Digital Biology

L. Ridgway Scott
University of Chicago

Ariel Fernández
Rice University

Release 0.3
DO NOT DISTRIBUTE

February 21, 2011

Contents

Chapter 1

How is biology digital?

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 requirements. 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.

We primarily study proteins and their interactions. These are often involved in signaling and function in a discrete (or digital, or quantized) way. In addition, proteins are discrete 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 chemical/physical mechanisms used are fundamentally continuous.

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 [285] 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 [324]. Resonance theory (Section 13.1) describes general bonding patterns as a combination of simple bonds (e.g, single and double bonds) [323]. The discrete behavior of DNA elucidated by Crick, Franklin, Watson, Wilkins and others [153, 421, 424] 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

[1] A ligand is anything that binds to something. We provide a glossary of terms like this in Chapter 25 rather than defining them in the text.

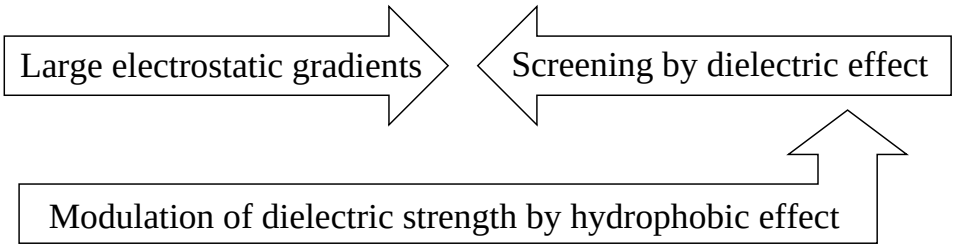


Figure 1.1: Three competing effects that determine protein behavior.

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 [134, 136, 141]. 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. 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 [95]. 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.

This book is not a typical introduction to a well developed field in which all the main results are already standard. Rather it is intended to stimulate 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 [285], "And thus the period of exploration begins."

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.

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 14.2.1, 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 [324] 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 (see Chapter 18). 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, 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.

Atom	Symbol	+/−	neutrons	outer	lacking	mass	radius
Hydrogen	H	1	0 (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	15.9994	1.52
Fluorine	F	9	10	7	1	18.998	1.47
Sodium	Na	11	12	1	7	22.9898	2.27
Magnesium	Mg	12	12 (13,14)	2	6	24.305	1.73
Phosphorus	P	15	16	5	3	30.974	1.80
Sulfur	S	16	16 (17,18,20)	6	2	32.065	1.80
Chlorine	Cl	17	18 (20)	7	1	35.45	1.75
Potassium	K	19	20 (22)	1	17	39.098	2.75
Calcium	Ca	20	20 (22-24)	2	16	40.08	2.00
Iron	Fe	26	30 (28,31,32)	8	10	55.845	1.10*
Copper	Cu	29	34 (36)	11	7	63.55	1.40*
Zinc	Zn	30	34 (36-38,40)	12	6	64.4	1.39*
Selenium	Se	34	46 (40,42-44)	16	2	78.96	1.90
Iodine	I	53	74	17	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 'mass' give the atomic mass in Daltons (see Chapter 14 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.

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.

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 (see Chapter 18).

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 5.3 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 atoms 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.

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 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 4.8).

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. 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.

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 Solvated struggle

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 [316], but proteins must struggle to live with it [121, 248].

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 [154], 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 [1, 382, 423], 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 $n = 1, 2$ 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 Biological 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 [139]. 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

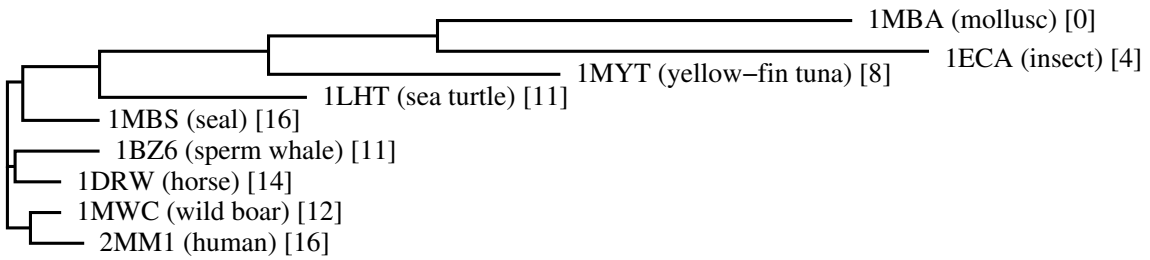


Figure 2.1: Number of dehydrons [shown in square brackets] in the protein myoglobin found in various species [139], which are presented in an evolutionary tree determined by sequence alignment distances.

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 [235]. Indeed, the original study of the structure of the peptide bond (see section 8-4 of [323]) 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 [127]. 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 2.1 depicts the number of dehydrons found in the protein myoglobin (or its analog) in various species [139]. Similar trends are seen with other proteins in Table 11.4.

The number of under-wrapped hydrogen bonds appears to be evolving (increasingly), providing increasing opportunities for interaction in advanced species. This provides additional understanding 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 [139].

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 [195] 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 [142] 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 [253]. 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 [399], 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 - looking for patterns where one event is connected to another event
- Sequence or path analysis - looking for patterns where one event leads to another later event
- Classification - looking for new patterns (May result in a change in the way the data is organized but that's ok)
- Clustering - finding and visually documenting groups of facts not previously known

Our conclusion? Data mining involves looking at data. 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 [399], 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 [204].

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 [97]. 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 [235] shows clearly both the radial and the angular dependence of the hydrogen bond. Similar evidence is found in later papers; Figure 3 in [400] 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 [28]. Figure 8 of [437] 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 [299]. 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 [306, 377]. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations has also been reported [231]. 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 will not make significant use of such metrics, but we review in Section **??** some possible approaches.

