

Thus the energy difference can be expressed as

$$\Delta E = \sum_{i \in \Delta I} q_i\, \phi^1(\mathbf{r}_i) + \sum_{i \in I_2} q_i \sum_{j \in \Delta I} q_j\, \phi_j^h(\mathbf{r}_i) - \sum_{i \in \Delta I} q_i^2\, \phi_i^h(\mathbf{r}_i). \tag{16.67}$$

Therefore we are only required to compute ϕ^1 and ϕ_i^h for $i \in \Delta I$.

16.2.5 Thermodynamic smoothing

Another physical interpretation of smoothing the delta function is available. One drawback of the ϵ-smoothing in Section 16.2.2 is that it is motivated mainly by the distribution of the electrons. The proton nucleus would still look like a point charge in this picture. However, it is useful to ask what this model is supposed to be representing. Unless we are at absolute zero, all of the atoms (meaning the nuclei as well) are moving due to thermal agitation. Thus the uncertainty in position of the nuclei in thermal motion is far greater than it is due only to quantum effects.

PDB files list **temperature factors** for each atom, also known as **B-factors**, which indicate the observed thermal motion (or maybe just the blurriness) of each atom. These smoothed distributions of charge could potentially be used to provide a more physically appropriate smoothing of the charge delta functions. The dielectric model is after all an ensemble average over time, and if the temperature factors accurately represent the mobility of atoms, then this could be factored into the model in a potentially useful way.

The B-factors are different for each atom, and thus we would have an approximation that varied from location to location. However, now the charge would be more legitimately distributed (and continuous), so the need for correction by elimination of self-charge terms would be eliminated. On the other hand, the assessment of the electrostatic contribution to enthalpy would now be potentially temperature dependent.

16.3 Nonlinear dielectric models

Large electric gradients near the surface could cause saturation of the dielectric effect, leading to nonlinear models of the dielectric coefficient [24, 25, 163, 211, 232, 237, 354, 355, 391].

Nonlinear models arise because the polarization field $\nabla\phi_\gamma$ in the relationship (16.8) cannot continue to increase indefinitely for $\nabla\phi$ arbitrarily large. The field $\nabla\phi_\gamma$ is due to the fortuitous orientation of the solvent charge groups (e.g., water molecules), but once perfect alignment is achieved, no more improvement can result. (There is an additional effect due to the polarization

of the electron distribution of the individual solvent charge groups, but this also will saturate with sufficiently large field strength.) We can write this saturation property mathematically by taking a limit in (16.8), viz.,

$$\lim_{|\nabla\phi|\to\infty} (1 - \varepsilon)\nabla\phi = \lim_{|\nabla\phi|\to\infty} \nabla\phi_\gamma = C, \tag{16.68}$$

for some constant C depending only on the direction of $\nabla\phi$. Let us write ε as a function of $\nabla\phi$ by introducing a vector $\xi = \nabla\phi$. Then (16.68) becomes

$$\lim_{|\xi|\to\infty} (1 - \varepsilon(\xi))\xi = C. \tag{16.69}$$

This constant can potentially depend on the direction of approach to infinity for a simple lattice, but in general it will be isotropic. The value of C can be estimated by considering the case of perfect alignment of water molecules. The dipole of water is 1.85 Debye, and the (maximum) density of water corresponds to about one water molecule per 30 Ångstroms cubed (Å^3), corresponding to a box of side 3.1 Ångstroms. A more intuitive measure of dipole strength is q_e–Å, where q_e is the charge of an electron. Thus the dipole of water is 0.386 q_e–Å. Thus the maximum strength of the water dipole in bulk is about $0.04 q_e$–Å^{-2}.

One simple model that satisfies (16.69) is

$$\varepsilon(\xi) = \varepsilon_0 + \frac{\varepsilon_1}{1 + \lambda|\xi|} \tag{16.70}$$

for some constants ε_0, ε_1, and λ. This model (16.70) is easy to analyze [364], but we do not intend to suggest it as a serious model. Rather it should be viewed as a prototype nonlinear dielectric model, one that can be easily analyzed and implemented in existing codes for the purposes of software verification.

The Langevin-Debye model [94, 354, 355] defines

$$\varepsilon(\xi) = c_0 + c_1 \frac{L(c_2|\xi|)}{|\xi|}, \tag{16.71}$$

where c_0, c_1, and c_2 are physical constants and L is the Langevin function

$$L(x) = \coth(x) - \frac{1}{x} = \frac{e^x + e^{-x}}{e^x - e^{-x}} - \frac{1}{x} \approx \tfrac{1}{3}x - \tfrac{1}{45}x^3 + \cdots \tag{16.72}$$

Other nonlinear coefficient models are reviewed in [211].

Both the nonlocal and nonlinear models of the dielectric response have the effect of representing frequency dependence of the dielectric effect. The gradient $|\nabla\phi(x)|$ provides a proxy for frequency content, although it will not reflect accurately high-frequency, low-power electric fields. Thus a combination of nonlocal and nonlinear models for the dielectric response might need to be used in general to capture the full effect.

16.4   Exercises

Exercise 16.1 *Plot the Langevin function (16.72) on the interval $[0, 30]$. Compare the expression (16.71) with the function (16.70). More precisely, compare $3L(x)/x$ with $1/(1 + x)$ on various intervals.*

Exercise 16.2 *Consider the Langevin function L defined in (16.72), and define*

$$M(x) = \frac{1}{3\sqrt{1 + (x/3)^2}}. \tag{16.73}$$

Prove that the ratio $r(x) = (L(x)/x)/M(x)$ satisfies $0.93171 \leq r(x) \leq 1$ for $x \geq 0$.

Exercise 16.3 *Assume that the $\mathbf{r}_i$ and q_i for $i \in I_\gamma$ can be enumerated as $i = (j, k)$, where j is the index for the group and k is the index within each group, with $\mathbf{r}_i := \mathbf{r}_j - \mathbf{r}_{jk}$ and $q_i := q_{jk}$ where the j-th group of charges q_{jk} sums to zero for all j:*

$$\sum_k q_{jk} = 0. \tag{16.74}$$

Show that the expression for the charge density can be simplified as

$$\gamma(\mathbf{r}) = \sum_j \sum_k q_{jk} \delta(\mathbf{r} - \mathbf{r}_j - \mathbf{r}_{jk}) \tag{16.75}$$

Exercise 16.4 *The expression δ in (16.1) can be interpreted in several ways. As a first abstraction, we can take it to be the Dirac delta function, which provides a rigorous model of a point charge [363]. In [192], a mollification of the Dirac delta function is introduced, which makes it possible to reason classically about expressions involving δ. This is a very useful device, and it can also be given a physical interpretation. We can think of δ representing the actual charge cloud that would be seen at a quantum scale. With this interpretation, there is an assumption being made, namely, that the local charge distribution can be represented by a single function $\delta(\mathbf{r})$, independent of the charge q and independent of the atom in question. This is of course not exact, but it gives a physical interpretation to the mollifier used in [192]. A closer approximation might be obtained by letting q be fractional, with positions $\mathbf{r}_{jk}$ chosen to improve the representation [279].*

Exercise 16.5 *Let us suppose that the charge groups are homogeneous in the sense that*

$$\mathbf{r}_{jk} = \mathcal{R}(\theta_j)\rho_k \tag{16.76}$$

for fixed vectors ρ_k and for some angle $\theta_j \in S_2$ (where S_2 denotes the unit 2-sphere), and further that $q_{jk} = q_k$ independent of j. This would be the case for water, for example [279]. Then

$$\sum_k q_{jk} \delta(\mathbf{r} - \mathbf{r}_j - \mathbf{r}_{jk}) = \sum_k q_k \delta(\mathbf{r} - \mathbf{r}_j - \mathcal{R}(\theta_j)\rho_k)$$
$$= \mathcal{F}(\theta_j, \mathbf{r} - \mathbf{r}_j) \tag{16.77}$$

where $\mathcal{F}$ is defined by

$$\mathcal{F}(\theta, \mathbf{r}) = \sum_k q_k \delta(\mathbf{r} - \mathcal{R}(\theta)\rho_k). \tag{16.78}$$

Now suppose that δ is rotationally invariant. Then

$$\begin{aligned}
\mathcal{F}(\theta, \mathcal{R}(\theta)\mathbf{r}) &= \sum_k q_k \delta(\mathcal{R}(\theta)\mathbf{r} - \mathcal{R}(\theta)\rho_k) \\
&= \sum_k q_k \delta(\mathbf{r} - \rho_k) \\
&= \nabla \cdot \mathcal{W}(\mathbf{r}),
\end{aligned} \tag{16.79}$$

with $\mathcal{W}(\mathbf{r}) = \nabla\psi(\mathbf{r})$ where ψ solves a Poisson equation of the form

$$\Delta\psi = \sum_k q_k \delta(\mathbf{r} - \rho_k). \tag{16.80}$$

If δ is the Dirac δ-function, then $\mathcal{W}$ is a generalized multipole expression

$$\mathcal{W}(\mathbf{r}) = -\sum_k q_k \frac{\mathbf{r} - \rho_k}{|\mathbf{r} - \rho_k|^3}. \tag{16.81}$$

Exercise 16.6 *Then*

$$\begin{aligned}
\sum_k q_{jk}\delta(\mathbf{r} - \mathbf{r}_j - \mathbf{r}_{jk}) &= \mathcal{F}(\theta_j, \mathbf{r} - \mathbf{r}_j) \\
&= \nabla \cdot \mathcal{W}(\mathcal{R}(\theta_j)^t(\mathbf{r} - \mathbf{r}_j)).
\end{aligned} \tag{16.82}$$

Therefore we have an exact representation of the dielectric field γ defined in (16.75), viz.,

$$\gamma(\mathbf{r}) = \sum_j \nabla \cdot \mathcal{W}(\mathcal{R}(\theta_j)^t(\mathbf{r} - \mathbf{r}_j)). \tag{16.83}$$

Exercise 16.7 *Suppose that $\nabla \cdot (\epsilon \mathbf{e}) = \rho$ where ρ is a continuous function in a neighborhood of $\partial\Omega$ and ϵ is a piecewise constant function with values ϵ_0 inside Ω and ϵ_1 outside Ω. Let e_1 be the limit of $\mathbf{n} \cdot \mathbf{e}$ from the outside of Ω, and let e_0 be the limit of $\mathbf{n} \cdot \mathbf{e}$ from the inside of Ω, where $\mathbf{n}$ is the unit normal to $\partial\Omega$. Prove that $\epsilon_0 e_0 = \epsilon_1 e_1$ on $\partial\Omega$. Show that this does not depend on whether $\mathbf{n}$ is the inner or outer normal.*

Chapter 17

Units

It is helpful to pick the right set of units in order to reason easily about a physical subject. In different contexts, different units are appropriate. Small boat enthusiasts will recognize the need to determine whether depth on a chart is labeled in feet or fathoms. It is common in the United States coastal waters to label the depths in feet where mostly small boats will be expected to be found. But commercial vessels might prefer to think in fathoms (a fathom is six feet) since their depth requirements will be some number of fathoms (and thus a much larger number of feet). The phrase "mark twain" was used by riverboats for whom twelve feet of water provided safe passage.

A mistake about units could be disastrous. In boating, thinking that a depth given in feet is really fathoms could lead to a grounding of the boat. Perhaps it is reasonable to expect that such gross errors would never occur, but a commercial airline was once forced to land at a converted airfield in Gimli, Manitoba because the weight of the fuel was computed in pounds instead of kilograms. As a result, the plane lost all power, including most of its electrical power needed for running the plane, at a high altitude. The aircraft and occupants survived the incident, but the former became known as the Gimli Glider.

In astronomy, we may measure distances in light-years. But this is the wrong unit for our discussion. Just like the choice between fathoms for commercial vessels and feet for small pleasure boats, we need to find the right size for our mental models.

17.1 Basic units *vs.* derived units

We encounter units for many things: length, time, mass, charge, viscosity, energy, and so forth. There are only so many of these that are independent. Once we choose a set of units, others must be derived from them. In Table 17.1, we give a simple example of a particular choice of basic and

basic units	length, time, mass, charge, temperature
derived units	energy, viscosity, kinematic viscosity, permittivity, speed of light

Table 17.1: An example of basic *versus* derived units.

basic units	meter, second, kilogram, ...
derived units	joule (energy), newton (force), ...

Table 17.2: The SI system of basic and derived units.

derived units.

As a simple example, energy (E) is measured in units of mass (m) times velocity (v) squared, and velocity has units length (ℓ) over time (t):

$$E = mv^2 = m(\ell/t)^2. \tag{17.1}$$

However, there is no canonical definition of which units are basic and which are derived.

We might want certain units to have a prescribed value, and thus we take those units to be basic. For example, we might want the permittivity of free space to be one, as we discuss in Section 17.2.3.

17.1.1 SI units

One standard set of units is the SI system. Mass is measured in (kilo)grams, distance in meters, time in seconds. There are other basic units as well, but let us stop here as it allows us to define the standard units of energy and force.

The SI standard unit of energy is the joule. Energy has units mass times velocity squared, as we know from Einstein's famous relation. A joule is a newton-meter, the work related to applying the force of a newton for a distance of a meter. A newton is one kilogram-meter/second^2, so a joule as one kilogram-(meter/second)^2.

These quantities are familiar macroscopic measures. A kilogram is the weight of a good book, and a meter per second is understandable as a walking speed: 3.6 kilometers per hour. Thus a joule is the energy required to get a book up to walking speed. In many cases, the older unit **calorie** is used, which differs from the joule by a small factor: one calorie is 4.1868 joules.

17.2 Biochemical units

There are natural units associated with biochemical phenomena which relate more to the nanoscale. For example, the frequently used unit for energy is **kcal/mole**. This of course refers to one-thousand calories per mole of particles, or per 6.022×10^{23} particles, which is Avogadro's number. That is a big number, but we can squash it down with the right word: it is 0.6022 yotta-particles (**yotta** is a prefix which means 10^{24}, just as kilo means 10^3 or nano means 10^{-9}). The kcal is 4.1868 kilojoules, or 3.9683 Btu (British thermal units, a unit used in describing the power of both residential and commercial heating and cooling systems).

17.2.1 Molecular length and mass units

At the molecular scale, the typical units of mass (e.g., the gram) are much too large to be meaningful. Moreover, the units such as the meter and gram are based on macro-scale quantities. The Ångstrom is just a shorthand notation for 10^{-10} meters, and it is the commonly used distance in molecular discussions. A more appropriate length scale for the atomic scale [390] might be the Bohr radius $a_0 = 0.529189$Å, which is based on properties of the electron distribution for the hydrogen atom, but at least we see that it is on the same order as an Ångstrom. Similarly, a more natural mass unit would be based on an atomic mass, e.g., the **dalton** (or **Da**) which is essentially the mass of the hydrogen atom. More precisely, it is one-twelfth of the mass of carbon twelve. The dalton is almost identical to the previous standard known as the **atomic mass unit** (or **amu**). The dalton mass unit $= 1.66053 10^{-27}$ kilograms.

17.2.2 Molecular time units

A natural time scale for biochemistry is the femtosecond (10^{-15} second) range. This is the temporal scale to observe the dynamics of molecules above the quantum level. For example, time-stepping schemes for molecular dynamics simulation are often a few femtoseconds, although some systems (e.g., liquid argon) appear to be stable for timesteps up to 100 femtoseconds. The **svedberg** is a time unit equal to 100 femtoseconds (10^{-13} second).

This svedberg provides a time scale that resolves molecular motion, but does not over resolve it: it is a scale at which to see details evolving the way a mechanical system would evolve in our everyday experience. We perceive things happening in a fraction of a second and are aware of motions that take place over many seconds. Runners and other athletes are timed to hundredths of a second, so we can think of that as a timestep for our perception. Thus our typical perception of motion covers 10^4 or 10^5 of our perceptual timesteps. By this reckoning, there are about 2×10^{11} timesteps in a typical human lifetime. Biological events, such as protein folding, take up to 10^{11} svedbergs, and even more. Note that a typical human height is about 2×10^{10} Ångstroms.

There is a natural length scale associated with any temporal scale when electromagnetic waves will be of interest. Just like the light-year, it is natural to consider the distance light travels in the natural time unit here, the svedberg, about 2.9979×10^{-5} meters, or 30 micrometers. This may seem odd. You might have expected a spatial unit on the order of an atomic unit such as the Ångstrom, but this is 0.03 millimeters, a scale we can almost resolve with a magnifying glass. This means that light is still very fast at these molecular scales. We hesitate to give this length a name, but it is clearly a light-svedberg.