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

$$\sin(\log(\tan(\cos(J_1(e^x + \sqrt{\pi x}))))). \tag{2.3}$$

Unfortunately, the classical paradigm is limited by our ability to absorb new names. While the names in (2.3) 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

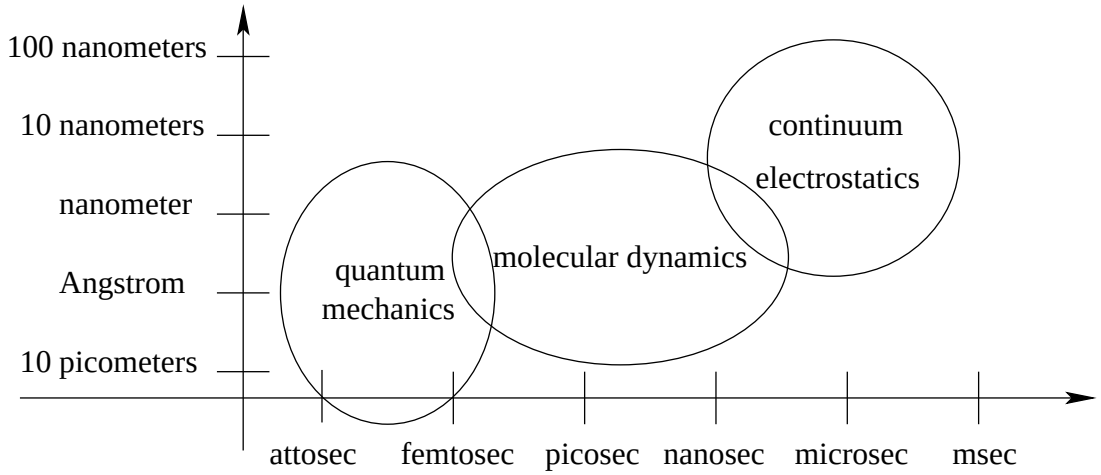


Figure 2.2: Cartoon of spatial and temporal scales of biomolecular models. See the text for more details.

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 2.2. 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 [190]. Scale separation often simplifies the interactions among the different scales. The differences often occur in 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 2.2. The smallest and fastest scale is that of quantum chemistry (Chapter 18). 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 [89, 371]. 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 2.2 the time scale for so-called Brownian dynamics models which are based on a continuum dielectric model [268].

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 2.2 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 [120]. 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 [9]. Electrostatic models hope to capture the expected impact of dielectric

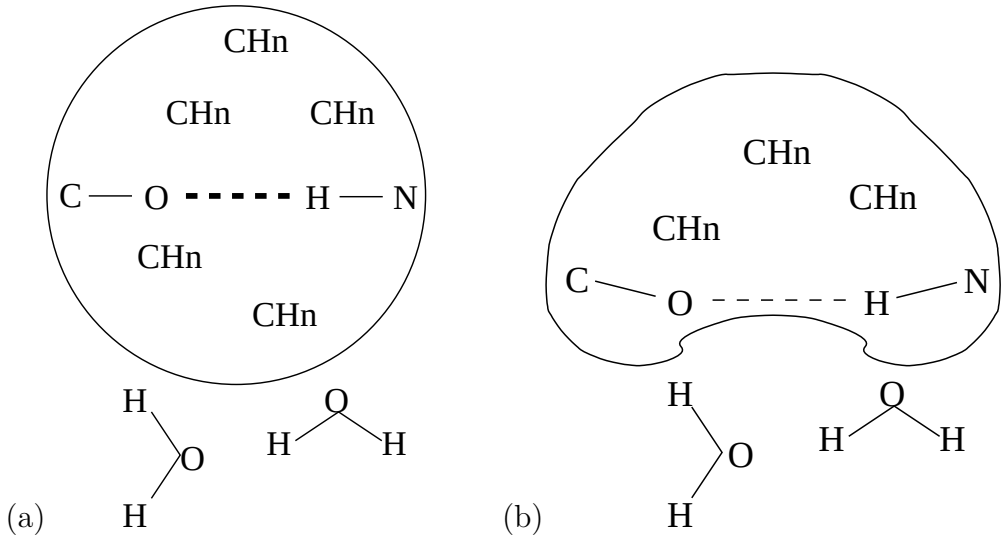


Figure 2.3: (a) Well wrapped hydrogen bond (b) Underwrapped hydrogen bond.

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 [371].

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 [98, 422].

2.5 Hydrophobic interactions

Hydrophobic interactions are sometimes said to be more important than even the hydrogen bond [214]. Although not completely understood, the **hydrophobic force** [42] derives from the **hydrophobic effect** [398]. This effect is one of the central topics of our study. However, the hydrophobic effect has many manifestations in protein behavior.

The hydrophobic effect [42, 398] was proposed in the 1950's [399] 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 [94, 95]. It was seen that the role of hydrophobic modulation of solvent dielectric is critical to the hydrophobic force [94].

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 5.3) 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 [94, 95]. These arguments are compelling, but they suggest a nonspecific interaction. Indeed, hydrophobic attraction leads to nonspecific binding [141].

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 [121].

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 [120].

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 [141].

More generally, solvent mediation can alter any electrostatic force via dielectric effects (Chapter 19). 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 19), 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.