If we pick the svedberg as time unit and the Ångstrom as spatial unit, then the natural velocity scale is the Ångstrom per svedberg, which is equal to $10^{-10+13} = 1000$ meters per second, about three times the speed of sound in air at sea level. With the dalton as mass unit, the natural energy unit in these units is one dalton-(Ångstrom per svedberg)^2. One dalton-(Å/svedberg)$^2 = 1.66 \times 10^{-21}$ joules=0.239 kcal/mole. Thus the chosen units of mass, length and time lead to a nearly unit value for the commonly used unit of energy, kcal/mole.

basic units	Ångstrom, svedberg (10^{-13}s), dalton (mass)
derived units	(energy) dalton-(Å/svedberg)^2 =0.239 kcal/mole

Table 17.3: Some units relevant for biochemistry.

17.2.3 Charge units

The natural unit of charge for protein chemistry is the charge of the electron, q_e. When we look at macromolecules, we can resolve individual units and their charges. The coulomb is an aggregate charge constant defined so that $q_e = 1.602 \times 10^{-19}$ C. That is, C= $6.242 \times 10^{18}q_e$. The actual definition of a coulomb is the charge associated with an ampere flowing for a second. Thus a hundred amp-hour battery has 360,000 coulombs of charge, or about $2.25 \times 10^{24}q_e$, which corresponds to 3.7 moles of electrons.

Permittivity has units charge-squared per energy-length:

$$\text{permittivity} = \frac{\text{charge}^2}{\text{energy-length}} = \frac{\text{charge}^2\text{time}^2}{\text{mass-length}^3}. \tag{17.2}$$

Thus it is possible to have the permittivity, charge and energy be one in any units by varying the spatial (length) unit:

$$\text{length} = \frac{\text{charge}^2}{\text{energy-permittivity}}. \tag{17.3}$$

However, it would not be possible to specify length, permittivity, energy, and charge independently.

The permittivity of free space ε_0 is 8.8542×10^{-12}F m^{-1} (farads per meter). A farad is a coulomb squared per newton-meter. That is, we also have

$$\begin{aligned}
\varepsilon_0 &= 8.8542 \times 10^{-12}\text{C}^2\text{N}^{-1}\text{m}^{-2} \\
&= 3.450 \times 10^{26}q_e^2\text{N}^{-1}\text{m}^{-2} = 3.450 \times 10^{26}q_e^2\text{J}^{-1}\text{m}^{-1} \\
&= 1.444 \times 10^{27}q_e^2\text{cal}^{-1}\text{m}^{-1} = 1.444 \times 10^{30}q_e^2\text{kcal}^{-1}\text{m}^{-1} \\
&= 2.40 \times 10^6 q_e^2(\text{kcal/mole})^{-1}\text{m}^{-1} = 2.40 \times q_e^2(\text{kcal/mole})^{-1}\mu\text{m}^{-1} \\
&= 0.72 q_e^2(\text{kcal/mole})^{-1}\text{lfs}^{-1},
\end{aligned} \tag{17.4}$$

where 'lfs' stands for light-femtosecond, the distance travelled by light (in a vacuum) in a femtosecond. Thus we see that in the units in which energy is measured in kcal/mol, charge is measured in units of the charge of the electron, q_e, and length is the light-femtosecond (lfs), we find the permittivity of free space to be on the order of unity. It is noteworthy that Debye [94] used units so that $\varepsilon_0 = 1$, together with energy measured in kcal/mol and charge measured in units of q_e. This means that the implied spatial unit is 1.39 lfs, or about 417 nanometers, or just under half a micron, a length related to the Debye screening length in water [187]. If this is the chosen spatial unit, then $\varepsilon_0 = 1$ in these units. For reference, very large viruses [440] are between one and two tenths of a micron in diameter.

basic units	kcal/mole (energy), q_e (charge), ε (permittivity)
derived units	(length) 417 nanometers

Table 17.4: The units implied by Debye's assumptions.

basic units	ε_0 (permittivity), a_0 (length), m_e (mass), 10^{-3} svedberg (time)
derived units	hartree=$\hbar^2$ (energy)

Table 17.5: Quantum chemistry units.

We can compute permittivity in a more common length scale as

$$\varepsilon_0 = 0.00024 q_e^2 (\text{kcal/mole})^{-1} \text{Å}^{-1}, \tag{17.5}$$

but we see it is now quite small.

17.2.4 Conversion constants

Boltzmann's constant, $k_B = 1.380 \times 10^{-23}$ joules per degree Kelvin, relates energy to temperature. This seems really small, so let us convert it to the "kcal/mole" energy unit. We get

$$
\begin{aligned}
k_B &= 1.380 \times 10^{-23} \text{J/K} \\
&= \frac{1.380 \times 6.022}{4.1868} \text{cal/mole-K} \\
&= 1.984 \, \text{cal/mole-K}
\end{aligned}
\tag{17.6}
$$

For example, at a temperature of T=303K, we have $k_B T = 0.601$ kcal/mole.

If temperature is in degrees Kelvin, velocities are measured in Ångstroms per picosecond (around 224 miles per hour), and masses in daltons, then $k_B \approx 0.831$.

Planck's constant,

$$h = 6.626068 \times 10^{-34} \, \text{m}^2\text{-kg/s} = 39.90165 \, \text{Ångstroms}^2\text{-dalton/picosecond}, \tag{17.7}$$

has units energy-time, which is a unit of **action**. The other Planck constant $\hbar = h/2\pi$ is then $\hbar = 6.35055$ Ångstroms^2-dalton/picosecond.

The ratio of Planck's constant to Boltzmann's constant has an interesting interpretation. It is $h/k_B = 4.80 \times 10^{-11}$ seconds per degree Kelvin, or 48 picoseconds per degree Kelvin.

17.3 Quantum chemistry units

For one particle of mass m moving in a force field having potential V, it can be written

$$\iota\hbar \frac{\partial \psi}{\partial t} = \mathcal{H}\psi \tag{17.8}$$

where $\hbar$ is Planck's constant divided by 2π, ι is the imaginary unit and

$$\mathcal{H}\phi(\mathbf{r}) := -\frac{\hbar^2}{2m}\nabla^2\phi(\mathbf{r}) + V(\mathbf{r})\phi(\mathbf{r}). \tag{17.9}$$

The Schrödinger equation (17.8) has three terms which must have the same units in order to be dimensionally correct. If we divide (17.8) by $\hbar$, then the diffusion term is multiplied by the constant $\hbar/2m$. Fortunately, $\hbar/m$ has units of length-square over time, as required. In the Schrödinger equation (17.8) we have implicitly assumed that the permittivity of free space $\varepsilon_0 = 1/4\pi$. We can do this, as noted above, but we need to choose the right spatial and energy units to make it all work out. Unfortunately, if the energy unit is kcal/mole, the natural scale for biochemistry, then the spatial unit is quite large, four orders of magnitude larger than the typcial scale of interest.

A more typical choice of spatial unit at the quantum scale [390] would be to use the Bohr radius $a_0 = 0.529189$Å. This scale only differs by a factor of about two from what we have been considering so far. But the natural unit of mass is the mass of the electron $m_e = 9.10938 \times 10^{-31}$ kg$= 5.48579 \times 10^{-4}$ dalton. In these units, Planck's constant is

$$h = 39.90165 \text{ Ångstroms}^2\text{-dalton/picosecond} = 137.45\, a_0^2\text{-}m_e/\text{femtosecond}. \tag{17.10}$$

If we also adopt the hartree[1] [372] E_h for the unit of energy, and we adopt the mass of the electron m_e as the unit of mass, then things are better. By definition, the Hartree E_h is

$$E_h = m_e c^2 \alpha^2 = 4.356 \times 10^{-18}\text{joules} = 1.040 \times 10^{-21}\text{kcal} = 626.5\,\text{kcal/mole}, \tag{17.11}$$

(cf. Table 3.2) where $\alpha \approx .007$ is a dimensionless number known as the **fine structure constant**, cf. Exercise 17.7. Moreover, we also have the coefficient of the potential in (17.8) equal to E_h; that is, $e^2/(4\pi\varepsilon_0 a_0) = E_h$.

The time-derivative term in (17.8) is multiplied by $\hbar$, which fortunately has units of energy times time. Planck's constant $h = 6.626068 \times 10^{-34}$ joule-seconds $= 1.521 \times 10^{-16}$ Hartree-seconds $= 0.1521$ Hartree-femtoseconds. Dividing by 2π, we find that Planck's constant $\hbar = 0.02421$ Hartree-femtoseconds. That is, if we take the time unit to be femtoseconds, then the coefficient of the time derivative term is $= 0.02421$, or about one over forty. This is a small term. It implies that changes can happen on the scale of a few tens of attoseconds, whereas on the scale of a few femtoseconds (the typical time step of molecular dynamics simulations), the time-derivative term in (17.8) can plausibly be ignored, or rather time-averaged. To cast this in terms of the units suggested for biochemistry, the natural timescale for quantum chemistry is about 10^{-3} smaller, about a milli-svedberg.

In thinking of the Schrödinger equation in classical terms as describing the probability of an electron's position as it flies around the nucleus, it is interesting to think about the time scale for such a motion. At the speed of light, it takes an attosecond to go 3 Ångstroms. The time-scale of the Schrödinger equation is 24 attoseconds, and in this time anything moving at the speed of light would go 72 Ångstroms. If the Schrödinger equation represents the average behavior of electrons moving around the nucleus at anything approaching the speed of light, then they can make many circuits in this basic time unit of the Schrödinger equation. So it is plausible that it represents such an average of dynamic behavior.

---

[1]Douglas Hartree (1897-1958) pioneered approximation methods for quantum chemistry calculations.

17.4   Laboratory units

In a laboratory, it would be confusing to use units appropriate at the molecular scale. A typical mass unit would be a milligram (mg), and a typical volume unit would be a milliliter (mL). Inside cells, protein concentrations can exceed 100 mg per mL.

The terms 'millimolar' and 'micromolar' are used to express the fractional concentration of one substance in another, e.g., water. A mole of water has 6.022×10^{23} particles, which is Avogadro's number, and this many atoms of oxygen, and twice this number of hydrogens, weighs 18.0153 grams. At 4 degress Centigrade, where water has its maximum density, one gram of water occupies one cubic centimeter, or one milliliter. Thus a mole of water occupies 0.0180153 liters (at 4° C), so a liter of water has 55.508 moles of water, near 4 degress C. Thus we say that the molar concentration of water is 55.508 moles per liter.

The molar concentration of a mixture is more complicated. By definition, a solute (e.g., salt) has a molar concentration of xM (moles per liter) if there are x moles of the solute in a liter of solute and solvent (e.g., water). This mixture is easy to create: you prepare x moles of the solute and add solvent until you fill a liter container. You figure out what x moles weighs based on its atomic weight. But it is not known a priori how much solvent you will need to add, since the way they mix is not predictable in advance.

A one-molar solution of salt (Na-Cl) in water would require one mole of salt, or 58.4425 grams, based on the combined atomic weights of sodium and chlorine. Correspondingly, a 200mM solution would require only 11.6885 grams. At 4° C, a liter of water weights about a kilogram. So it is reasonable to assume that a 200mM solution of Na-CL and water would be mostly water. The number of atoms in a 200mM solution is 0.2 times Avagodro's number; the number of atoms of water in a liter is 55.508 times Avagodro's number. If we ignore the number of waters that would flow over the one liter container when the salt is added, the ratio of water molecules to salt molecules would be 277.54. For a one molar solution (1M), the number would be about five times smaller, nearer to 55 waters per Na-Cl pair. Accounting for the displacement of water by salt, these numbers would be slightly smaller.

17.5   Mathematical units

There is a natural set of units that might be called mathematical units. They are based on the observation that many named constants are really just conversion factors. For example, Boltzmann's constant really just converts temperature to energy. Thus with the right temperature scale, Boltzmann's constant is one (cf. Exercise 17.4). Similarly, Planck's constant has units energy times time, and it will be one with the right relationship between energy and time. This places a constraint on the relationship between mass, length, and time. A natural mass unit is the dalton, since it is roughly the mass of the smallest atom. With the dalton as the mass unit, the largest masses in the Schrödinger equation are of order one, although the smallest (i.e., the electrons) have a tiny mass in this unit. It is natural to take the speed of light to be one, so this sets a relationship between length and time.

If we divide Planck's constant by the speed of light we get $\hbar/c = 0.212 \times 10^{-15}$ dalton-meters.

basic units	speed of light, Planck's constant, Boltzmann's constant
derived units	time unit=0.7 yoctoseconds, length unit=0.2 femtometers

Table 17.6: Some units to simplify mathematical equations.

If we want $\hbar = 1$ and $c = 1$ [323], then we need to have the length unit to be 0.212×10^{-15} meters= 0.212 femtometers. The diameter of a proton is approximately one femtometer.

If we divide Planck's constant by the speed of light squared we get $\hbar/c^2 = 0.7066 \times 10^{-24}$ dalton-second. If we want $\hbar = 1$ and $c = 1$, then we need to have the time unit to be 0.7066×10^{-24} seconds=0.7066 yoctoseconds. If these independent calculations are correct, we would find that the speed of light is about 0.3 femtometers per yoctosecond. A femtometer per yoctosecond is 10^9 meters per second, so we have agreement.

To summarize, if we take length to be measured in multiples of L=0.212 femtometers, time to be measured in multiples of t=0.7066 yoctoseconds, and mass in daltons, then $c = \hbar = 1$. See Exercise 17.5 for the similar case where the unit of mass is the mass of the electron. As noted above, a joule in these units is 6.7006×10^9 dalton-$(L/t)^2$. Similarly, in these units $k_B = 9.2468 \times 10^{-14} K^{-1}$.

17.6 Evolutionary units

There are also other time scales of interest in biology, and geology. The **molecular clock** refers to the time it takes for a single point mutation to occur in DNA. This is measured by comparing divergent genomes of related species, for which the time of divergence is estimated from the fossil record. The unit of measure for sequence divergence is **percentage divergence**, which refers to the fraction of times each individual sequence entry is expected to have been modified in a given time unit. Typically, this is a very small number, so the time unit is often taken to be large. Thus a 2% sequence divergence per million years means that the probability of mutation of each individual sequence entry is only 0.02 in a million years. However, in a hundred million years, we would expect each entry to be modified twice.

Estimates for molecular clocks vary on the order of one percent divergence per 10^6 years, although turtle mitochondrial DNA mutation tends to be slower [22], possibly due to their slower internet connections. Fortunately, this is a slow scale from a human perspective. However, over geologic time, it is significant.

It is interesting to note that using typical estimates of the age of the earth [3], there has not been enough time for this type of mutation to cause a complete change to a typical chromosome. Time is usually measured in units of Mya (millions of years ago), or bya (billions of years ago); the latter unit is often replaced by the shorter Ga (giga-annum). The age of the earth is estimated to be at least 4.5 Ga [3].

We can put these two pieces of data together to estimate how many times a complete genome might have been modified. If the variation is occuring at a rate of one percent per 10^6 years, then in 10^8 years it could become completely modified. But given the estimated age of the earth, no genome would be expected to have been fully modified more than about forty five times via single

point mutations.

From the point of view of a dynamical system, a genome that has been modified 45 times might well have reached a stable equilibrium (if there is one). On the other hand, more recently evolved species would not yet be at a point where many cylces have taken place, so their genomes might be still far from an equilibrium point. Species that are older that 10^6 years would have experienced substantial modification (more than one base-pair in a hundred modified). But species much younger than 10^8 years might exhibit little limiting behvior, still undergoing substantial modification due to random mutations.

17.7   Other physical properties

We now consider some other physical properties that are measured in much the same way that basic units are. Some of them could themselves be treated as basic units, such as viscosity. Others deal with more complex issues, such as the pH scale which describes a mixture of materials.

17.7.1   The pH scale

At a pH of k, there are 10^{-k} moles of hydronium ions (and hydroxyl ions) per liter of water. A mole of water weighs 18.0153 grams. At 4 degress Centigrade, where water has its maximum density, one gram of water occupies one cubic centimeter, or one milliliter. Thus a mole of water occupies 0.0180153 liters (at 4° C), so a liter of water has 55.508 moles of water. Thus the ratio of hydronium ions to water molecules at a pH of k is roughly one hydronium ion per $5.5508 \times 10^{k+1}$ water molecules. Humans seem happiest at pH seven, which corresponds to a ratio of approximately one hydronium ion per half billion water molecules. However, the pH in cells can be much lower.

17.7.2   Polarity and polarization

The Debye is the standard unit for dipole moment, and is 3.338×10^{-30} coulomb-meters. A more useful unit would be a q_e-Ångstrom, where q_e is the charge of an electron, and this turns out to be about 4.8 Debye. Recall that a coulomb is $6.242 \times 10^{18} q_e$. Thus, a Debye is 0.2084 q_e-Ångstrom. The dipole moment of water ranges from about 1.9 Debye to 3.5 Debye depending on the enviroment [172, 85].

Polarization is the effect of an external field to change the strength of a dipole. An interesting feature is that the polarization coefficient has units of volume (i.e., length cubed). Thus there is a natural motif that can be used to illustrate the polarizability of an object: the volume of its representation. For example, if we are representing atoms as spheres, the volume of the sphere could be taken to be its polarization coefficient.

Polarization is a tensor, and it need not be isotropic. However, in many cases, a scalar approximation is appropriate. The polarizability of water is $\alpha \approx 1.2 \text{Å}^3$.

17.7.3   Water density

Water is a molecule with a complex shape, but it is possible to estimate the volume that an individual molecule occupies. A mole of water, which consists of 6.022141×10^{23} water molecules, weighs 18.0153 grams. Thus a gram of water has 0.33427×10^{23} water molecules. At 4 degress Centigrade, where water has its maximum density, one gram of water occupies one cubic centimeter, or 10^{24}Å^3. Therefore 0.33427×10^{23} water molecules occupy 10^{24}Å^3, at 4° C; one water molecule thus occupies $29.916 = 10/0.33427\text{Å}^3$. This corresponds to a cube of just over 3.1 Ångtroms on a side. It is interesting to compare this distance with a typical O-O distance (3.0Å, cf. Table 6.1).

The density of water is the number of water molecules per unit volume. If a single water molecule occupies 29.92Å^3, this corresponds to 0.0334 water molecules per Å^3, or a density of 0.0334Å^{-3}. Near protein surfaces, the density can be significantly higher, and there is a correlation between the average dipole orientation of nearby waters and the local density [188].

17.7.4   Fluid viscosity and diffusion

Fluids display an aggregate behavior known as **viscosity**. Fluid dynamicists [38] call the viscosity μ and physicists [247] call it η. The units of the coefficient of viscosity (often called **dynamic viscosity**) are mass per length-time. A standard unit of viscosity is the poise,[2] which is one gram per centimeter-second. One poise is 0.1 Pascal-second, where a Pascal is a unit of pressure or stress. One pascal is one newton per meter-squared, where we recall that a newton (one kilogram-meter/second-squared) is a measure of force.

The viscosity of water at 293 degrees Kelvin (20 degrees Centigrade) is about one centipoise, or about 0.001 Pascal-second. The viscosity of olive oil is about 80 times larger, so the ratio of viscosities of olive oil and water is roughly the ratio the dielectric of water and vacuum. The viscosity of air is 0.0018 centipoise, a factor of over five-hundred smaller.

17.7.5   Kinematic viscosity

Another scaling factor is significant in fluid flow, namely the fluid density. The ratio of viscosity (or dynamic viscosity) and density is called **kinematic viscosity**, usually labelled ν. This has units length-squared per time, since density has units of mass per length-cubed. Thus kinematic viscosity has the same units as a spatial diffusion constant. The stoke is one centimeter-squared per second. The kinematic viscosity of water is about one millimeter-squared per second, or one centistoke, whereas the kinematic viscosity of air is rougly two times *larger*. That is, air is more viscous than water! The viscosity of fluids varies significantly with temperature, but we have provided values at roughly the same temperature (293 K) for comparison.

Viscous drag is the effective force of viscosity in opposing motion. It provides a retarding force in the direction oppositve to the motion. The drag coefficient has the units of force divided by velocity, or mass per time unit.

---

[2]The unit of viscosity is named for Jean Louis Marie Poiseuille (1799–1869) who, together with Gotthilf Heinrich Ludwig Hagen (1797–1884) established the basic properties of viscous flow in simple geometries.

17.7.6   Diffusion

17.8   Exercises

Exercise 17.1 *Using the Bohr radius $a_0 = 0.5291772\mathring{A}$ as the basic unit of length, the svedberg as time unit, and the amu as mass unit, compute the unit of energy, one amu-(bohradius/svedberg)^2 in terms of the unit kcal/mole.*

Exercise 17.2 *Determine a basic unit of length L such that, with the svedberg as time unit, and the dalton as mass unit, then the unit of energy of one dalton-(L/svedberg)^2 is exactly one kcal/mole. Compare L with the van der Waals radius of different atoms (which are closest?).*

Exercise 17.3 *Determine a three-dimensional volume which can be used to tile space and fits a water molecule better than a cubic box. Use this volume to estimate the density of water.*

Exercise 17.4 *Suppose we take the dalton as the mass unit and that we choose space and time units so that the speed of light and Planck's constant are both one. What is the temperature scale that makes Boltzmann's constant equal to one?*

Exercise 17.5 *Suppose we take the mass of the electron as the mass unit, and that we want units so that the speed of light and Planck's constant are both one. What are the corresponding time and length scales?*

Exercise 17.6 *Suppose temperature is in degrees Kelvin and mass is in daltons. Determine a velocity scale such that $k_B = 1$.*

Exercise 17.7 *The fine structure constant is*

$$\alpha = \frac{q_e^2}{2\,h\,\varepsilon_0\,c},\tag{17.12}$$

where m_e and q_e are the mass and charge of the electron, respectively, ε_0 is the permittivity of free space, h is Planck's constant, and c is the speed of light in a vacuum. Prove that α is dimensionless. Determine other combinations of various physical constants that are also dimensionless.

Exercise 17.8 *The Rydberg constant R_∞ is*

$$R_\infty = \frac{m_e\,q_e^4}{8\,\varepsilon_0^2\,h^3\,c},\tag{17.13}$$

where m_e and q_e are the mass and charge of the electron, respectively, ε_0 is the permittivity of free space, h is Planck's constant, and c is the speed of light in a vacuum. Prove that the hartree $E_h = 2R$.

Exercise 17.9 *We have two equations for the Hartree E_h, namely, $E_h = m_e c^2 \alpha^2$ and $q_e^2/(4\pi\varepsilon_0 a_0) = E_h$. Show that these are compatible, that is, $4\pi\varepsilon_0 = q_e^2/a_0 m_e c^2 \alpha^2$.*

Chapter 18

Notes

The following notes are intended to complement the discussion in the indicated chapters.

Chapter 2

We are indebted to the noted photographer Ron Scott for the suggestion of the grain of sand in an oyster to explain epidiorthotric effects. Regarding epiorthotric forces in psychosocial contexts, we have in mind something like a sense of insecurity that often drives highly talented people. Groucho Marx expressed this via the conundrum that he would never join a club that would stoop so low as to have him as a member. A prolific inventor, Larkin Burneal Scott (1917–1991) expressed it in a jest as follows: "I don't have an inferiority complex; I *am* inferior."

Chapter 3

The origin of the octet rule is the Periodic Table, initiated in 1869 by Dmitri Ivanovitch Mendelev. It is useful to reflect on the history of the development of the understanding of different types of electrostatic forces and structural features of proteins. In Table 18.1, we present these together with approximate dates of emergence and representative citations. Dates should be considered only approximate and not a definitive statement about priority. Similarly, the references presented are only intended to give a sense of some central contributions.

Chapter 4

Chapter 5

Chapter 8

Chapter 7

Chapter 13

Chapter 9

Force/structure name	emergence date(s)	selected references
van der Waals	1893	
Keesom dipole	1912	
covalent bond	1916	
Debye induction	1920	
hydrogen bond	1920-33-44	[249]
London dispersion	1930	
alpha/beta structure	1936	
hydrophobic force	1954–present	[100]
cation-π interaction	1996	[105]
dehydron	2003	[140, 129, 142]

Table 18.1: A brief history of the development of understanding of the principle bonds that are significant for protein structure and interaction.

Chapter 19

Glossary

The following definitions are intended only as an informal description. Other sources should be consulted for definitive meanings. Our objective is simply to provide a rapid, if approximate, way for the reader to return to the main part of the book without needing to resort to another text or an on-line resource. The definitions often depend on other definitions; terms that are used in one definition and are defined separately are inidicated in a separate font *like this*. Many terms are defined fully in the text, and the index provides pointers to these explications. If a term is not found here, consult the index next.

An **aliphatic chain** is a group of atoms in which the electron distribution is localized around each atom, e.g., the non-cyclic chains of CH_n groups in Leucine and Isoleucine.

Allostery is derived from the Greek meaning "other shape." An **allosteric** effect is one that is induced at one part of a molecule by an effect (e.g., ligand binding) at another.

An **amino acid** is a molecule that forms the basis of the sidechain of a protein.

An **amide group** is the $N - H$ pair in a peptide bond as shown in Figure 13.1.

Something is **amphiphilic** if it is a combination of hydrophobic and hydrophilic parts.

An **antigen** is the entity to which antibodies bind. In general, these should be entities that are foreign such as bacteria.

An **aromatic chain** is a system in which the electron distribution is distributed around many atoms, e.g., the cyclic chains of CH_n groups in Phenylalanine.

A protein **antagonist** .....

A **beta-hairpin** (or β-hairpin) is ....

The **backbone** is the name for the sequence of C_α carbons in a protein chain, that is, the lower left and upper right C's in Figure 13.1.

A **bidentate** interaction involves two separate interactions between heavy atoms, such as can occur between terminal atoms on Arg and either Glu or Asp.

A **capsid** is the outer coat of a virus, typically a protein complex.

A **carbonyl group** is the $C = O$ pair in a peptide bond as shown in Figure 13.1.

C. elegans, or more completely **Caenorhabditis elegans**, is a worm.

A **chain** is an individual protein in a *protein complex*

A **coil** (a.k.a., loop) in a protein structure is a sequence of residues without alpha-helix or

beta-sheet structure. It can be quite long (several tens of residues) and is generally believed to be without a predetermined geometry, i.e., it can easily change its shape.

The **conformation** of a protein is the three-dimensional shape that it adopts. A change in conformation means that a new shape is adopted.

A **covalent** bond is an electrostatic bond in which the electrons of different atoms become intertwined and can no longer be identified as belonging to a distinct atom.

A **crystal** is a lattice of objects, such as proteins, that can form under certain conditions. The repeated (periodic) structure in particular allows them to be imaged using X-rays.

A **dimer** is an object made of two *monomers*, typically the same or very similar.

A protein **domain** is the basic unit of *tertiary strucutre*. A single protein can consist of a single domain or many domains. A protein *fold* is a synonym for domain.

DNA is the acronym for DioxyriboNucleicAcid.

E. coli, or more completely **Escherichia coli**, is a bacterium commonly found in food.

Electron density refers to the fact that electrons can not be located exactly, but rather a probabilistic description is used in quantum mechanics to describe where they spend a fraction of their time.

Endocytosis is a process of cell surface folding that ingests a substance.

The term **epidiorthotric** was introduced in Section 2.2.3 to refer to an effect that occurs as the result of a repair of a defect.

An **epitope** is a small region of a protein involved in a binding event, such as the part of an *antigen* where an antibody binds.

A protein **fold** is the basic unit of *tertiary strucutre*. A single protein can consist of a single fold or many folds. A protein *domain* is a synonym for fold.

Homo sapien (sometimes written **H. sapien**) is the formal biological name for a human being.

Something is **hydrophobic** if it repels water.

Something is **hydrophilic** if it attracts water.

Hydrophobic packing refers to the placement of carbonaceous groups in the vicinity.

A **hydroxyl** group is the OH group at the ends of the sidechains of serine, threonine and tyrosine.

A **ligand** is anything that binds to something.

A **loop** in a protein structure is an alternate designation for a *coil*.

A **missense mutation** is gene codon mutation that is not synonymous and not a *nonsense mutation*.

A **moity** is a portion of a whole, usually with a defined property or structure.

A **monodentate** interaction involves only separate interactions between heavy atoms, as opposed to *bidentate* interactions. Interactions between terminal atoms on Arg and either Glu or Asp can be either monodentate or bidentate.

A **monomer** is a single unit, e.g., a peptide, that can join with one or more other monomers of the same type to form a larger complex. The term can be used for something as small as a single molecule or as large as a protein.

A **motif** is a characteristic feature.

A **multimer** is something formed from small units, often called *monomers*, cf. *polymer*.

Mus musculus is the formal biological name for the common house mouse.

A **noncovalent** bond, or interaction, is an electrostatic interaction in which the electrons of the atoms in the bond remain sufficiently apart to remain identified with their atoms, even though they may be strongly correlated.

A **nonsense mutation** is gene codon mutation to a stop codon.

A **packing defect** is a defficiency in wrapping, that is, a lack of adequate carbonaceous groups in the vicinity.

A **partial charge** is a model to account for the fact that electrons may be unevenly distributed in a molecule.

A **peptide** is the basic unit of a protein.

Something is **polar** if it has a positive *partial charge* on one side and a negative *partial charge* on the other.

A **polymer** is something formed from small units, often called *monomers*, cf. *multimer*.

To **polymerize** is to form a larger system from small units, e.g., a chain such as a protein.

A **polypeptide** is the result of polymerizing peptides.

A graph has **power-law distribution** if the number of vertices with degree k is rougly $k^{-\gamma}$ for a fixed γ.

The **primary structure** of a protein is the sequence of its amino acids.

A **protein complex** is a collection of two or more proteins that are bound together.

The **quartenary structure** of a protein system is the three-dimensional arrangement of the different protein *chains* of the system.

An amino acid **residue** is part of the amino acid that remains when it is cleaved to form a sidechain on a protein.

RNA is the acronym for RiboNucleicAcid.

A graph is **scale free** if obeys a *power-law distribution*, that is, if the number of verticeswith degree k is rougly $k^{-\gamma}$ for a fixed γ.

A **small world** graph is one in which most of the vertices have low degree, such as a graph with a *power-law distribution*.

The **secondary structure** of a protein is the set of alpha helices, beta sheets, turns and loops.

A **sidechain** is another name for *residue*.

The **solvent accessible surface** of a protein is the surface obtained by rolling a ball of a fixed diameter (usually related to the size of a water molecule) around a protein. Such a surface need not be connected.

A protein is **soluble** if it can form a stable and functional form in water.

An **somatic mutation** is one that occurs after birth, and is neither inherited or passed on to offspring, such as occur during antibody maturation.

A **steric** effect is one that involves the shape of an object (the word steric derives from the Greek word for 'shape;' also see the explanation of *allostery*).

Protein **structure** is hierarchical, involving *primary*, *secondary*, *tertiary*, and, for a *protein complex*, *quartenary* structure.

A **subunit** of a *protein complex* is one of the proteins in the collection.

The **tertiary structure** of a protein is the three-dimensional shape of the fully folded proteins.

A **tetramer** is an object made of four *monomers*, typically the same or very similar.

The **three-letter code** for RNA and DNA is the sequence of three letters that code for a particular amino acid.

A **trimer** is an object made of three *monomers*, typically the same or very similar.

A **turn** in a protein structure is a short sequence (typically four) of residues without alpha-helix or beta-sheet structure.

UWHB is the acronym for underwrapped hydrogen bond, a.k.a. dehdydron.

Vicinal means 'in the vicinity' or nearby.

A **widget** is an object that can be used in larger systems in a generic way.

Bibliography

[1] H. Adams, K.D.M. Harris, G.A. Hembury, C.A. Hunter, D. Livingstone, and J.F. McCabe. How strong is a π-facial hydrogen bond? *Chemical Communications*, (22):2531–2532, 1996.

[2] Noam Agmon. Tetrahedral displacement: The molecular mechanism behind the Debye relaxation in water. *Journal of Physical Chemistry*, 100(1):1072–1080, 1996.

[3] C. J. Allègre, G. Manhès, and Christa Göpel. The age of the earth. *Geochimica et Cosmochimica Acta*, 59(8):1445–1456, 1995.

[4] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford Science Publications, Clarendon Press, Oxford, 1989.

[5] Patrick Aloy and Robert B. Russell. Interrogating protein interaction networks through structural biology. *Proceedings of the National Academy of Sciences, USA*, 99:5896–5901, 2002.

[6] D. E. Anderson, J. H. Hurley, H. Nicholson, W. A. Baase, and B. W. Matthews. Hydrophobic core repacking and aromatic-aromatic interaction in the thermostable mutant of T4 lysozyme Ser 117 → Phe. *Protein Science*, 2(8):1285–1290, 1993.

[7] D. Eric Anderson, Wayne J. Becktel, and F. W. Dahlquist. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, 29(9):2403–2408, 1990.

[8] Antonina Andreeva, Dave Howorth, John-Marc Chandonia, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia, and Alexey G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.*, 36:D419–425, 2008.

[9] G. Apic, J. Gough, and S. A. Teichmann. An insight into domain combinations. *Bioinformatics*, 17 (Suppl.1):S83–89, 2001.

[10] Jon Applequist, James R. Carl, and Kwok-Kueng Fung. Atom dipole interaction model for molecular polarizability. application to polyatomic molecules and determination of atom polarizabilities. *Journal of the American Chemical Society*, 94(9):2952–2960, 1972.

[11] D.H. Ardell and G. Sella. No accident: genetic codes freeze in error–correcting patterns of the standard genetic code. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1427):1625, 2002.

[12] Isaiah T. Arkin, Huafeng Xu, Morten O. Jensen, Eyal Arbely, Estelle R. Bennett, Kevin J. Bowers, Edmond Chow, Ron O. Dror, Michael P. Eastwood, Ravenna Flitman-Tene, Brent A. Gregersen, John L. Klepeis, Istvan Kolossvary, Yibing Shan, and David E. Shaw. Mechanism of Na+/H+ antiporting. *Science*, 317(5839):799–803, 2007.