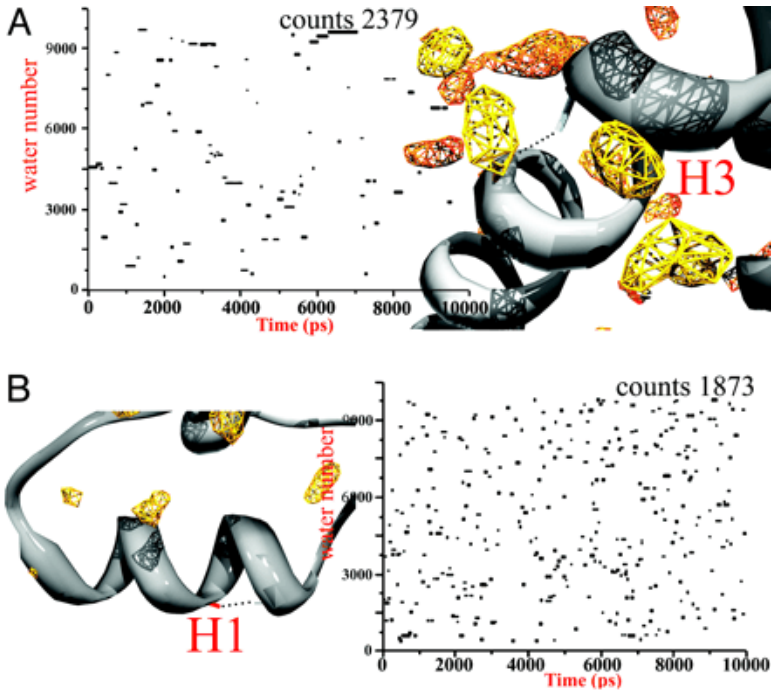


Figure 2.4: Dynamics of water near hydrogen bonds, reproduced from Fig. 5 in [86]. (A) Hydrogen bond (H3) is well wrapped. (B) Hydrogen bond (H1) is underwrapped.

2.5.2 Dehydrons

In [139], 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 [42, 398], and its role in modulating the dielectric effect (Chapter 19). 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 [137]. The concept of wrapping of a hydrogen bond is depicted informally in Figure 2.3. A well wrapped hydrogen bond, Figure 2.3(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 2.3(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

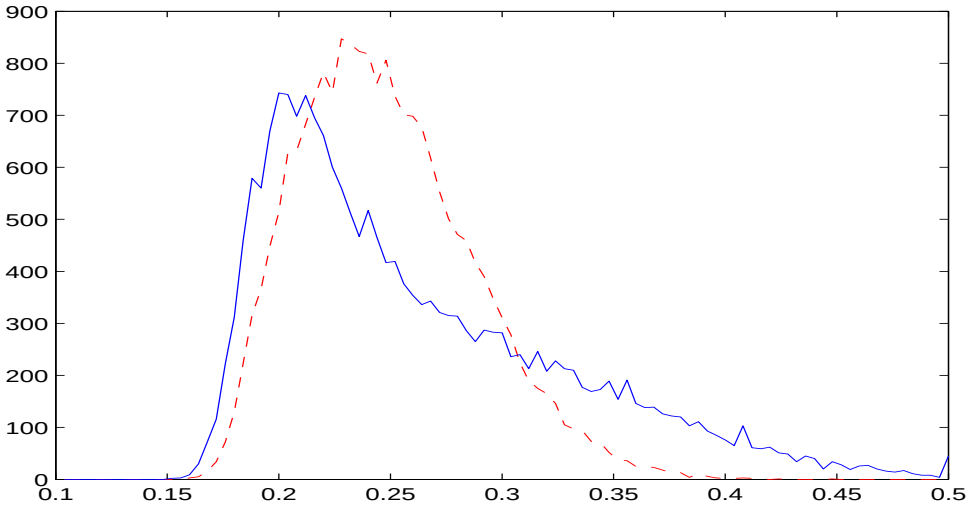


Figure 2.5: Distribution of bond lengths for two hydrogen bonds formed in a structure of the sheep prion [86]. The horizontal axis is measured in nanometers, whereas the vertical axis represents numbers of occurrences taken from a simulation with 20,000 data points with bin widths of 0.1 Ångstrom. The 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.

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 [139] of the wrappers (Section 8.6).

We have already seen in Figure 2.1 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 [139] and in a cartoon in Figure 11.1.

2.5.3 Dynamics of dehydrons

The extent of wrapping changes the nature of hydrogen bond [86] and the structure of nearby water [129]. Hydrogen bonds that are not protected from water do not persist [86]. Figure 5 of [86] shows the striking difference of water residence times for well wrapped and underwrapped hydrogen bonds. Private communication with the authors of [86] 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 2.5 are quite different, confirming our prediction based on Figure 2.3 that

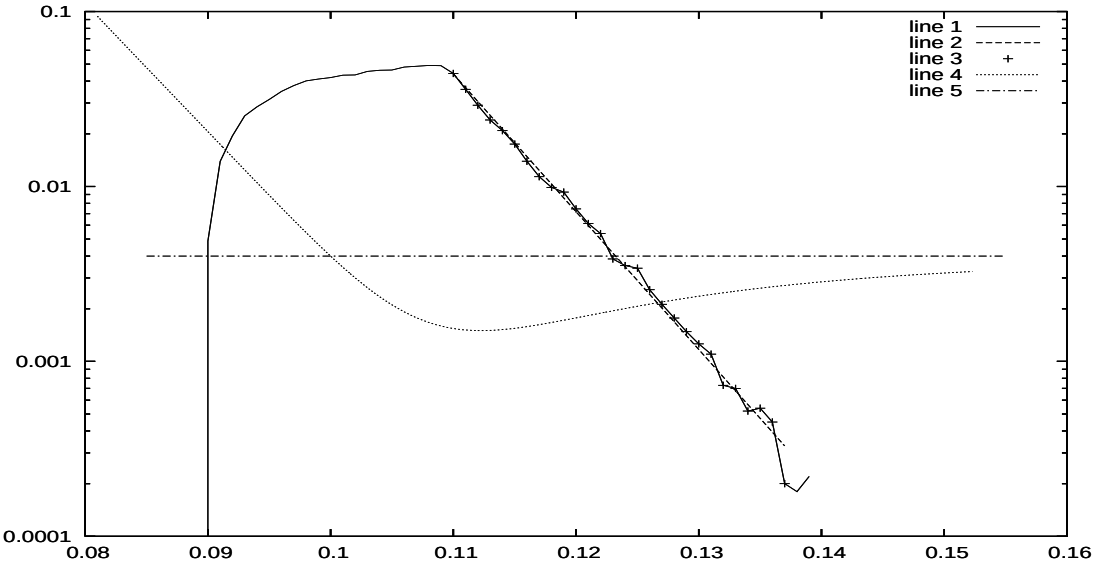


Figure 2.6: Simulation of a random walk with a restoring force. Shown is the distribution of values x_i defined in (2.4) for 10^5 time steps i, starting with $x_1 = 0.1$, scaled by a factor of 10^{-3}. Also shown is a graph of $\phi + 0.03$ where ϕ is the potential (2.4). The dot-dashed horizontal line 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.

the coupling of the hydrogen bond characteristics with the water environment would be different.