[13] Michelle R. Arkin and James A. Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature Reviews-Drug Discovery*, 3(4):301–317, 2004.

[14] Jóhanna Arnórsdóttir, Asta Rós Sigtryggsdóttir, Sigrithur H. Thorbjarnardóttir, and Magnús M. Kristjánsson. Effect of proline substitutions on stability and kinetic properties of a cold adapted subtilase. *J. Biochem*, 145(3):325–329, 2009.

[15] Kiyoshi Asai, Satoru Hayamizu, and Ken'ichi Handa. Prediction of protein secondary structure by the hidden Markov model. *Comput. Appl. Biosci.*, 9(2):141–146, 1993.

[16] R. Aurora and G. D. Rose. Helix capping. *Protein Science*, 7(1):21–38, 1998.

[17] F. Avbelj and J. Moult. Role of electrostatic screening in determining protein main chain conformational preferences. *Biochemistry*, 34(3):755–764, 1995.

[18] Franc Avbelj. Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins. *Journal of Molecular Biology*, 300(5):1335–1359, 2000.

[19] Franc Avbelj and Robert L. Baldwin. Role of backbone solvation in determining thermodynamic β propensities of the amino acids. *Proceedings of the National Academy of Sciences, USA*, 99(3):1309–1313, 2002.

[20] Franc Avbelj, Simona Golic Grdadolnik, Joze Grdadolnik, and Robert L. Baldwin. Intrinsic backbone preferences are fully present in blocked amino acids. *Proceedings of the National Academy of Sciences, USA*, 103(5):1272–1277, 2006.

[21] Franc Avbelj, Peizhi Luo, and Robert L. Baldwin. Energetics of the interaction between water and the helical peptide group and its role in determining helix propensities. *Proceedings of the National Academy of Sciences, USA*, 97:10786–10791, 2000.

[22] J.C. Avise, B.W. Bowen, T. Lamb, A.B. Meylan, and E. Bermingham. Mitochondrial DNA evolution at a turtle's pace: evidence for low genetic variability and reduced microevolutionary rate in the Testudines. *Mol Biol Evol*, 9(3):457–473, 1992.

[23] R. Azriel and E. Gazit. Analysis of the minimal amyloid-forming fragment of the islet amyloid polypeptide. An experimental support for the key role of the phenylalanine residue in amyloid formation. *Journal of Biological Chemistry*, 276(36):34156–34161, Sep 2001.

[24] C. Azuara, E. Lindahl, P. Koehl, H. Orland, and M. Delarue. PDB_Hydro: incorporating dipolar solvents with variable density in the Poisson-Boltzmann treatment of macromolecule electrostatics. *Nucleic Acids Research*, 34(Web Server issue):W38, 2006.

[25] C. Azuara, H. Orland, M. Bon, P. Koehl, and M. Delarue. Incorporating dipolar solvents with variable density in Poisson-Boltzmann electrostatics. *Biophysical Journal*, 95(12):5587–5605, 2008.

[26] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joöl Janin. Dissecting subunit interfaces in homodimeric proteins. *Proteins: Structure, Function, and Genetics*, 53:708–719, 2003.

[27] Ranjit Prasad Bahadur, Pinak Chakrabarti, Francis Rodier, and Joöl Janin. A dissection of specific and non-specific protein-protein interfaces. *Journal of Molecular Biology*, 336:943–955, 2004.

[28] Y. Bai and S.W. Englander. Hydrogen bond strength and β-sheet propensities: The role of a side chain blocking effect. *Proteins-Structure Function and Genetics*, 18(3):262–266, 1994.

[29] Yawen Bai, John S. Milne, Leland Mayne, and S. Walter Englander. Primary structure effects on peptide group hydrogen exchange. *Proteins-Structure Function and Genetics*, 17:75–86, 1993.

[30] E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, 44:97–179, 1984.

[31] R. L. Baldwin. In search of the energetic role of peptide hydrogen bonds. *Journal of Biological Chemistry*, 278(20):17581–17588, 2003.

[32] R.L. Baldwin. Energetics of protein folding. *Journal of molecular biology*, 371(2):283–301, 2007.

[33] Robert L. Baldwin. Protein folding: Making a network of hydrophobic clusters. *Science*, 295:1657–1658, 2002.

[34] Y.-E. A. Ban, H. Edelsbrunner, and J. Rudolph. Interface surfaces for protein-protein complexes. In *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 205–212, 2004.

[35] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[36] A. L. Barabasi. *Linked: The New Science of Networks*. Perseus, New York, 2002.

[37] K. Bartik, C. Redfield, and C. M. Dobson. Measurement of the individual pK_a values of acidic residues of hen and turkey lysozymes by two-dimensional ^{1}H NMR. *Biophysical Journal*, 66:1180–1184, Apr 1994.

[38] George K. Batchelor. *An introduction to fluid dynamics*. Cambridge Univ Pr, 2000.

[39] Paul A. Bates, Pawel Dokurno, Paul S. Freemont, and Michael J. E. Sternberg. Conformational analysis of the first observed non-proline cis-peptide bond occurring within the complementarity determining region (CDR) of an antibody. *Journal of Molecular Biology*, 284(3):549 – 555, 1998.

[40] Enrique R. Batista, Sotiris S. Xantheas, and Hannes Jonsson. Multipole moments of water molecules in clusters and ice Ih from first principles calculations. *Journal of Chemical Physics*, 111(13):6011–6015, 1999.

[41] Arieh Y. Ben-Naim. *Hydrophobic Interactions*. Springer, 1980.

[42] R. Berisio, V. S. Lamzin, F. Sica, K. S. Wilson, A. Zagari, and L. Mazzarella. Protein titration in the crystal state. *Journal of Molecular Biology*, 292:845–854, 1999.

[43] D. Berleant, M. White, E. Pierce, E. Tudoreanu, A. Boeszoermenyi, Y. Shtridelman, and J.C. Macosko. The genetic code—more than just a table. *Cell Biochemistry and Biophysics*, 55(2):107–116, 2009.

[44] Charles C. Bigelow. On the average hydrophobicity of proteins and the relation between it and protein structure. *Journal of Theoretical Biology*, 16(2):187 – 211, 1967.

[45] Christophe Biot, Eric Buisine, and Marianne Rooman. Free-energy calculations of proteinligand cationπ and aminoπ interactions: From vacuum to proteinlike environments. *Journal of the American Chemical Society*, 125(46):13988–13994, 2003. PMID: 14611235.

[46] Marco Bittelli, Markus Flury, and Kurt Roth. Use of dielectric spectroscopy to estimate ice content in frozen porous media. *Water Resources Research*, 40:W04212, 2004.

[47] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280:1–9, 1998.

[48] A. Bondi. van der Waals volumes and radii. *Journal of Physical Chemistry*, 68(3):441–451, 1964.

[49] P. A. Bopp, A. A. Kornyshev, and G. Sutmann. Static nonlocal dielectric function of liquid water. *Physical Review Letters*, 76:1280–1283, 1996.

[50] Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H. Lackner, Jrg Bhl, Valerie Wood, Kara Dolinsk, and Mike Tyers. The BioGRID Interaction Database: 2008 update. *Nucl. Acids Res.*, 36(database issue):D637–640, 2008.

[51] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer-Verlag, third edition, 2008.

[52] E. Breslow, V. Mombouyran, R. Deeb, C. Zheng, J.P. Rose, BC Wang, and RH Haschemeyer. Structural basis of neurophysin hormone specificity: Geometry, polarity, and polarizability in aromatic ring interactions. *Protein Science*, 8(4):820–831, 1999.

[53] Dawn J. Brooks, Jacques R. Fresco, Arthur M. Lesk, and Mona Singh. Evolution of amino acid frequencies in proteins over deep time: Inferred order of introduction of amino acids into the genetic code. *Molecular Biology and Evolution*, 19:1645–1655, 2002.

[54] William Fuller Brown, Jr. Dielectrics. In *Handbuch der Physik*, volume 14, pages 1–154. Springer, 1956.

[55] R. G. Bryant. The dynamics of water-protein interactions. *Annu. Rev. Biophys. Biomol. Struct.*, 25:29–53, 1996.

[56] M. Bryliński, L. Konieczny, and I. Roterman. Ligation site in proteins recognized in silico. *Bioinformation*, 1(4):127–129, 2006.

[57] M. Bryliński, K. Prymula, W. Jurkowski, M. Kochańczyk, E. Stawowczyk, L. Konieczny, and I. Roterman. Prediction of functional sites based on the fuzzy oil drop model. *PLoS computational biology*, 3(5):e94, 2007.

[58] A. D. Buckingham and P. W. Fowler. Do electrostatic interactions predict structures of van der Waals molecules? *Journal of Chemical Physics*, 79(12):6426–6428, 1983.

[59] A. D. Buckingham and P. W. Fowler. A model for the geometries of van der Waals complexes. *Canadian Journal of Chemistry*, 63:2018–2025, 1985.

[60] S.K. Burley and G.A. Petsko. Amino-aromatic interactions in proteins. *FEBS Letters*, 203(2):139–143, 1986.

[61] Stephen K. Burley and Gregory A. Petsko. Electrostatic interactions in aromatic oligopeptides contribute to protein stability. *Trends in Biotechnology*, 7(12):354 – 359, 1989.

[62] A. Buzzell. Action of urea on Tobacco Mosaic Virus II. The bonds between protein subunits. *Biophysical Journal*, 2(2P1):223–233, 1962.

[63] Christopher Bystroff, Vesteinn Thorsson, and David Baker. Hmmstr: a hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology*, 301(1):173 – 190, 2000.

[64] Orlando M. Cabarcos, Corey J. Weinheimer, and James M. Lisy. Competitive solvation of K^+ by benzene and water: Cation-π interactions and π-hydrogen bonds. *Journal of Chemical Physics*, 108(13):5151–5154, 1998.

[65] Daniel R. Caffrey, Shyamal Somaroo, Jason D. Hughes, Julian Mintseris, and Enoch S. Huang. Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Science*, 13(1):190–202, 2004.

[66] James W. Caldwell and Peter A. Kollman. Cation-π interactions: Nonadditive effects are critical in their accurate representation. *Journal of the American Chemical Society*, 117(14):4177–4178, 1995.

[67] Heather A. Carlson and J. Andrew McCammon. Accommodating Protein Flexibility in Computational Drug Design. *Mol Pharmacol*, 57(2):213–218, 2000.

[68] Fiona J. Carver, C.A. Hunter, D.J. Livingstone, J.F. McCabe, and E.M. Seward. Substituent effects on edge-to-face aromatic interactions. *Chemistry–A European Journal*, 8(13):2847–2859, 2002.

[69] Ana Cauerhff, Fernando A. Goldbaum, and Bradford C. Braden. Structural mechanism for affinity maturation of an anti-lysozyme antibody. *Proceedings of the National Academy of Sciences, USA*, 101(10):3539–3544, 2004.

[70] P. Chakrabarti and R. Bhattacharyya. Geometry of nonbonded interactions involving planar groups in proteins. *Progress in biophysics and molecular biology*, 95(1-3):83–137, 2007.

[71] Pinak Chakrabarti and Joël Janin. Dissecting protein-protein recognition sites. *Proteins: Structure, Function, and Genetics*, 47:334–343, 2002.

[72] Jakub Chalupsky, Jiri Vondrasek, and Vladimir Spirko. Quasiplanarity of the peptide bond. *The Journal of Physical Chemistry A*, 112(4):693–699, 2008.

[73] David Chandler. Interfaces and the driving force of hydrophobic assembly. *Nature*, 437(7059):640–647, 2005.

[74] Dmitry A. Cherepanov. Force oscillations and dielectric overscreening of interfacial water. *Physical Review Letters*, 93:266104, 2004.

[75] Hwanho Choi, Hongsuk Kang, and Hwangseo Park. New angle-dependent potential energy function for backbonebackbone hydrogen bond in proteinprotein interactions. *Journal of Computational Chemistry*, 31(5):897–903, 2010.

[76] Lillian T. Chong, Yong Duan, Lu Wang, Irina Massova, and Peter A. Kollman. Molecular dynamics and free-energy calculations applied to affinity maturation in antibody 48G7. *Proceedings of the National Academy of Sciences*, 96(25):14330–14335, 1999.

[77] Jill E. Chrencik, Jillian Orans, Linda B. Moore, Yu Xue, Li Peng, Jon L. Collins, G. Bruce Wisely, Millard H. Lambert, Steven A. Kliewer, and Matthew R. Redinbo. Structural disorder in the complex of human Pregnane X Receptor and the macrolide antibiotic Rifampicin. *Mol Endocrinol*, 19(5):1125–1134, 2005.

[78] T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383–386, 1995.

[79] M. Cohen, D. Reichmann, H. Neuvirth, and G. Schreiber. Similar chemistry, but different bond preferences in inter versus intra-protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 72(2):741–753, 2008.

[80] Michael S. Cohen, Chao Zhang, Kevan M. Shokat, and Jack Taunton. Structural bioinformatics-based design of selective, irreversible kinase inhibitors. *Science*, 308(5726):1318–1321, 2005.

[81] Philip Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nature Reviews-Drug Discovery*, 1(4):309–315, 2002.

[82] C. Colovos and T. O. Yeates. Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, 2(9):1511–1519, 1993.

[83] PR Connelly, RA Aldape, FJ Bruzzese, SP Chambers, MJ Fitzgibbon, MA Fleming, S. Itoh, DJ Livingston, MA Navia, JA Thomson, et al. Enthalpy of hydrogen bond formation in a protein-ligand binding reaction. *Proceedings of the National Academy of Sciences, USA*, 91(5):1964–1968, 1994.

[84] S. Costantini, G. Colonna, and A.M. Facchiano. Esbri: A web server for evaluating salt bridges in proteins. *Bioinformation*, 3(3):137, 2008.

[85] François-Xavier Coudert, Rodolphe Vuilleumiear, and Anne Boutin. Dipole moment, hydrogen bonding and IR spectrum of confined water. *ChemPhysChem*, 7(12):2464–2467, 2006.

[86] Christopher J. Cramer. *Essentials of Computational Chemistry*. Wiley, second edition, 2004.

[87] Thomas E. Creighton. *Proteins: Structures and molecular properties*. W. H. Freeman, 1993.

[88] Peter B. Crowley and Adel Golovin. Cation-π interactions in protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 59:231–239, 2005.

[89] B.I. Dahiyat, D. Benjamin Gordon, and S.L. Mayo. Automated design of the surface positions of protein helices. *Protein science*, 6(6):1333–1337, 1997.

[90] Matthew Davies, Christopher Toseland, David Moss, and Darren Flower. Benchmarking pK_a prediction. *BMC Biochemistry*, 7(1):18, 2006.

[91] Alfonso De Simone, Guy G. Dodson, Chandra S. Verma, Adriana Zagari, and Franca Fraternali. Prion and water: Tight and dynamical hydration sites have a key role in structural stability. *Proceedings of the National Academy of Sciences, USA*, 102:7535–7540, 2005.

[92] Alfonso De Simone, Roberta Spadaccini, Piero A. Temussi, and Franca Fraternali. Toward the understanding of MNEI sweetness from hydration map surfaces. *Biophys. J.*, 90(9):3052–3061, 2006.

[93] Alfonso De Simone, Adriana Zagari, and Philippe Derreumaux. Structural and hydration properties of the partially unfolded states of the prion protein. *Biophys. J.*, 93(4):1284–1292, 2007.

[94] P. Debye. *Polar Molecules.* Dover, New York, 1945.

[95] W. L. DeLano, M. H. Ultsch, A. M. de Vos, and J. A. Wells. Convergent solutions to binding at a protein-protein interface. *Science*, 287:1279–1283, 2000.

[96] George D. Demetri. Structural reengineering of imatinib to decrease cardiac risk in cancer therapy. *The Journal of Clinical Investigation*, 117(12):3650–3653, 2007.

[97] A. H. DePace, A. Santoso, P. Hillner, and J. S. Weissman. A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion. *Cell*, 93(7):1241–1252, Jun 1998.

[98] Cyril Deremble and Richard Lavery. Macromolecular recognition. *Current Opinion in Structural Biology*, 15(2):171–175, 2005.

[99] Florin Despa and R. Stephen Berry. The origin of long range attraction between hydrophobes in water. *Biophysical Journal*, 92:373–378, 2007.

[100] Florin Despa, Ariel Fernández, and R. Stephen Berry. Dielectric modulation of biological water. *Physical Review Letters*, 93:269901, 2004.

[101] M. Di Giulio. On the origin of the genetic code. *Journal of Theoretical Biology*, 187(4):573–581, 1997.

[102] Christopher M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[103] Matthew Dobson and Mitchell Luskin. Iterative solution of the quasicontinuum equilibrium equations with continuation. *Journal of Scientific Computing*, 37(1):19–41, 2008.

[104] J.E. Donald, D.W. Kulp, and W.F. DeGrado. Salt bridges: Geometrically specific, designable interactions. *Proteins: Structure, Function, and Bioinformatics*, 79(3):898–915, 2011.

[105] Dennis A. Dougherty. Cation-pi interactions in chemistry and biology: a new view of benzene, Phe, Tyr, and Trp. *Science*, 271(5246):163–168, 1996.

[106] John Doyle. Beyond the sperical cow. *Nature*, 411:151–152, May 2001.

[107] Xiaoqun Joyce Duan, Ioannis Xenarios, and David Eisenberg. Describing biological protein interactions in terms of protein states and state transitions. *Mol. Cell. Proteomics*, 1:104–116, 2002.

[108] Y. Duan, C. Wu, S. Chowdhury, M.C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, 2003.

[109] Johannes J. Duistermaat. *Fourier Integral Operators*. Springer, Berlin, 1996.

[110] J. Dunbar, H. P. Yennawar, S. Banerjee, J. Luo, and G. K. Farber. The effect of denaturants on protein structure. *Protein Sci.*, 6:1727–1733, Aug 1997.

[111] Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[112] David Eisenberg, Robert M. Weiss, and Thomas C. Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Sciences, USA*, 81(1):140–144, 1984.

[113] A. Elgsaeter, B. T. Stokke, A. Mikkelsen, and D. Branton. The molecular basis of erythrocyte shape. *Science*, 234:1217–1223, 1986.

[114] Luciana Esposito, Alfonso De Simone, Adriana Zagari, and Luigi Vitagliano. Correlation between ω and ψ dihedral angles in protein structures. *Journal of Molecular Biology*, 347(3):483 – 487, 2005.

[115] E. Eyal, S. Gerzon, V. Potapov, M. Edelman, and V. Sobolev. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *Journal of Molecular Biology*, 351(2):431–442, 2005.

[116] Miles A. Fabian, William H. Biggs, Daniel K. Treiber, Corey E. Atteridge, Mihai D. Azimioara, Michael G. Benedetti, Todd A. Carter, Pietro Ciceri, Philip T. Edeen, Mark Floyd, Julia M. Ford, Margaret Galvin, Jay L. Gerlach, Robert M. Grotzfeld, Sanna Herrgard, Darren E. Insko, Michael A. Insko, Andiliy G. Lai, Jean-Michel Lelias, Shamal A. Mehta, Zdravko V. Milanov, Anne Marie Velasco, Lisa M. Wodicka, Hitesh K. Patel, Patrick P. Zarrinkar, and David J. Lockhart. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nature Biotechnology*, 23(3):329–336, 2005.

[117] M. Fandrich, M. A. Fletcher, and C. M. Dobson. Amyloid fibrils from muscle myoglobin. *Nature*, 410(6825):165–166, Mar 2001.

[118] A. Fernández, A. Sanguino, Z. Peng, A. Crespo, E. Ozturk, X. Zhang, S. Wang, W. Bornmann, and G. Lopez-Berestein. Rational drug redesign to overcome drug resistance in cancer therapy: imatinib moving target. *Cancer research*, 67(9):4028, 2007.

[119] A. Fernández, A. Sanguino, Z. Peng, E. Ozturk, J. Chen, A. Crespo, S. Wulf, A. Shavrin, C. Qin, J. Ma, et al. An anticancer C-Kit kinase inhibitor is reengineered to make it more active and less cardiotoxic. *Journal of Clinical Investigation*, 117(12):4044, 2007.

[120] Ariel Fernández. Cooperative walks in a cubic lattice: Protein folding as a many-body problem. *J. Chem. Phys.*, 115:7293–7297, 2001.

[121] Ariel Fernández. Intramolecular modulation of electric fields in folding proteins. *Phys. Lett. A*, 299:217–220, 2002.

[122] Ariel Fernández. Buffering the entropic cost of hydrophobic collapse in folding proteins. *Journal of Chemical Physics*, 121:11501–11502, 2004.

[123] Ariel Fernández. Keeping dry and crossing membranes. *Nature Biotechnology*, 22:1081–1084, 2004.

[124] Ariel Fernández. Direct nanoscale dehydration of hydrogen bonds. *Journal of Physics D: Applied Physics*, 38:2928–2932, 2005.

[125] Ariel Fernández. Incomplete protein packing as a selectivity filter in drug design. *Structure*, 13:1829–1836, 2005.

[126] Ariel Fernández. What factor drives the fibrillogenic association of beta-sheets? *FEBS Letters*, 579:6635–6640, 2005.

[127] Ariel Fernández. *Transformative concepts for drug design.* Springer, 2010.

[128] Ariel Fernández and R. Stephen Berry. Extent of hydrogen-bond protection in folded proteins: a constraint on packing architectures. *Biophysical Journal*, 83(5):2475–2481, Nov 2002.

[129] Ariel Fernández and R. Stephen Berry. Proteins with H-bond packing defects are highly interactive with lipid bilayers: Implications for amyloidogenesis. *Proceedings of the National Academy of Sciences, USA*, 100:2391–2396, 2003.

[130] Ariel Fernández and R. Stephen Berry. Molecular dimension explored in evolution to promote proteomic complexity. *Proceedings of the National Academy of Sciences, USA*, 101:13460–13465, 2004.

[131] Ariel Fernández and R. Stephen Berry. Golden rule for buttressing vulnerable soluble proteins. *Journal of Proteome Research*, 9(5):2643–2648, 2010. PMID: 20364868.

[132] Ariel Fernández and M. Boland. Solvent environment conducive to protein aggregation. *FEBS Letters*, 529:298–303, 2002.

[133] Ariel Fernández, Jianping Chen, and Alejandro Crespo. Solvent-exposed backbone loosens the hydration shell of soluble folded proteins. *Journal of Chemical Physics*, 126(24):245103, 2007.

[134] Ariel Fernández, Alejandro Crespo, Sridhar Maddipati, and L. Ridgway Scott. Bottom-up engineering of peptide cell translocators based on environmentally modulated quadrupole switches. *ACS Nano*, 2:61–68, 2008.

[135] Ariel Fernández, Jösef Kardos, L. Ridgway Scott, Yuji Goto, and R. Stephen Berry. Structural defects and the diagnosis of amyloidogenic propensity. *Proceedings of the National Academy of Sciences, USA*, 100(11):6446–6451, 2003.

[136] Ariel Fernández and Sridhar Maddipati. A priori inference of cross reactivity for drug-targeted kinases. *Journal of Medicinal Chemistry*, 49(11):3092–3100, 2006.

[137] Ariel Fernández, Kristina Rogale Plazonic, L. Ridgway Scott, and Harold A. Scheraga. Inhibitor design by wrapping packing defects in HIV-1 proteins. *Proceedings of the National Academy of Sciences, USA*, 101:11640–11645, 2004.

[138] Ariel Fernández, Angela Sanguino, Zhenghong Peng, Alejandro Crespo, Eylem Ozturk, Xi Zhang, Shimei Wang, William Bornmann, and Gabriel Lopez-Berestein. Rational drug redesign to overcome drug resistance in cancer therapy: Imatinib moving target. *Cancer Res*, 67(9):4028–4033, 2007.

[139] Ariel Fernández, Angela Sanguino, Zhenghong Peng, Eylem Ozturk, Jianping Chen, Alejandro Crespo, Sarah Wulf, Aleksander Shavrin, Chaoping Qin, Jianpeng Ma, Jonathan Trent, Yvonne Lin, Hee-Dong Han, Lingegowda S. Mangala, James A. Bankson, Juri Gelovani, Allen Samarel, William Bornmann, Anil K. Sood, and Gabriel Lopez-Berestein. An anticancer C-Kit kinase inhibitor is reengineered to make it more active and less cardiotoxic. *The Journal of Clinical Investigation*, 117(12):4044–4054, 2007.

[140] Ariel Fernández and Harold A. Scheraga. Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. *Proceedings of the National Academy of Sciences, USA*, 100(1):113–118, Jan 2003.

[141] Ariel Fernández and L. Ridgway Scott. Adherence of packing defects in soluble proteins. *Physical Review Letters*, 91:18102(4), 2003.

[142] Ariel Fernández and L. Ridgway Scott. Dehydron: a structurally encoded signal for protein interaction. *Biophysical Journal*, 85:1914–1928, 2003.

[143] Ariel Fernández and L. Ridgway Scott. Under-wrapped soluble proteins as signals triggering membrane morphology. *Journal of Chemical Physics*, 119(13):6911–6915, 2003.

[144] Ariel Fernández and L. Ridgway Scott. Modulating drug impact by wrapping target proteins. *Expert Opinion on Drug Discovery*, 2:249–259, 2007.

[145] Ariel Fernández, L. Ridgway Scott, and R. Stephen Berry. The nonconserved wrapping of conserved protein folds reveals a trend towards increasing connectivity in proteomic networks. *Proceedings of the National Academy of Sciences, USA*, 101(9):2823–2827, 2004.

[146] Ariel Fernández, L. Ridgway Scott, and R. Stephen Berry. Packing defects as selectivity switches for drug-based protein inhibitors. *Proceedings of the National Academy of Sciences, USA*, 103:323–328, 2006.

[147] Ariel Fernández, L. Ridgway Scott, and Harold A. Scheraga. Amino-acid residues at protein-protein interfaces: Why is propensity so different from relative abundance? *J. Phys. Chem. B*, 107(36):9929–9932, 2003.

[148] Ariel Fernández, Tobin R. Sosnick, and Andres Colubri. Dynamics of hydrogen bond desolvation in protein folding. *Journal of Molecular Biology*, 321(4):659–675, Aug 2002.

[149] Ariel Fernández, Xi Zhang, and Jianping Chen. Hitting an imatinib-resistant target through an imatinib redesign. *Journal of Biological Chemistry*, Accelerated Publication:accepted, 2006.

[150] Josephine C. Ferreon and Vincent J. Hilser. The effect of the polyproline II (PPII) conformation on the denatured state entropy. *Protein Science*, 12(3):447–457, 2003.

[151] Robert D. Finn, John Tate, Jaina Mistry, Penny C. Coggill, Stephen John Sammut, Hans-Rudolf Hotz, Goran Ceric, Kristoffer Forslund, Sean R. Eddy, Erik L. L. Sonnhammer, and Alex Bateman. The Pfam protein families database. *Nucl. Acids Res.*, 36:D281–288, 2008.

[152] J. E. Fitzgerald, A. K. Jha, T. R. Sosnick, and K. F. Freed. Polypeptide motions are dominated by peptide group oscillations resulting from dihedral angle correlations between nearest neighbors. *Biochemistry*, 46(3):669–682, 2007.

[153] Anton F. Fliri, William T. Loging, Peter F. Thadeio, and Robert A. Volkmann. Biological spectra analysis: Linking biological activity profiles to molecular structure. *Proceedings of the National Academy of Sciences, USA*, 102(2):261–266, 2005.

[154] B. Folch, M. Rooman, and Y. Dehouck. Thermostability of salt bridges versus hydrophobic interactions in proteins probed by statistical potentials. *Journal of chemical information and modeling*, 48(1):119–127, 2008.

[155] L. R. Forrest and B. Honig. An assessment of the accuracy of methods for predicting hydrogen positions in protein structures. *Proteins: Structure, Function, and Bioinformatics*, 62(2):296–309, 2005.

[156] William R. Forsyth, Jan M. Antosiewicz, and Andrew D. Robertson. Empirical relationships between protein structure and carboxyl pK_a values in proteins. *Proteins–Structure, Function, and Genetics*, 48(2):388–403, 2002.

[157] Rosalind Franklin and R. G. Gosling. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature*, 172:156–157, 1953.

[158] Felix Franks. *Water: a matrix for life*. Royal Society of Chemistry, 2000.

[159] J.S. Franzen and R.E. Stephens. The effect of a dipolar solvent system on interamide hydrogen bonds. *Biochemistry*, 2(6):1321–1327, 1963.

[160] Ernesto Freire. The propagation of binding interactions to remote sites in proteins: Analysis of the binding of the monoclonal antibody D1.3 to lysozyme. *Proceedings of the National Academy of Sciences, USA*, 96:10118–10122, 1999.

[161] Robert H. French. Origins and applications of London dispersion forces and Hamaker constants in ceramics. *Journal of the American Ceramic Society*, 83(9):2117–46, 2000.

[162] E. Fry, R. Acharya, and D. Stuart. Methods used in the structure determination of foot-and-mouth disease virus. *Acta Crystall. A*, 49:45–55, 1993.

[163] R.L. Fulton. The nonlinear dielectric behavior of water: Comparisons of various approaches to the nonlinear dielectric increment. *Journal of Chemical Physics*, 130:204503, 2009.

[164] Justin P. Gallivan and Dennis A. Dougherty. Cation-pi interactions in structural biology. *Proceedings of the National Academy of Sciences, USA*, 96(17):9459–9464, 1999.

[165] J. Gao, D.A. Bosco, E.T. Powers, and J.W. Kelly. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. *Nature Structural & Molecular Biology*, 16(7):684–690, 2009.

[166] Angel E. García and Kevin Y. Sanbonmatsu. α-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences, USA*, 99(5):2782–2787, 2002.

[167] Roxana E. Georgescu, Emil G. Alexov, and Marilyn R. Gunner. Combining conformational flexibility and continuum electrostatics for calculating pK_as in proteins. *Biophysical Journal*, 83(4):1731 – 1748, 2002.

[168] D. Gilis, S. Massar, N.J. Cerf, and M. Rooman. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. *Genome Biol*, 2(11):49–1, 2001.

[169] I. Gitlin, J.D. Carbeck, and G.M. Whitesides. Why are proteins charged? networks of charge–charge interactions in proteins measured by charge ladders and capillary electrophoresis. *Angewandte Chemie International Edition*, 45(19):3022–3060, 2006.

[170] Fabian Glaser, David M. Steinberg, Ilya A. Vakser, and Nir Ben-Tal. Residue frequencies and pairing preferences at protein-protein interfaces. *Proteins: Structure, Function, and Genetics*, 43:89–102, 2001.

[171] G. A. Grant, C. W. Luetje, R. Summers, and X. L. Xu. Differential roles for disulfide bonds in the structural integrity and biological activity of κ-bungarotoxin, a neuronal nicotinic acetylcholine receptor antagonist. *Biochemistry*, 37(35):12166–12171, 1998.

[172] J. K. Gregory, D. C. Clary, K. Liu, M. G. Brown, and R. J. Saykally. The water dipole moment in water clusters. *Science*, 275(5301):814–817, 1997.

[173] A. V. Grinberg and R. Bernhardt. Effect of replacing a conserved proline residue on the function and stability of bovine adrenodoxin. *Protein Eng.*, 11(11):1057–1064, 1998.

[174] Nick V. Grishin. Fold change in evolution of protein structures. *Journal of Structural Biology*, 134(2-3):167 – 185, 2001.

[175] J. A. Gruenke, R. T. Armstrong, W. W. Newcomb, J. C. Brown, and J. M. White. New insights into the spring-loaded conformational change of influenza virus hemagglutinin. *J. Virol.*, 76(9):4456–4466, 2002.

[176] J. I. Guijarro, M. Sunde, J. A. Jones, I. D. Campbell, and C. M. Dobson. Amyloid fibril formation by an SH3 domain. *Proceedings of the National Academy of Sciences, USA*, 95(8):4224–4228, Apr 1998.

[177] S. Günther, J. Von Eichborn, P. May, and R. Preissner. JAIL: a structure-based interface library for macromolecules. *Nucleic Acids Research*, 37(suppl 1):D338–D341, 2009.

[178] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge University Press, January 1997.

[179] Inbal Halperin, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions; coupling of structurally conserved residues and of hot spots across interfaces. implications for docking. *Structure*, 12:1027–1038, 2004.

[180] D. Hamelberg and J. A. McCammon. Fast peptidyl cis-trans isomerization within the flexible Gly-rich flaps of HIV-1 protease. *Journal of the American Chemical Society*, 127(40):13778–13779, 2005.

[181] Yehouda Harpaz, Mark Gerstein, and Cyrus Chothia. Volume changes on protein folding. *Structure*, 2:641–649, 1994.

[182] M. J. Hartshorn, C. W. Murray, A. Cleasby, M. Frederickson, I. J. Tickle, and H. Jhoti. Fragment-based lead discovery using X-ray crystallography. *Journal of Medicinal Chemistry*, 48(2):403–413, 2005.

[183] John Barrett Hasted. *Aqueous Dielectrics.* Chapman and Hall, 1974.

[184] B. Hayes. The invention of the genetic code. *American Scientist*, 86(1):8–14, 1998.

[185] J. P. Helfrich. Dynamic laser light scattering technology for the molecular weight and hydrodynamic radius characterization of proteins. *Pharmaceutical Laboratory*, 1:34–40, 1998.