The H-bond R208–E212 depicted in Fig. 5(A) [86] is well wrapped whereas V189–T193 depicted in Fig. 5(B) is a dehydron (see Fig 3a in [131] page 6448). Well-wrapped hydrogen bonds are visited by fewer water molecules but have longer-lasting water interactions (due to the structuring effect of the hydrophobes), whereas the behavior of dehydrons is more like that of bulk water: frequent re-bonding with different water molecules [86].

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 2.3(b) when water is near, to one more like that depicted in Figure 2.3(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 2.5. It is natural to expect 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 5.11 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 2.5 can be interpreted via a simulation which is depicted in Figure 2.6. This figure records the distribution of positions for a random walk subject to a restoring force defined by

$$x_{i+1} = x_i + \Delta t(f_i + \phi(x_i)) \tag{2.4}$$

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

$$\phi(x) = (0.1/x)^{12} - (0.1/x)^6. \tag{2.5}$$

The particular time step used in Figure 2.6 is $\Delta t = 0.02$; the simulation was initiated with $x_1 = 0.1$ and carried out for 10^5 steps.

The simulation (2.4) 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 (2.5). Such a system exhibits a distribution with an exponential decay, as verified in Figure 2.6 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 2.7. 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 force of attraction for a dehydron. This is explained in more detail in Chapter 9.

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

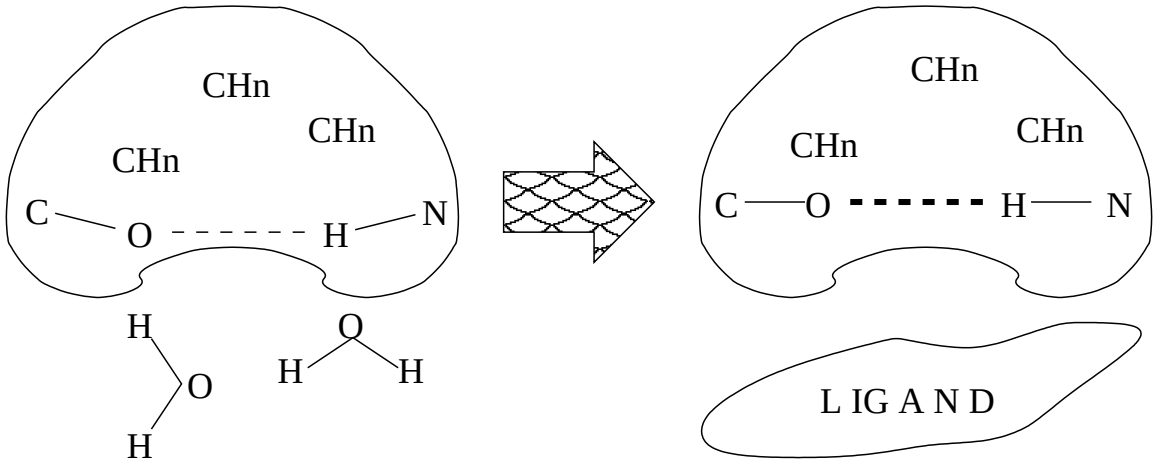


Figure 2.7: Cartoon showing dehydration due to ligand binding and the resulting strengthening of an underwrapped hydrogen bond.

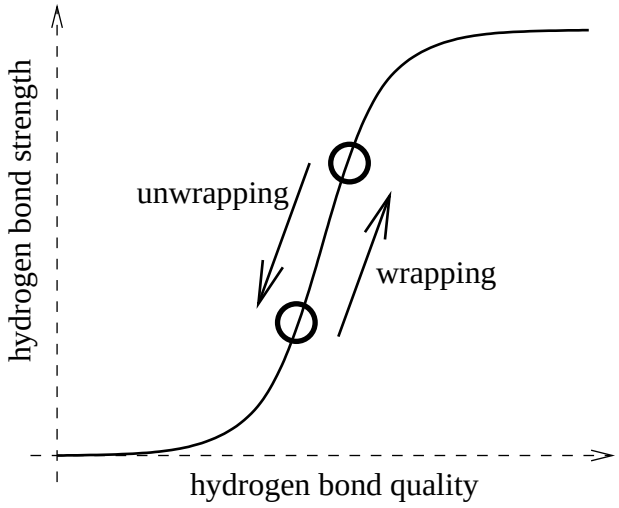


Figure 2.8: Cartoon depicting the capability of a dehydron to switch on and off when a ligand binds. The binding event may increase wrapping, in which case the hydrogen bond becomes stronger and more stable, or it may decrease wrapping, in which case the hydrogen bond becomes weaker and less stable.

the strength of the hydrogen bond depends in some way on this quality variable, as depicted in Figure 2.8.

Binding events can cause the quality of a hydrogen bond to change significantly, due to several factors. Large scale structural changes can occur which might alter the distances and angles that define 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. Wrapping can increase the quality, and unwrapping can decrease the quality. 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.

2.7 Entropy and thermodynamics

Certain aspects of protein biology are governed by the laws of thermodynamics. This reflects the fact that all molecules are in constant motion due to the thermal fluctuations that persist above absolute zero temperature. At physiological temperatures, these fluctuations are a significant part of the energy budget of biological systems. We do not attempt an introduction to this subject, and instead suggest consulting suitable texts that we list in Chapter 20. However, we do need to understand one simple principle related to the concept of entropy.

We will consider systems that are described by entities (e.g., atoms, or charge groups) at positions $x_i \in \mathbb{R}^3$, $i = 1, \ldots, n$, whose potential energy is given by a function $V(x_1, \ldots, x_n)$. For example, this could reflect the attraction of positive and negative electric charges, in which terms of the form $|x_i - x_j|^{-1}$ will contribute to V. Or it could involve more subtle forms of attraction that decay like $|x_i - x_j|^{-6}$. In any case, it will involve a sum of such terms. The key point is that the energy will be distinctly changed if some of the terms x_i get removed from the system. We will see that this can effectively happen when some external operation is applied to the system. In particular, opportunities to make bonds can get removed by the introduction of a non-interacting surface. A key example is water and hydrophobic surfaces.

It is not hard to understand how the restriction of degrees of freedom in a system can change the energy. Consider going on a hike with a backpack with several items dangling from the bottom. If these are small, they may just make an annoying noise. But if they are heavier, their motion may be a serious energy drain, as you must exert extra force to counterbalance their movements.

Immobilizing them (e.g., by tying them to the pack) can make the hike less tiring. Conversely, adding degrees of freedom can also be beneficial in certain instances. Runners typically use swinging arms as a means to impart extra force in a concerted motion together with the movement of their legs. Thus a change in the number of degrees of freedom can result in a change in the energy of a system. We will see that the concept of **entropy** relates to the number of degrees of freedom that a system can adopt, cf. Chapter 20.

2.8 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 [323].*

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 all pairs of carbonyl and amide groups in the peptide bonds (cf. Figure 4.3). What is the part of the distribution that corresponds to ones forming a hydrogen bond? (Hint: exclude the N's and O's that are near neighbors in the peptide bond backbone.)*

Exercise 2.5 *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 2.6 *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 2.7 *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 $m = 10$.*

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 [323, 324] are still natural references.

The order of forces, or bonds [341], 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

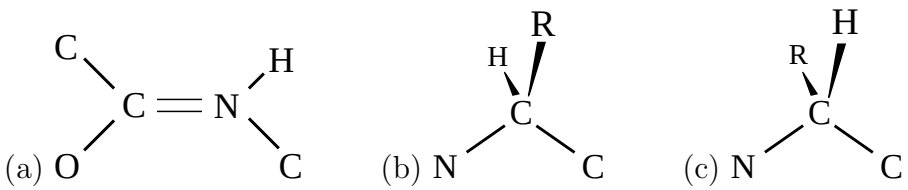


Figure 3.1: Single and double bonds. The double bond (a) often confers a planar geometry to the atoms; all six atoms in (a) are in the plane of the page. The upper-left and lower-right carbons represent C_α carbons in a peptide sequence, and would each have three additional single bonds (not shown). (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 [293].

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 4.4–4.5 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 [323, 324] 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

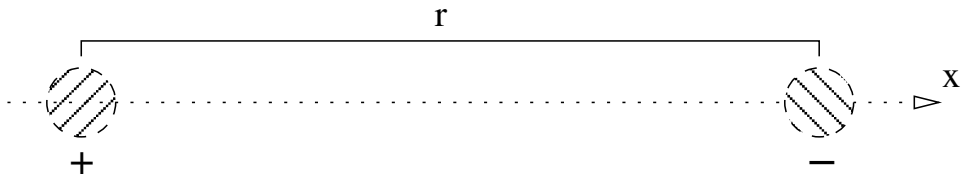


Figure 3.2: Salt bridges are ionic bonds.

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. We address the form of such a force of repulsion in Section 18.7. 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 [399]. The detailed structure of hydrogen bonds in biology is still being investigated [231, 299, 382, 423]. 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 [84, 159, 439]. 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 14) 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:

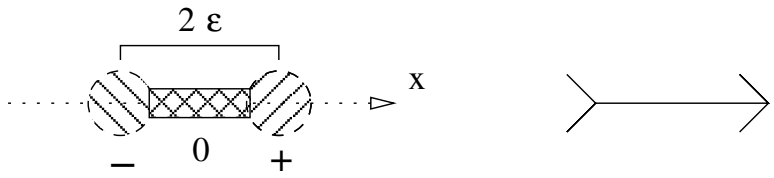


Figure 3.3: Left: configuration of a dipole consisting of a pair of charged molecules at $(\pm\epsilon, 0, 0)$. Right: abstract representation of a dipole as a vector.

$z'\mathbf{e}$. 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. (14.2).

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 14) 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. 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 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.

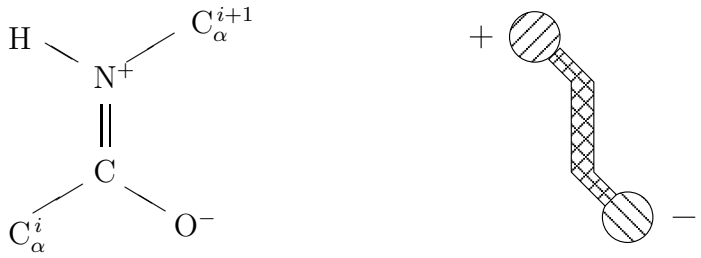


Figure 3.4: The dipole of the peptide backbone (in the 'trans' configuration, cf. Figure 4.3). On the left is the chemical description, and the right shows a cartoon of the dipole. The double bond between the central carbon and nitrogen keeps the peptide bond planar. There is a positive charge center near the hydrogen and a negative charge center near the oxygen.

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 [157]. 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 large r, that is, that

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

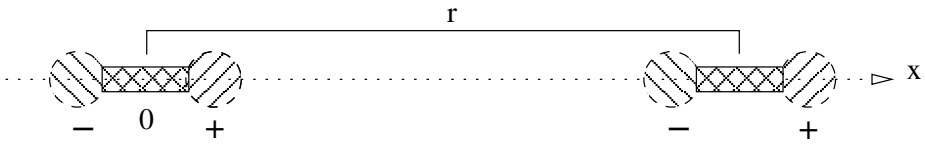


Figure 3.5: Single-file dipole-dipole configuration consisting of two pairs of molecules with charges ± 1 at $(\pm\epsilon, 0, 0)$ and $(r \pm \epsilon, 0, 0)$.