[186] David H. Herce, Lalith Perera, Thomas A. Darden, and Celeste Sagui. Surface solvation for an ion in a water cluster. *Journal of Chemical Physics*, 122(2):024513, 2005.

[187] M. Heuberger, T. Drobek, and N. D. Spencer. Interaction forces and morphology of a protein-resistant poly(ethylene glycol) layer. *Biophysical Journal*, 88(1):495–504, 2005.

[188] J. Higo and M. Nakasako. Hydration structure of human lysozyme investigated by molecular dynamics simulation and cryogenic X-ray crystal structure analyses: On the correlation between crystal water sites, solvent density, and solvent dipole. *Journal of computational chemistry*, 23(14):1323–1336, 2002.

[189] A. Hildebrandt, R. Blossey, S. Rjasanow, O. Kohlbacher, and H.-P. Lenhof. Novel formulation of nonlocal electrostatics. *Physical Review Letters*, 93(10):108104, Sep 2004.

[190] A. Hildebrandt, R. Blossey, S. Rjasanow, O. Kohlbacher, and H.P. Lenhof. Electrostatic potentials of proteins in water: a structured continuum approach. *Bioinformatics*, 23(2):e99, 2007.

[191] M. K. Hill, M. Shehu-Xhilaga, S. M. Crowe, and J. Mak. Proline residues within spacer peptide p1 are important for human immunodeficiency virus type 1 infectivity, protein processing, and genomic RNA dimer stability. *J. Virol.*, 76:11245–11253, Nov 2002.

[192] Nora E. Hill, Worth E. Vaughan, A. H. Price, and Mansel Davies. *Dielectric properties and molecular behaviour.* van Nostrand, London, 1969.

[193] S. Hirohashi and Y. Kanai. Cell adhesion system and human cancer morphogenesis. *Cancer science*, 94(7):575–581, 2003.

[194] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Science*, 1(3):409–417, 1992.

[195] B. Honig and A. S. Yang. Free energy balance in protein folding. *Adv. Protein Chem.*, 46:27–58, 1995.

[196] Andrew L. Hopkins, Jonathan S. Mason, and John P. Overington. Can we rationally design promiscuous drugs? *Current Opinion in Structural Biology*, 16(1):127–136, 2006.

[197] M. Hoshino, H. Katou, Y. Hagihara, K. Hasegawa, H. Naiki, and Y. Goto. Mapping the core of the beta(2)-microglobulin amyloid fibril by H/D exchange. *Nature Structural Biology*, 19:332–336, 2002.

[198] Thomas Hou and Petros Koumoutsakos. Special section on multiscale modeling in materials and life sciences. *SIAM J. Multiscale Modeling & Simulation*, 4(1):213–214, 2005.

[199] Sven Hovmöller, Tuping Zhou, and Tomas Ohlson. Conformations of amino acids in proteins. *Acta Crystallographica Section D*, 58(5):768–776, 2002.

[200] Zengjian Hu, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Genetics*, 39:331–342, 2000.

[201] M.A. Huntley and G.B. Golding. Simple sequences are rare in the Protein Data Bank. *Proteins: Structure, Function, and Bioinformatics*, 48(1):134–140, 2002.

[202] Beatrice M. P. Huyghues-Despointes, Tod M. Klingler, and Robert L. Baldwin. Measuring the strength of side-chain hydrogen bonds in peptide helixes: The Gln·Asp (i, i + 4) interaction. *Biochemistry*, 34:13267–13271, May 2002. doi: 10.1021/bi00041a001.

[203] L. M. Iakoucheva and K. A. Dunker. Order, disorder, and flexibility: prediction from protein sequence. *Structure*, 11:1316–1317, 2003.

[204] Jacob Israelachvili. *Intermolecular and surface forces*. Academic Press, second edition, 1991.

[205] A. Jabs, M.S. Weiss, and R. Hilgenfeld. Non-proline cis peptide bonds in proteins. *Journal of molecular biology*, 286(1):291–304, 1999.

[206] Joël Janin, S. Miller, and Cyrus Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *Journal of Molecular Biology*, 204:155–164, 1988.

[207] Mariusz Jaskolski, Miroslaw Gilski, Zbigniew Dauter, and Alexander Wlodawer. Stereochemical restraints revisited: how accurate are refinement targets and how much should protein structures be allowed to deviate from them? *Acta Crystallographica Section D*, 63(5):611–620, May 2007.

[208] George A. Jeffrey. *An introduction to hydrogen bonds*. Oxford Science Publications, Clarendon Press, Oxford, 1997.

[209] George A. Jeffrey. Hydrogen-bonding: An update. *Crystallography Reviews*, 9:135–176, Jan 2003. 10.1080/08893110310001621754.

[210] Jan H. Jensen, Hui Li, Andrew D. Robertson, and Pablo A. Molina. Prediction and rationalization of protein pK_a values using QM and QM/MM methods. *The Journal of Physical Chemistry A*, 109(30):6634–6643, 2005.

[211] A.K. Jha and K.F. Freed. Solvation effect on conformations of 1, 2: dimethoxyethane: Charge-dependent nonlinear response in implicit solvent models. *Journal of Chemical Physics*, 128:034501, 2008.

[212] S. Ji. The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Annals of the New York Academy of Sciences*, 870:411–417, 1999.

[213] L. Jiang and L. Lai. CH··· O hydrogen bonds at protein-protein interfaces. *Journal of Biological Chemistry*, 277(40):37732, 2002.

[214] S. Jones, A. Marin, and J. M. Thornton. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, 13:77–82, 2000.

[215] Susan Jones and Janet M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences, USA*, 93:13–20, 1996.

[216] Susan Jones and Janet M. Thornton. Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*, 272:121–132, 1997.

[217] Susan Jones and Janet M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*, 272:133–143, 1997.

[218] F. Jourdan, S. Lazzaroni, B.L. Méndez, P.L. Cantore, M. de Julio, P. Amodeo, N.S. Iacobellis, A. Evidente, and A. Motta. A left-handed α-helix containing both L-and D-amino acids: The solution structure of the antimicrobial lipodepsipeptide tolaasin. *Proteins: Structure, Function, and Bioinformatics*, 52(4):534–543, 2003.

[219] U. Kaatze, R. Behrends, and R. Pottel. Hydrogen network fluctuations and dielectric spectrometry of liquids. *Journal of Non-Crystalline Solids*, 305(1):19–28, 2002.

[220] R. Kaufmann, U. Junker, M. Schilli-Westermann, C. Klötzer, J. Scheele, and K. Junker. Meizothrombin, an intermediate of prothrombin cleavage potently activates renal carcinoma cells by interaction with par-type thrombin receptors. *Oncology Reports*, 10:493–496, 2003.

[221] W. Kauzmann. Some factors in the interpretation of protein denaturation. In *Advances in Protein Chemistry*, volume 14, pages 1–63. Academic Press, 1959.

[222] Johan Kemmink and Thomas E. Creighton. The physical properties of local interactions of tyrosine residues in peptides and unfolded proteins. *Journal of Molecular Biology*, 245(3):251 – 260, 1995.

[223] F.N. Keutsch, J.D. Cruzan, and R.J. Saykally. The water trimer. *Chemical reviews*, 103(7):2533–2578, 2003.

[224] Jinrang Kim, Junjun Mao, and M.R. Gunner. Are acidic and basic groups in buried proteins predicted to be ionized? *Journal of Molecular Biology*, 348(5):1283 – 1298, 2005.

[225] S. S. Kim, T. J. Smith, M. S. Chapman, M. C. Rossmann, D. C. Pevear, F. J. Dutko, P. J. Felock, G. D. Diana, and M. A. McKinlay. Crystal structure of human rhinovirus serotype 1A (HRV1A). *Journal of Molecular Biology*, 210:91–111, 1989.

[226] Sangtae Kim and Seppo J. Karrila. *Microhydrodynamics : Principles and Selected Applications*. Dover Publications, 2005.

[227] W. Klopper, J. G. C. M. van Duijneveldt-van de Rijdt, and F. B. van Duijneveldt. Computational determination of equilibrium geometry and dissociation energy of the water dimer. *Physical Chemistry Chemical Physics*, 2:2227– 2234, 2000.

[228] I.M. Klotz. Solvent water and protein behavior: View through a retroscope. *Protein Science*, 2(11), 1993.

[229] I.M. Klotz and J.S. Franzen. Hydrogen bonds between model peptide groups in solution. *Journal of the American Chemical Society*, 84(18):3461–3466, 1962.

[230] Zachary A. Knight and Kevan M. Shokat. Features of selective kinase inhibitors. *Chemistry & Biology*, 12(6):621–637, 2005.

[231] O. Koch, M. Bocola, and G. Klebe. Cooperative effects in hydrogen-bonding of protein secondary structure elements: A systematic analysis of crystal data using secbase. *Proteins: Structure, Function, and Bioinformatics*, 61(2), 2005.

[232] P. Koehl, H. Orland, and M. Delarue. Computing ion solvation free energies using the dipolar Poisson model. *The Journal of Physical Chemistry B*, 113(17):5694–5697, 2009.

[233] Peter A. Kollman and Leland C. Allen. The theory of the hydrogen bond. *Chemical Reviews*, 72(3):283–303, 1972.

[234] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:218–223, 2002.

[235] J. Korlach, P. Schwille, W. W. Webb, and G. W. Feigenson. Characterization of lipid bilayer phases by confocal microscopy and fluorescence correlation spectroscopy. *Proceedings of the National Academy of Sciences, USA*, 96:8461–8466, 1999.

[236] A. A. Kornyshev and A. Nitzan. Effect of overscreening on the localization of hydrated electrons. *Zeitschrift für Physikalische Chemie*, 215(6):701–715, 2001.

[237] Alexei A. Kornyshev and Godehard Sutmann. Nonlocal dielectric saturation in liquid water. *Physical Review Letters*, 79:3435–3438, 1997.

[238] Tanja Kortemme, Alexandre V. Morozov, and David Baker. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *Journal of Molecular Biology*, 326:1239–1259, 2003.

[239] Gennady V. Kozhukh, Yoshihisa Hagihara, Toru Kawakami, Kazuhiro Hasegawa, Hironobu Naiki, and Yuji. Goto. Investigation of a peptide responsible for amyloid fibril formation of β_2-microglobulin by *achromobacter* protease I. *Journal of Biological Chemistry*, 277(2):1310–1315, Jan 2002.

[240] Lawrence M. Krauss. *Fear of Physics*. Basic Books, 1994.

[241] S. Krishnaswamy and Michael G. Rossmann. Structural refinement and analysis of mengo virus. *Journal of Molecular Biology*, 211:803–844, 1990.

[242] J. Kroon, J. A. Kanters, J. G. C. M. van Duijneveldt-van De Rijdt, F. B. van Duijneveldt, and J. A. Vliegenthart. O-H $\cdot\cdot$ O hydrogen bonds in molecular crystals a statistical and quantum-chemical analysis. *Journal of Molecular Structure*, 24:109–129, Jan 1975.

[243] Sandeep Kumar and Ruth Nussinov. Salt bridge stability in monomeric proteins. *Journal of Molecular Biology*, 293(5):1241–1255, 1999.

[244] Sandeep Kumar and Ruth Nussinov. Relationship between ion pair geometries and electrostatic strengths in proteins. *Biophysical Journal*, 83(3):1595 – 1612, 2002.

[245] W. Kunz. *Specific Ion Effects*. World Scientific Pub Co Inc, 2010.

[246] J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157:105–132, 1982.

[247] L. D. Landau and E. M. Lifshitz. Fluid mechanics. *Course of theoretical physics*, 6, 1987.

[248] Paula I. Lario and Alice Vrielink. Atomic resolution density maps reveal secondary structure dependent differences in electronic distribution. *Journal of the American Chemical Society*, 125:12787–12794, Sep 2003. doi: 10.1021/ja0289954.

[249] Wendell M. Latimer and Worth H. Rodebush. Polarity and ionization from the standpoint of the lewis theory of valence. *Journal of the American Chemical Society*, 42:1419–1433, 1920. doi: 10.1021/ja01452a015.

[250] Christian Laurence and Michel Berthelot. Observations on the strength of hydrogen bonding. *Perspectives in Drug Discovery and Design*, 18:3960, 2000.

[251] Andrew Leaver-Fay, Yuanxin Liu, Jack Snoeyink, and Xueyi Wang. Faster placement of hydrogens in protein structures by dynamic programming. *J. Exp. Algorithmics*, 12:2.5:1–2.5:16, June 2008.

[252] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, 55:379–400, 1971.

[253] Eun Cheol Lee, Byung Hee Hong, Ju Young Lee, Jong Chan Kim, Dongwook Kim, Yukyung Kim, P. Tarakeshwar, and Kwang S. Kim. Substituent effects on the edge-to-face aromatic interactions. *Journal of the American Chemical Society*, 127(12):4530–4537, 2005.

[254] J. Lehmann. Physico-chemical constraints connected with the coding properties of the genetic system. *Journal of Theoretical Biology*, 202(2):129–144, 2000.

[255] S. Leikin and A. A. Kornyshev. Theory of hydration forces. Nonlocal electrostatic interaction of neutral surfaces. *Journal of Chemical Physics*, 92(6):6890–6898, 1990.

[256] P. A. Leland, K. E. Staniszewski, C. Park, B. R. Kelemen, and R. T. Raines. The ribonucleolytic activity of angiogenin. *Biochemistry*, 41(4):1343–1350, 2002.

[257] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol*, 104:59–107, 1976.

[258] Michael Levitt and Max F. Perutz. Aromatic rings act as hydrogen bond acceptors. *Journal of Molecular Biology*, 201:751–754, 1988.

[259] Yaakov Levy and Jose N. Onuchic. Water and proteins: A love-hate relationship. *Proceedings of the National Academy of Sciences, USA*, 101(10):3325–3326, 2004.

[260] Hui Li, Andrew D. Robertson, and Jan H. Jensen. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins–Structure, Function, and Bioinformatics*, 61(4):704–721, 2005.

[261] Yili Li, Hongmin Li, Feng Yang, Sandra J. Smith-Gill, and Roy A. Mariuzza. X-ray snapshots of the maturation of an antibody response to a protein antigen. *Nature Structural and Molecular Biology*, 10:482–488, Jun 2003. 10.1038/nsb930.

[262] Laurent Limozin and Erich Sackmann. Polymorphism of cross-linked actin networks in giant vesicles. *Physical Review Letters*, 89(16):168103, Oct 2002.

[263] M. Lisal, J. Kolafa, and I. Nezbeda. An examination of the five-site potential (TIP5P) for water. *Journal of Chemical Physics*, 117:8892–8897, November 2002.

[264] Stephen J. Littler and Simon J. Hubbard. Conservation of orientation and sequence in protein domain-domain interactions. *Journal of Molecular Biology*, 345(5):1265–1279, 2005.

[265] Bernard A. Liu, Karl Jablonowski, Monica Raina, Michael Arcé, Tony Pawson, and Piers D. Nash. The human and mouse complement of SH2 domain proteins — establishing the boundaries of phosphotyrosine signaling. *Molecular Cell*, 22(6):851–868, 2006.

[266] Jing Liu, L. Ridgway Scott, and Ariel Fernández. Interactions of aligned nearest neighbor protein side chains. *Journal of Bioinformatics and Computational Biology*, 7:submitted–, 2006.

[267] Jing Liu, L. Ridgway Scott, and Ariel Fernández. Interactions of aligned nearest neighbor protein side chains. *Journal of Bioinformatics and Computational Biology*, 7:submitted–, 2006.

[268] K. Liu, M. G. Brown, C. Carter, R. J. Saykally, J. K. Gregory, and D. C. Clary. Characterization of a cage form of the water hexamer. *Nature*, 381:501–503, 1996.

[269] Loredana Lo Conte, S. E. Brenner, T. J. Hubbard, Cyrus Chothia, and A. G. Murzin. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, 30:264–267, 2002.

[270] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Mol. Biol.*, 285:2177–2198, 1999.

[271] Vakhtang V. Loladze and George I. Makhatadze. Energetics of charge-charge interactions between residues adjacent in sequence. *Proteins: Structure, Function, and Bioinformatics*, 79(12):3494–3499, 2011.

[272] Simon C. Lovell, J. Michael Word, Jane S. Richardson, and David C. Richardson. The penultimate rotamer library. *Proteins: Structure, Function, and Genetics*, 40:389–408, 2000.

[273] Hui Lu, Long Lu, and Jeffrey Skolnick. MULTIPROSPECTOR: An algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins: Structure, Function, and Genetics*, 49:350–364, 2002.

[274] Peizhi Luo and Robert L. Baldwin. Interaction between water and polar groups of the helix backbone: An important determinant of helix propensities. *Proceedings of the National Academy of Sciences, USA*, 96(9):4930–4935, 1999.

[275] Buyong Ma, Tal Elkayam, Haim Wolfson, and Ruth Nussinov. Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences, USA*, 100(10):5772–5777, 2003.