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 $r >> \epsilon$ along the x-axis,

$$\begin{aligned} F_x(r,0,0) &= -(r-\epsilon)^{-2} + (r+\epsilon)^{-2} \\ &= \frac{-(r+\epsilon)^2 + (r-\epsilon)^2}{(r-\epsilon)^2(r+\epsilon)^2} = \frac{-4\epsilon r}{(r-\epsilon)^2(r+\epsilon)^2} \\ &\approx -4\epsilon r^{-3} = -2\mu r^{-3}. \end{aligned} \tag{3.12}$$

The attractive force experienced by a dipole displaced on the x-axis at a distance r, with a

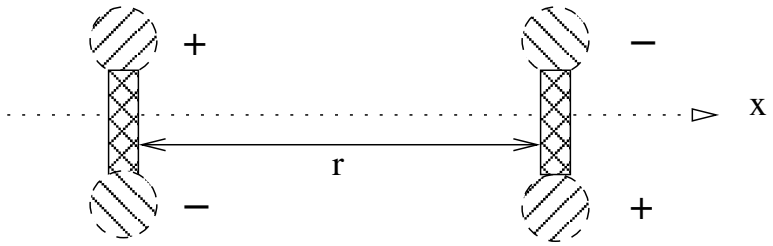


Figure 3.6: Parallel dipole-dipole configuration with ± 1 charges at $(0, \pm\epsilon, 0)$ and $(r, \mp\epsilon, 0)$.

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

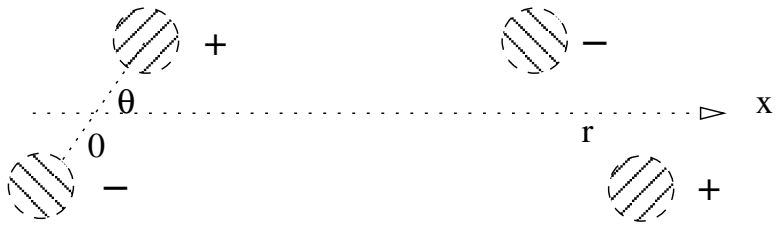


Figure 3.7: General θ-dependent dipole-dipole configuration. The connective material between the charge centers has been omitted for simplicity.

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

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

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 $r >> \epsilon$ 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

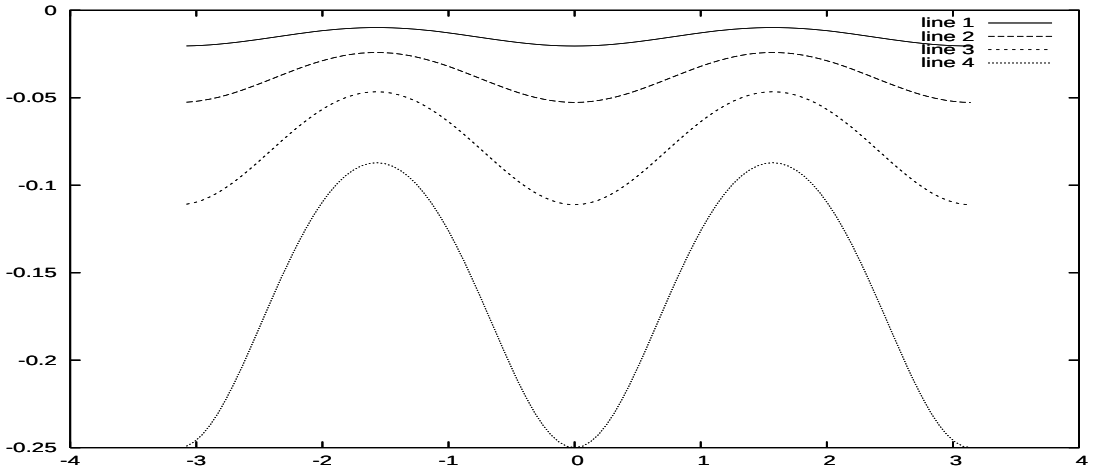


Figure 3.8: Potential energy variation $v(\rho, \theta)$ as defined in (3.23) (vertical axis) of dipoles as a function of θ (horizontal axis) for the configurations shown in Figure 3.7 for $\rho = 0.02$ (top), $0.05, 0.1, 0.2$ (bottom), where ρ is defined in (3.22).

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 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

$$\rho = (2\epsilon/r)^2. \tag{3.22}$$

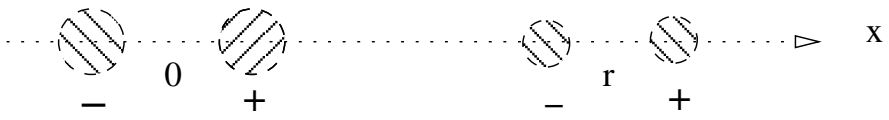


Figure 3.9: Single-file dipole-dipole configuration with different dipole strengths.

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 3.7.

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. 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.

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.

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 3.1: 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 [323] κ are given in kcal/mol, together with the repulsion potential energy $V(\rho)$ at the van der Waals radius ρ.

Debye forces and London dispersion forces [157] 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 [157] which results from both dipoles being induced. This often takes the form of a symmetry breaking, and we give a derivation of the dependence of the magnitude of the induced dipole on distance in Section 18.5.2.

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, but defer to Section 18.5.2 a detailed discussion. The repulsive term r^{-12} is a convenient model, whereas other terms are more accurate [81] (cf. Section 18.7).

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 3.1. Note that $V(\rho/1.2) \approx -3V(\rho)$, and $V(\rho/2) = -3968V(\rho)$, so the repulsion is quite strong in this model.

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 [157]. More subtly, two polarizable molecules can induce dipoles in each other, via what are called London dispersion forces [157].

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 3.2: Experimental [8] and derived [410] values for polarizabilities of some molecules. Units are Å^3.

Figure 3.10: 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.

Essentially all materials are polarizable. This just means that the distributions of electrons can be distorted by an electric field. Table 3.2 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 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. The behavior in (3.27) will be deduced by a perturbation technique for small $\mathbf{F}$ from the concepts in Section 18.5.

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.

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 (to be explained in Section 18.5), 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. We will see in Section 18.5.3 that this value of r can be interpreted in a mathematical, if not physical, sense.