[276] Jennifer C. Ma and Dennis A. Dougherty. The cation-π interaction. *Chemical Reviews*, 97(5):1303–1324, 1997.

[277] Malcolm W. MacArthur and Janet M. Thornton. Deviations from planarity of the peptide bond in peptides and proteins. *Journal of Molecular Biology*, 264(5):1180 – 1195, 1996.

[278] J. D. Madura, J. M. Briggs, R. C. Wade, M. E. Davis, B. A. Luty, A. Ilin, J. Antosciewicz, M. K. Gilson, B. Bagheri, L. R. Scott, and J. A. McCammon. Electrostatics and diffusion of molecules in solution: Simulations with the University of Houston Brownian Dynamics program. *Comp. Phys. Commun.*, 91:57–95, Sept. 1995.

[279] Michael W. Mahoney and William L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *J. Chem. Phys.*, 112:8910–8922, 2000.

[280] M. Maleki, M.M. Aziz, and L. Rueda. Analysis of obligate and non-obligate complexes using desolvation energies in domain-domain interactions. In *Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics*, page 2. ACM, 2011.

[281] R. J. Mallis, K. N. Brazin, D. B. Fulton, and A. H. Andreotti. Structural characterization of a proline-driven conformational switch within the Itk SH2 domain. *Nat. Struct. Biol.*, 9:900–905, Dec 2002.

[282] K. Manikandan and S. Ramakumar. The occurrence of C–H$\cdots$O hydrogen bonds in α-helices and helix termini in globular proteins. *Proteins: Structure, Function, and Bioinformatics*, 56(4):768–781, 2004.

[283] Raimund Mannhold, Gennadiy I. Poda, Claude Ostermann, and Igor V. Tetko. Calculation of molecular lipophilicity: State-of-the-art and comparison of log P methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences*, 98(3):861–893, 2009.

[284] P. E. Mason, G. W. Neilson, C. E. Dempsey, A. C. Barnes, and J. M. Cruickshank. The hydration structure of guanidinium and thiocyanate ions: Implications for protein stability in aqueous solution. *Proceedings of the National Academy of Sciences, USA*, 100(8):4557–4561, 2003.

[285] C. Matzler and U. Wegmuller. Dielectric properties of freshwater ice at microwave frequencies. *Journal of Physics D: Applied Physics*, 20(12):1623–1630, 1987.

[286] I. K. McDonald and Janet M. Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238:777–793, 1994.

[287] Lawrence P. McIntosh, Greg Hand, Philip E. Johnson, Manish D. Joshi, Michael Körner, Leigh A. Plesniak, Lothar Ziser, Warren W. Wakarchuk, and Stephen G. Withers. The pK_a of the general acid/base carboxyl group of a glycosidase cycles during catalysis: A ^{13}C-NMR study of bacillus circulans xylanase. *Biochemistry*, 35(12):9958–9966, 1996.

[288] Victoria J. McParland, Arnout P. Kalverda, Steve W. Homans, and Sheena E. Radford. Structural properties of an amyloid precursor of beta(2)-microglobulin. *Nature Structural Biology*, 9(5):326–331, May 2002.

[289] Carver Mead and Lynn Conway. *Introduction to VLSI Systems*. Addison-Wesley, 1979.

[290] S. Mecozzi, A. P. West, and D. A. Dougherty. Cation-π interactions in aromatics of biological and medicinal interest: electrostatic potential surfaces as a useful qualitative guide. *Proceedings of the National Academy of Sciences, USA*, 93(20):10566–10571, 1996.

[291] Ernest L. Mehler and Frank Guarnieri. A self-consistent, microenvironment modulated screened Coulomb potential approximation to calculate pH-dependent electrostatic effects in proteins. *Biophysical Journal*, 77:3–22, Jul 1999.

[292] I. Mihalek, I. Reš, and O. Lichtarge. On itinerant water molecules and detectability of protein–protein interfaces through comparative analysis of homologues. *Journal of molecular biology*, 369(2):584–595, 2007.

[293] C. Millot and A.J. Stone. Towards an accurate intermolecular potential for water. *Molecular Physics*, 77(3):439–462, 1992.

[294] D. J. Mitchell, L. Steinman, D. T. Kim, C. G. Fathman, and J. B. Rothbard. Polyarginine enters cells more efficiently than other polycationic homopolymers. *Journal of Peptide Research*, 56(5):318–325, 2000.

[295] John B.O. Mitchell and James Smith. D-amino acid residues in peptides and proteins. *Proteins: Structure, Function, and Genetics*, 50(4):563–571, 2003.

[296] S. Miyazawa and R.L. Jernigan. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, 18(3):534–552, 1985.

[297] Andreas Möglich, Florian Krieger, and Thomas Kiefhaber. Molecular basis for the effect of urea and guanidinium chloride on the dynamics of unfolded polypeptide chains. *Journal of Molecular Biology*, 345(1):153 – 162, 2005.

[298] C. Momany, L. C. Kovari, A. J. Prongay, W. Keller, R. K. Gitti, B. M. Lee, A. E. Gorbalenya, L. Tong, J. McClure, L. S. Ehrlich, M. F. Carter, and M. G. Rossmann. Crystal structure of dimeric HIV-1 capsid protein. *Nature Structural Biology*, 3:763–770, 1996.

[299] F. A. Momany, R. F. McGuire, A. W. Burgess, and H. A. Scheraga. Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids. *Journal of Physical Chemistry*, 79(22):2361–2381, 1975.

[300] Alexandre V. Morozov, Tanja Kortemme, Kiril Tsemekhman, and David Baker. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. *Proceedings of the National Academy of Sciences, USA*, 101:6946–6951, 2004.

[301] James A. Morrill and Roderick MacKinnon. Isolation of a single carboxyl-carboxylate proton binding site in the pore of a cyclic nucleotide-gated channel. *J. Gen. Physiol.*, 114(1):71–84, 1999.

[302] D. W. Mount. *Bioinformatics.* Cold Spring Harbor Laboratory Press, 2001.

[303] A. S. Muresan, H. Diamant, and K.-Y. Lee. Effect of temperature and composition on the formation of nanoscale compartments in phospholipid membranes. *J. Am. Chem. Soc.*, 123:6951–6952, 2001.

[304] A. G. Murzin, S. E. Brenner, T. J. Hubbard, and C. Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

[305] Rebecca Nelson, Michael R. Sawaya, Melinda Balbirnie, Anders O. Madsen, Christian Riekel, Robert Grothe, and David Eisenberg. Structure of the cross-β spine of amyloid-like fibrils. *Nature*, 435:773–778, Jun 2005. 10.1038/nature03680.

[306] G. Némethy, I. Z. Steinberg, and H. A. Scheraga. Influence of water structure and of hydrophobic interactions on the strength of side-chain hydrogen bonds in proteins. *Biopolymers*, 1:43–69, 1963.

[307] Alexander Neumeister, Nicole Praschak-Rieder, Barbara Hesselmann, Oliver Vitouch, Manfred Rauh, Arnd Barocka, Johannes Tauscher, and Siegfried Kasper. Effects of tryptophan depletion in drug-free depressed patients who responded to total sleep deprivation. *Arch. Gen. Psychiatry*, 55(2):167–172, 1998.

[308] Anthony Nicholls, Kim A. Sharp, and Barry Honig. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Structure, Function, and Genetics*, 11(4):281–296, 1991.

[309] Kyoung Tai No, Oh Young Kwon, Su Yeon Kim, Mu Shik Jhon, and Harold A. Scheraga. A simple functional representation of angular-dependent hydrogen-bonded systems. 1. amide, carboxylic acid, and amide-carboxylic acid pairs. *Journal of Physical Chemistry*, 99:3478–3486, 1995.

[310] Martin E. M. Noble, Jane A. Endicott, and Louise N. Johnson. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science*, 303(5665):1800–1805, 2004.

[311] Irene M. A. Nooren and Janet M. Thornton. Diversity of proteinprotein interactions. *EMBO Journal*, 22:3486–3492, 2003.

[312] Irene M. A. Nooren and Janet M. Thornton. Structural characterisation and functional significance of transient protein-protein interactions. *Journal of Molecular Biology*, 325(5):991–1018, 2003.

[313] Marian Novotny and Gerard J. Kleywegt. A survey of left-handed helices in protein structures. *Journal of Molecular Biology*, 347(2):231 – 241, 2005.

[314] Yasuhiko Nozaki and Charles Tanford. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. *Journal of Biological Chemistry*, 246(7):2211–2217, 1971.

[315] Yanay Ofren and Burkhard Rost. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, 335:377–387, 2003.

[316] Utkan Ogmen, Ozlem Keskin, A. Selim Aytuna, Ruth Nussinov, and Attila Gursoy. Prism: protein interactions by structural matching. *Nucleic Acids Research*, 33(suppl 2):W331–W336.

[317] Diana O. Omecinsky, Katherine E. Holub, Michael E. Adams, and Michael D. Reily. Three-dimensional structure analysis of μ-agatoxins: Further evidence for common motifs among neurotoxins with diverse ion channel specificities. *Biochemistry*, 35(9):2836–2844, 1996.

[318] T. Ooi. Thermodynamics of protein folding: effects of hydration and electrostatic interactions. *Advan. Biophys.*, 30:105–154, 1994.

[319] C. Nick Pace, Gerald R. Grimsley, and J. Martin Scholtz. Protein ionizable groups: pK values and their contribution to protein stability and solubility. *Journal of Biological Chemistry*, 284(20):13285–13289, 2009.

[320] C. Nick Pace, Saul Treviño, Erode Prabhakaran, and J. Martin Scholtz. Protein structure, stability and solubility in water and other solvents. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1448):1225–1235, 2004.

[321] Doreen Pahlke, Dietmar Leitner, Urs Wiedemann, and Dirk Labudde. COPS–cis/trans peptide bond conformation prediction of amino acids on the basis of secondary structure information. *Bioinformatics*, 21(5):685–686, 2005.

[322] J. Park, M. Lappe, and S. A. Teichmann. Mapping protein family interactions: Intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *Journal of Molecular Biology*, 307:929–938, 2001.

[323] Wolfgang Pauli. The connection between spin and statistics. *Phys. Rev.*, 58(8):716–722, Oct 1940.

[324] Linus Pauling. *Nature of the Chemical Bond.* Cornell Univ. Press, third edition, 1960.

[325] Linus Pauling. *General Chemistry.* Dover, 1970.

[326] Jiri Pavlicek, Steven L. Coon, Surajit Ganguly, Joan L. Weller, Sergio A. Hassan, Dan L. Sackett, and David C. Klein. Evidence that proline focuses movement of the floppy loop of Arylalkylamine N-Acetyltransferase (EC 2.3.1.87). *J. Biol. Chem.*, 283(21):14552–14558, 2008.

[327] Michael Y. Pavlov, Richard E. Watts, Zhongping Tan, Virginia W. Cornish, Mäns Ehrenberg, and Anthony C. Forster. Slow peptide bond formation by proline and other n-alkylamino acids in translation. *Proceedings of the National Academy of Sciences*, 106(1):50–54, 2009.

[328] Gang Pei, Thomas M. Laue, Ann Aulabaugh, Dana M. Fowlkes, and Barry R. Lentz. Structural comparisons of meizothrombin and its precursor prothrombin in the presence or absence of procoagulant membranes. *Biochemistry*, 31:6990–6996, 1992.

[329] Stefan Persson, J. Antoinette Killian, and Göran Lindblom. Molecular ordering of interfacially localized tryptophan analogs in ester- and ether-lipid bilayers studied by ^{2}H-NMR. *Biophysical Journal*, 75(3):1365–1371, 1998.

[330] M. F. Perutz. The role of aromatic rings as hydrogen-bond acceptors in molecular recognition. *Philosophical Transactions of the Royal Society A*, 345(1674):105–112, 1993.

[331] D. Petrey and B. Honig. Free energy determinants of tertiary structure and the evaluation of protein models. *Protein Science*, 9(11):2181–2191, Nov 2000.

[332] Gregory A. Petsko and Dagmar Ringe. *Protein Stucture and Function.* New Science Press, 2004.

[333] Michael Petukhov, David Cregut, Clàudio M. Soares, and Luis Serrano. Local water bridges and protein conformational stability. *Protein Science*, 8:1982–1989, 1999.

[334] Amira Pierucci-Lagha, Richard Feinn, Vania Modesto-Lowe, Robert Swift, Maggie Nellissery, Jonathan Covault, and Henry R. Kranzler. Effects of rapid tryptophan depletion on mood and urge to drink in patients with co-morbid major depression and alcohol dependence. *Psychopharmacology*, 171(3):340–348, 2004.

[335] G. C. Pimentel and A. L. McClellan. Hydrogen bonding. *Annual Review of Physical Chemistry*, 22(1):347–385, 1971.

[336] Leigh A. Plesniak, Gregory P. Connelly, Lawrence P. Mcintosh, and Warren W. Wakarchuk. Characterization of a buried neutral histidine residue in *Bacillus circulans* xylanase: NMR assignments, pH titration, and hydrogen exchange. *Protein Science*, 5(11):2319–2328, 1996.

[337] Ekaterina V. Pletneva, Alain T. Laederach, D. Bruce Fulton, and Nenad M. Kostic. The role of cation-π interactions in biomolecular association. design of peptides favoring interactions between cationic and aromatic amino acid side chains. *Journal of the American Chemical Society*, 123(26):6232–6245, 2001.

[338] Jay W. Ponder and Frederic M. Richards. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193:775–791, Feb 1987.

[339] P.K. Ponnuswamy, M. Prabhakaran, and P. Manavalan. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 623(2):301 – 316, 1980.

[340] Paul Popelier. *Atoms in Molecules.* Prentice-Hall, 2000.

[341] S. B. Prusiner. Prions. *Proceedings of the National Academy of Sciences, USA*, 95(23):13363–13383, Nov 1998.

[342] Jiang Qian, Nicholas M. Luscombe, and Mark Gerstein. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model,. *Journal of Molecular Biology*, 313(4):673 – 681, 2001.

[343] Sergei Radaev and Peter Sun. Recognition of immunoglobulins by Fcγ receptors. *Molecular Immunology*, 38(14):1073–1083, 2002.

[344] Deepa Rajamani, Spencer Thiel, Sandor Vajda, and Carlos J. Camacho. Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences, USA*, 101:11287–11292, 2004.

[345] J. J. Ramsden. Review of optical methods to probe protein adsorption at solid-liquid interfaces. *J. Stat. Phys.*, 73:853–877, 1993.

[346] Steven W. Rick and R. E. Cachau. The nonplanarity of the peptide group: Molecular dynamics simulations with a polarizable two-state model for the peptide bond. *Journal of Chemical Physics*, 112(11):5230–5241, 2000.

[347] R. Riek, G. Wider, M. Billeter, S. Hornemann, R. Glockshuber, and K. Wuthrich. Prion protein NMR structure and familial human spongiform encephalopathies. *Proceedings of the National Academy of Sciences, USA*, 95(20):11667–11672, Sep 1998.

[348] Dagmar Ringe. What makes a binding site a binding site? *Curr. Opin. Struct. Biol.*, 5:825–829, 1995.

[349] G. G. Roberts, editor. *Langmuir-Blodgett Films.* Plenum Press, New York, 1990.

[350] John Ròmer, Thomas H. Bugge, Leif R. Lund, Matthew J. Flick, Jay L. Degen, and Keld Danò. Impaired wound healing in mice with a disrupted plasminogen gene. *Nature Medicine*, 2(3):287–292, 1996.

[351] Marianne Rooman, Jacky Liévin, Eric Buisine, and René Wintjens. Cation–π/H-bond stair motifs at protein–DNA interfaces. *Journal of Molecular Biology*, 319:6776, 2002.

[352] Lukasz Salwinski, Christopher S. Miller, Adam J. Smith, Frank K. Pettit, James U. Bowie, and David Eisenberg. The database of interacting proteins: 2004 update. *Nucl. Acids Res.*, 32(database issue):D449–451, 2004.

[353] S. Samsonov, J. Teyra, and M.T. Pisabarro. A molecular dynamics approach to study the importance of solvent in protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 73(2), 2008.

[354] Lars Sandberg, R. Casemyr, and Olle Edholm. Calculated hydration free energies of small organic molecules using a nonlinear dielectric continuum model. *J. Phys. Chem. B*, 106(32):7889–7897, 2002.

[355] Lars Sandberg and Olle Edholm. Nonlinear response effects in continuum models of the hydration of ions. *Journal of Chemical Physics*, 116:2936–2944, 2002.

[356] R. T. Sanderson. Electronegativity and bond energy. *Journal of the American Chemical Society*, 105:2259–2261, 1983.

[357] J.L.R. Santos, R. Aparicio, I. Joekes, J.L. Silva, J.A.C. Bispo, and C.F.S. Bonafe. Different urea stoichiometries between the dissociation and denaturation of tobacco mosaic virus as probed by hydrostatic pressure. *Biophysical Chemistry*, 134(3):214–224, 2008.

[358] Aleister J. Saunders, Gregory B. Young, and Gary J. Pielak. Polarity of disulfide bonds. *Protein Science*, 2(7):1183–1184, 1993.

[359] Lindsay Sawyer and Michael N. G. James. Carboxyl-carboxylate interactions in proteins. *Nature*, 295:79–80, Jan 1982. 10.1038/295079a0.

[360] John A. Schellman. Fifty years of solvent denaturation. *Biophysical Chemistry*, 96(2-3):91 – 101, 2002.

[361] J M Scholtz, D Barrick, E J York, J M Stewart, and R L Baldwin. Urea unfolding of peptide helices as a model for interpreting protein unfolding. *Proceedings of the National Academy of Sciences, USA*, 92(1):185–189, 1995.

[362] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucl. Acids Res.*, 32(database issue):D431–433, 2004.

[363] Laurent Schwartz. *Théorie des distributions*. Hermann, Paris, 1966.

[364] L. Ridgway Scott. Nonstandard dielectric response. Research Report UC/CS TR-2010-06, Dept. Comp. Sci., Univ. Chicago, 2010.

[365] L. Ridgway Scott, Mercedes Boland, Kristina Rogale, and Ariel Fernández. Continuum equations for dielectric response to macro-molecular assemblies at the nano scale. *Journal of Physics A: Math. Gen.*, 37:9791–9803, 2004.

[366] R. Scott. Finite element convergence for singular data. *Numer. Math.*, 21:317–327, 1973.

[367] R. Scott. Optimal L^∞ estimates for the finite element method on irregular meshes. *Math. Comp.*, 30:681–697, 1976.

[368] J. Seelig and A. Seelig. Lipid conformation in model membranes and biological membranes. *Q. Rev Biophys*, 13(1):19–61, Feb 1980.

[369] Z. Shi, C. A. Olson, and N. R. Kallenbach. Cation-π interaction in model α-helical peptides. *Journal of the American Chemical Society*, 124(13):3284–3291, 2002.

[370] Zhengshuang Shi, Kang Chen, Zhigang Liu, Tobin R. Sosnick, and Neville R. Kallenbach. PII structure in the model peptides for unfolded proteins – studies on ubiquitin fragments and several alanine-rich peptides containing QQQ, SSS, FFF, and VVV. *Proteins - Structure, Function, and Bioinformatics*, 63(2):312–321, 2006.

[371] Kiyokazu SHIMOMURA, Tomiko FUKUSHIMA, Tamotsu DANNO, Kazuo MATSUMOTO, Munetsugu MIYOSHI, and Yoshio KOWA. Inhibition of intestinal absorption of phenylalanine by phenylalaninol. *J Biochem*, 78(2):269–275, 1975.

[372] H. Shull and G. G. Hall. Atomic units. *Nature*, 184:1559–1560, Nov 1959. 10.1038/1841559a0.

[373] U. C. Singh and Peter A. Kollman. A water dimer potential based on *ab initio* calculations using Morokuma component analyses. *Journal of Chemical Physics*, 83(8):4033–4040, 1985.

[374] N. Sinha, S. Mohan, C.A. Lipschultz, and S.J. Smith-Gill. Differences in electrostatic properties at antibody-antigen binding sites: implications for specificity and cross-reactivity. *Biophysical journal*, 83(6):2946–2968, 2002.

[375] O. Sipahioglu, S. A. Barringer, I. Taub, , and A. P. P. Yang. Characterization and modeling of dielectric properties of turkey meat. *Journal of Food Science*, 68(2):521–527, 2003.

[376] G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, 12:28–35, 2002.

[377] Jared D. Smith, Christopher D. Cappa, Kevin R. Wilson, Benjamin M. Messer, Ronald C. Cohen, and Richard J. Saykally. Energetics of hydrogen bond network rearrangements in liquid water. *Science*, 306:851–853, 2004.

[378] Peter Sondermann, Robert Huber, Vaughan Oosthuizen, and Uwe Jacob. The 3.2-Å crystal structure of the human IgG1 Fc fragment–FcγRIII complex. *Nature*, 406:267–273, 2000.

[379] B. J. Stapley and T. P. Creamer. A survey of left-handed polyproline II helices. *Protein Sci.*, 8:587–595, Mar 1999.

[380] Thomas Steiner. Competition of hydrogen-bond acceptors for the strong carboxyl donor. *Acta Crystallographica Section B*, 57(1):103–106, Feb 2001.

[381] Thomas Steiner and Gertraud Koellner. Hydrogen bonds with π-acceptors in proteins: frequencies and role in stabilizing local 3D structures. *Journal of Molecular Biology*, 305(3):535 – 557, 2001.

[382] DF Stickle, LG Presta, KA Dill, and GD Rose. Hydrogen bonding in globular proteins. *Journal of molecular biology*, 226(4):1143–1159, 1992.

[383] A. J. Stone. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters*, 83(2):233 – 239, 1981.

[384] A. J. Stone. *The theory of intermolecular forces*. Oxford University Press, USA, 1997.

[385] A. J. Stone and M. Alderton. Distributed multipole analysis: methods and applications. *Molecular Physics*, 56(5):1047–1064, 1985.

[386] Kulandavelu Subramanian, Srinivasakannan Lakshmi, Krishnamurthy Rajagopalan, Gertraud Koellner, and Thomas Steiner. Cooperative hydrogen bond cycles involving O—H···π and C—H···O hydrogen bonds as found in a hydrated dialkyne. *Journal of Molecular Structure*, 384(2-3):121 – 126, 1996.

[387] P. A. Suci and G. G. Geesey. Comparison of adsorption behavior of two mytilus edulis foot proteins on three surfaces. *Colloids Surf. B Biointerfaces*, 22:159–168, 2001.

[388] S. Sudarsanam and S. Srinivasan. Sequence-dependent conformational sampling using a database of ϕ_{i+1} and ψ_i angles for predicting polypeptide backbone conformations. *Protein Eng.*, 10(10):1155–1162, 1997.

[389] Monica Sundd, Nicole Iverson, Beatriz Ibarra-Molero, Jose M.Sanchez-Ruiz, and Andrew D.Robertson. Electrostatic interactions in ubiquitin: Stabilization of carboxylates by lysine amino groups. *Biochemistry*, pages 7586–7596, 2002.

[390] Attila Szabo and Neil S. Ostlund, editors. *Modern Quantum Chemistry*. Dover, 1996.

[391] I. Szalai, S. Nagy, and S. Dietrich. Nonlinear dielectric effect of dipolar fluids. *Journal of Chemical Physics*, 131:154905, 2009.

[392] Charles Tanford. *Hydrophobic Effect*. Wiley, 1973.

[393] Charles Tanford and Jacqueline Reynolds. *Nature's Robots, a history of proteins*. Oxford Univ. Press, 2001.

[394] Robin Taylor and Olga Kennard. Hydrogen-bond geometry in organic crystals. *Accounts of Chemical Research*, 17(9):320–326, 1984.

[395] Richard L. Thurlkill, Gerald R. Grimsley, J. Martin Scholtz, and C. Nick Pace. Hydrogen bonding markedly reduces the pK of buried carboxyl groups in proteins. *Journal of Molecular Biology*, 362(3):594–604, 2006.

[396] Richard L. Thurlkill, Gerald R. Grimsley, J. Martin Scholtz, and C. Nick Pace. pK_a values of the ionizable groups of proteins. *Protein Science*, 15(5):1214–1218, 2006.

[397] Ian J. Tickle. Experimental determination of optimal root-mean-square deviations of macromolecular bond lengths and angles from their restrained ideal values. *Acta Crystallographica Section D*, 63(12):1274–1281, Dec 2007.

[398] T. Tokushima, Y. Harada, O. Takahashi, Y. Senba, H. Ohashi, L.G.M. Pettersson, A. Nilsson, and S. Shin. High resolution X-ray emission spectroscopy of liquid water: The observation of two structural motifs. *Chemical Physics Letters*, 460(4-6):387 – 400, 2008.

[399] Chung-Jung Tsai, S. L. Lin, Haim J. Wolfson, and Ruth Nussinov. Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect. *Protein Science*, 6:53–64, 1997.

[400] Jerry Tsai, Robin Taylor, Cyrus Chothia, and Mark Gerstein. The packing density in proteins: Standard radii and volumes. *Journal of Molecular Biology*, 290:253–266, 1999.

[401] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36(5):823–841, Sep 1930.

[402] P.T. Van Duijnen and M. Swart. Molecular and atomic polarizabilities: Thole's model revisited. *J. Phys. Chem. A*, 102(14):2399–2407, 1998.

[403] Michal Vieth, Richard E. Higgs, Daniel H. Robertson, Michael Shapiro, Ellen A. Gragg, and Horst Hemmerle. Kinomics–structural biology and chemogenomics of kinase inhibitors and targets. *Biochimica et Biophysica Acta*, 1697:243–257, 2004.

[404] Jorge A. Vila, Daniel R. Ripoll, Myriam E. Villegas, Yury N. Vorobjev, and Harold A. Scheraga. Role of hydrophobicity and solvent-mediated charge-charge interactions in stabilizing alpha-helices. *Biophysical Journal*, 75:2637–2646, 1998.

[405] D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nature Structural Biology*, 8:559–566, 2001.

[406] Dennis Vitkup, Eugene Melamud, John Moult, and Chris Sander. Completeness in structural genomics. *Nature Structural and Molecular Biology*, 8:559–566, Jun 2001. 10.1038/88640.

[407] D. Voet and J. G. Voet. *Biochemistry*. John Wiley & Sons, New York, 1990.

[408] Christine Vogel, Carlo Berzuini, Matthew Bashton, Julian Gough, and Sarah A. Teichmann. Supra-domains: Evolutionary units larger than single protein domains. *Journal of Molecular Biology*, 336(3):809 – 823, 2004.

[409] J. Von Eichborn, S. Günther, and R. Preissner. Structural features and evolution of protein-protein interactions. In *Genome Inform*, volume 22, pages 1–10, 2010.

[410] Xiaoyuan. Wang, Mikhail. Bogdanov, and William. Dowhan. Topology of polytopic membrane protein subdomains is dictated by membrane phospholipid composition. *EMBO J.*, 21(21):5673–5681, Nov 2002.

[411] A. Warshel and A. Papazyan. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.*, 8:211–217, 1998.

[412] Arieh Warshel. Electrostatic origin of the catalytic power of enzymes and the role of preorganized active sites. *Journal of Biological Chemistry*, 273:27035–27038, 1998.

[413] James D. Watson and Francis H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[414] Gary J. Wedemayer, Phillip A. Patten, Leo H. Wang, Peter G. Schultz, and Raymond C. Stevens. Structural insights into the evolution of an antibody combining site. *Science*, 276(5319):1665–1669, 1997.

[415] S. Weggler, V. Rutka, and A. Hildebrandt. A new numerical method for nonlocal electrostatics in biomolecular simulation. *Journal of Computational Physics*, 229(11):4059–4074, 2010.

[416] W. Ren Weinan E and E. Vanden-Eijnden. A general strategy for designing seamless multiscale methods. *Journal of Computational Physics*, 1(1):1–1, 2010.

[417] Ph. Wernet, D. Nordlund, U. Bergmann, M. Cavalleri, M. Odelius, H. Ogasawara, L.Å. Näslund, T. K. Hirsch, L. Ojamäe, P. Glatzel, L. G. M. Pettersson, and A. Nilsson. The structure of the first coordination shell in liquid water. *Science*, 304(5673):995–999, 2004.

[418] Maurice Wilkins, A. R. Stokes, and H. R. Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171:738–740, 1953.

[419] Robert W. Williams, Albert Chang, Davor Juretic, and Sheila Loughran. Secondary structure predictions and medium range interactions. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, 916(2):200 – 204, 1987.

[420] René Wintjens, Jacky Liévin, Marianne Rooman, and Eric Buisine. Contribution of cation-π interactions to the stability of protein-DNA complexes. *Journal of Molecular Biology*, 302:394–410, 2000.

[421] S. T. Wlodek, T. W. Clark, L. R. Scott, and J. A. McCammon. Molecular dynamics of acetylcholinesterase dimer complexed with tacrine. *J. Am. Chem. Soc.*, 119:9513–9522, 1997.

[422] Stanislaw T. Wlodek, Tongye Shen, and J. A. McCammon. Electrostatic steering of substrate to acetylcholinesterase: Analysis of field fluctuations. *Biopolymers*, 53(3):265–271, 2000.

[423] Gerd Wohlfahrt. Analysis of pH-dependent elements in proteins: Geometry and properties of pairs of hydrogen-bonded carboxylic acid side-chains. *Proteins: Structure, Function, and Bioinformatics*, 58:396–406, 2005.

[424] Yuri I. Wolf, Steven E. Brenner, Paul A. Bash, and Eugene V. Koonin. Distribution of protein folds in the three superkingdoms of life. *Genome Res*, 9(1):17–26, Jan 1999.

[425] Yuri I. Wolf, Nick V. Grishin, and Eugene V. Koonin. Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, 299(4):897 – 905, 2000.

[426] J. Wong. Coevolution theory of the genetic code at age thirty. *BioEssays*, 27(4):416–425, 2005.

[427] J.M. Word, S.C. Lovell, J.S. Richardson, and D.C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *Journal of molecular biology*, 285(4):1735–1747, 1999.

[428] Chia Kuei Wu, Bing Hu, John P. Rose, Zhi-Jie Liu, Tam L. Nguyen, Changsheng Zheng, Esther Breslow, and Bi-Cheng Wang. Structures of an unliganded neurophysin and its vasopressin complex: Implications for binding and allosteric mechanisms. *Protein Science*, 10(9):1869–1880, 2001.

[429] Stefan Wuchty. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18(9):1694–1702, Sep 2001.

[430] Dong Xu, Chung-Jung Tsai, and Ruth Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering*, 10(9):999–1012, 1997.

[431] Wai-Ming Yau, William C. Wimley, Klaus Gawrisc, and Stephen H. White. The preference of tryptophan for membrane interfaces. *Biochemistry*, 37:14713–14718, 1998.

[432] Hitomi Yuki, Yoshikazu Tanaka, Masayuki Hata, Hidenori Ishikawa, Saburo Neya, and Tyuji Hoshino. Implementation of π-π interactions in molecular dynamics simulation. *Journal of Computational Chemistry*, 28(6):1091–1099, 2007.

[433] R. Zahn, A. Liu, T. Luhrs, R. Riek, C. von Schroetter., F. Lopez Garcia, M. Billeter, L. Calzolai, G. Wider, and K. Wüthrich. NMR solution structure of the human prion protein. *Proceedings of the National Academy of Sciences, USA*, 97(1):145–150, Jan 2000.

[434] A. Zarrine-Afsar, A. Mittermaier, L. E. Kay, and A. R. Davidson. Protein stabilization by specific binding of guanidinium to a functional arginine-binding surface on an SH3 domain. *Protein Sci.*, 15:162–170, Jan 2006.

[435] Wenge Zhong, Justin P. Gallivan, Yinong Zhang, Lintong Li, Henry A. Lester, and Dennis A. Dougherty. From ab initio quantum mechanics to molecular neurobiology: A cation-π binding site in the nicotinic receptor. *Proceedings of the National Academy of Sciences, USA*, 95(21):12088–12093, 1998.

[436] Ruhong Zhou, Xuhui Huang, Claudio J. Margulis, and Bruce J. Berne. Hydrophobic collapse in multidomain protein folding. *Science*, 305(5690):1605–1609, 2004.

[437] Z. H. Zhou, M. L. Baker, W. Jiang, M. Dougherty, J. Jakana, G. Dong, G. Lu, and W. Chiu. Electron cryomicroscopy and bioinformatics suggest protein fold models for rice dwarf virus. *Nature Structural Biology*, 8:868–873, 2001.

[438] Z. H. Zhou, W. Chiu, K. Haskell, H. Spears, J. Jakana, F. J. Rixon, and L. R. Scott. Parallel refinement of herpesvirus B-capsid structure. *Biophysical Journal*, 73:576–588, January 1997.

[439] Z. H. Zhou, M. Dougherty, J. Jakana, J. He, F. J. Rixon, and W. Chiu. Seeing the herpesvirus capsid at 8.5 Å. *Science*, 288:877–880, 2000.

[440] Z. H. Zhou, Sharon J. Macnab, Joanita Jakana, L. R. Scott, W. Chiu, and F. J. Rixon. Identification of the sites of interaction between the scaffold and outer shell in herpes simplex virus-1 capsids by difference electron imaging. *Proceedings of the National Academy of Sciences, USA*, 95:2778–2783, March 1998.

[441] Matjaz Zorko and Ülo Langel. Cell-penetrating peptides: mechanism and kinetics of cargo delivery. *Advanced Drug Delivery Reviews*, 57(4):529–545, 2005.

Index