To achieve a correct description of the van der Waals force, we need to use a quantum mechanics description of the atomic interactions (Chapter 18). We will derive in (18.70) a result that confirms

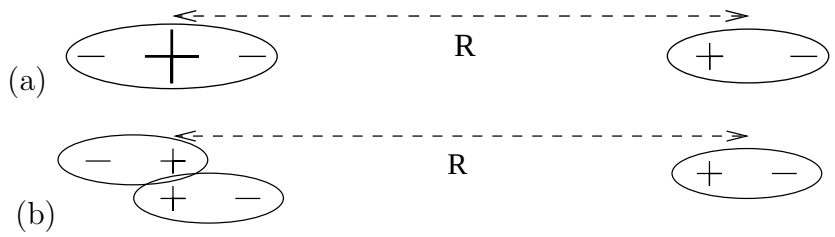


Figure 3.12: (a)Interaction between a dipole (right) and a neutral group (left). The neutrally charged group on the left has charges $+2c$ at the origin and $-c$ located at $(\pm\delta, 0, 0)$, and the dipole on the right has charges ± 1 at $(R \mp \frac{1}{2}\mu, 0, 0)$. (b) Equivalent representation of the interaction between a dipole and a neutral group, in which the neutral group in (a) is written as two dipoles.

the basic dependence of the dipole strength on r, namely

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

where an expression for the constant c_1 will be made explicit. In addition, we will 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 will 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 *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 3.3 *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 3.4 *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 3.5 *Consider the expression in (3.21). Prove that, for any $\rho < 1$, it has a maximum when $\theta = 0$ and a minimum when $\theta = \pi/2$.*

Exercise 3.6 *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.*

Exercise 3.7 *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 [197]. This can be done with one distance parameter and three angular parameters.)*

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

Exercise 3.9 *Show that the in-line 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.

Chapter 4

Protein basics

It is not our intention to provide a complete introduction to the structure of proteins. Instead, we suggest consulting texts [82, 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 4.4—4.5. 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 4.2.

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,

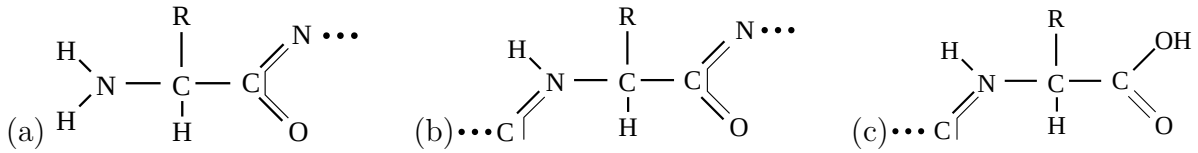


Figure 4.1: The basic units of a peptide sequence. (a) The initial, or N-terminal, unit. (a) The typical (internal) unit. (c) The final, or C-terminal, unit. The letter R stands for 'residue' and can be any of the twenty depicted in Figures 4.4—4.5. The atoms are not all co-planar. In particular, the 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.

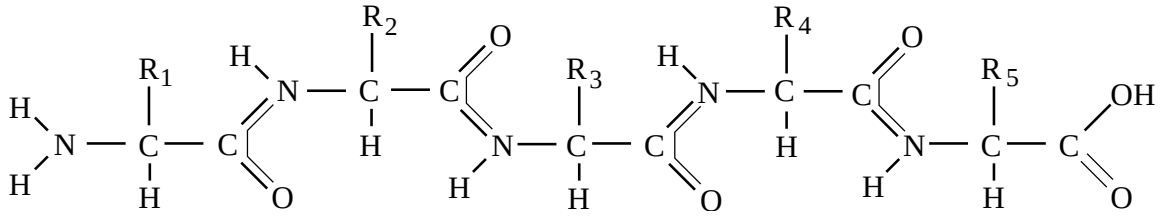


Figure 4.2: Schematic representation of a peptide sequence with five units (all in the trans conformation). The double bonds (double lines) between C and N denote the transition point between one peptide unit and the next. The residues are numbered in increasing order starting at the N-terminal end.

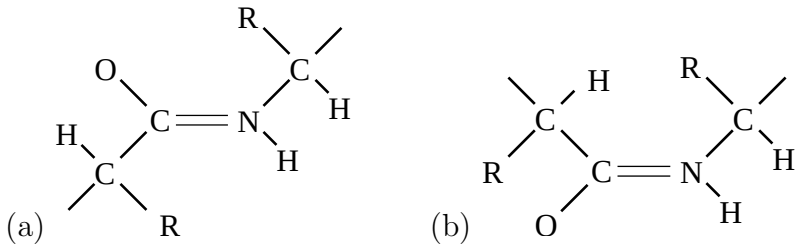


Figure 4.3: The rigid state of the peptide bond: (a) trans form, (b) cis form [39]. The letter R stands for 'residue' and can be any of the twenty depicted in Figures 4.4—4.5. The double bond between the central carbon and nitrogen keeps the peptide bond planar. Compare Figure 13.1.

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 4.3 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, 181, 317]. 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 4.2. This particular bond will be considered in more detail in Chapter 13. In Figure 4.3 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 4.3. 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.

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.

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 descriptions of the sidechains for the amino acids can be found in Figures 4.4—4.5.

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

[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.

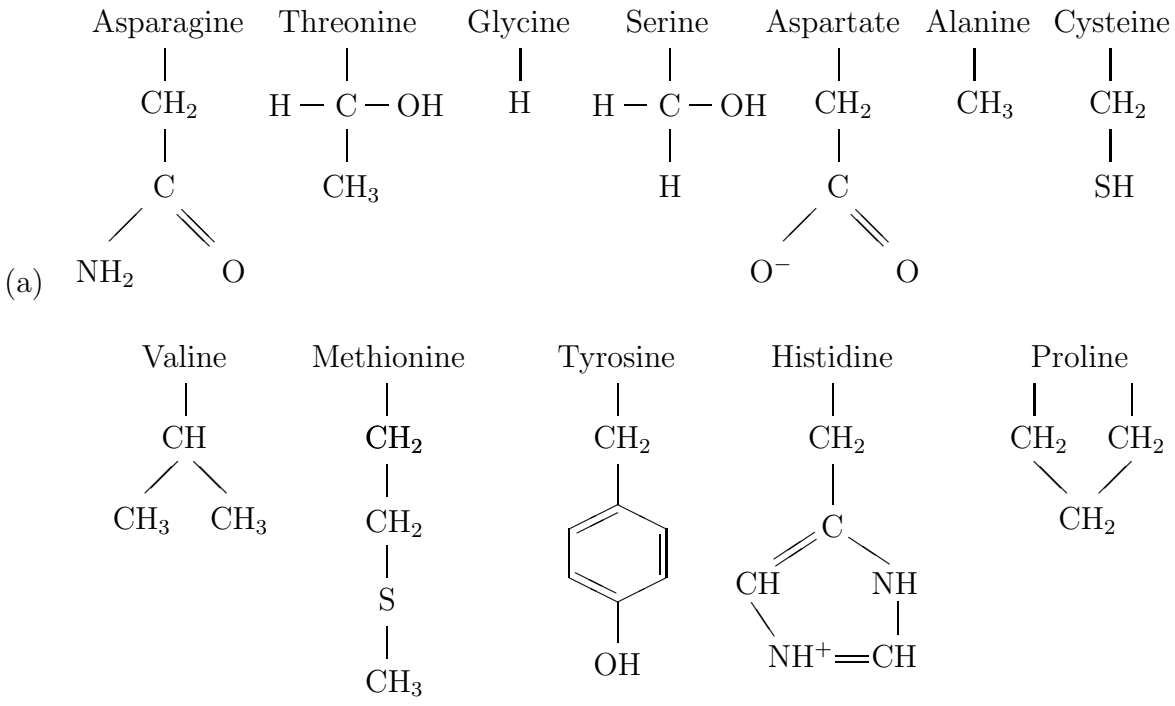


Figure 4.4: Periodic table of amino acid sidechains (residues). Not shown is the C_α carbon (see Figure 4.3), located at the top of the residue where the name appears. (a) The smallest, and most likely to be involved in protein-ligand interactions, ordered from the left (asparagine). (b) The middle ground in terms of interactivity.

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 [42, 398], 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 4.4—4.5). 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.

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.

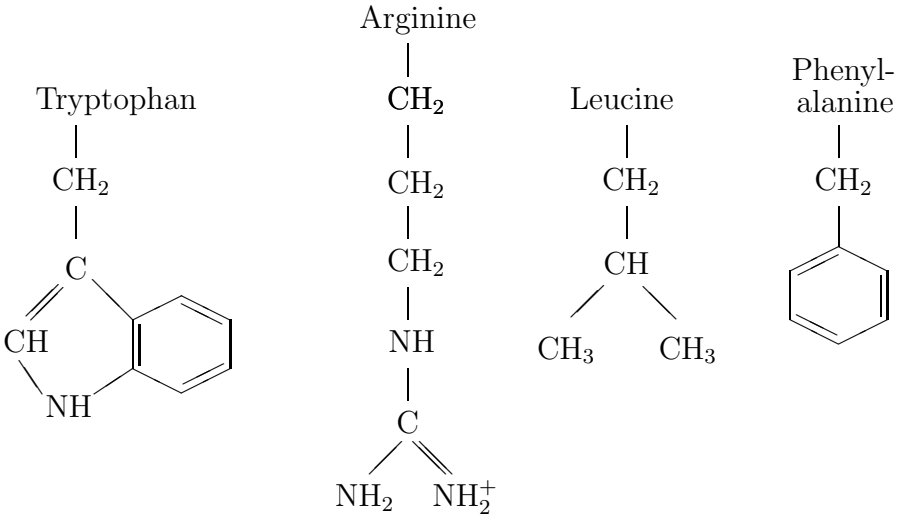


Figure 4.5: Periodic table of amino acid sidechains (residues). Not shown is the C_α carbon (see Figure 4.3), located at the top of the residue where the name appears. (c) Some amino acids less likely to be interactive. (d) The amino acids least likely to be involved in interactions.

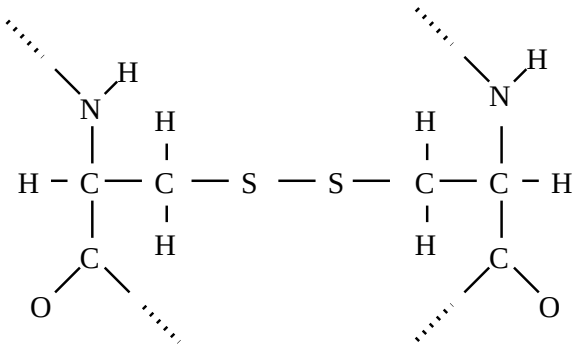


Figure 4.6: Schematic representation of a disulfide bond. The dotted lines indicate continuation of the peptide sequence. We leave as Exercise 4.4 to complete the drawing including the resonant bond character linking O-C-N as in Figure 4.1.

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**. Thus there are four (or six, depending 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.

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 4.6. 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 [170, 312]. These proteins would be highly disordered without the disulfide bridges.

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

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 [404] (cf. Table 1.2 of [82]). 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 [182] of sidechains between protein core and water (Section 5.3).

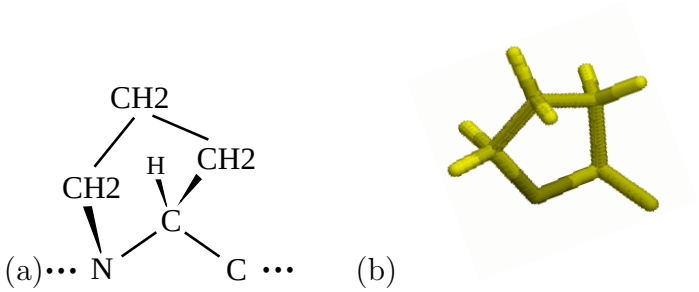


Figure 4.7: Proline sidechain configuration: (a) cartoon and (b) configuration of Pro165 in the PDB file 1QM0 (the hydrogen attached to the C_α carbon is not shown.

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 [82]. But there are some special properties of sidechains that deserve special mention here for emphasis.

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.

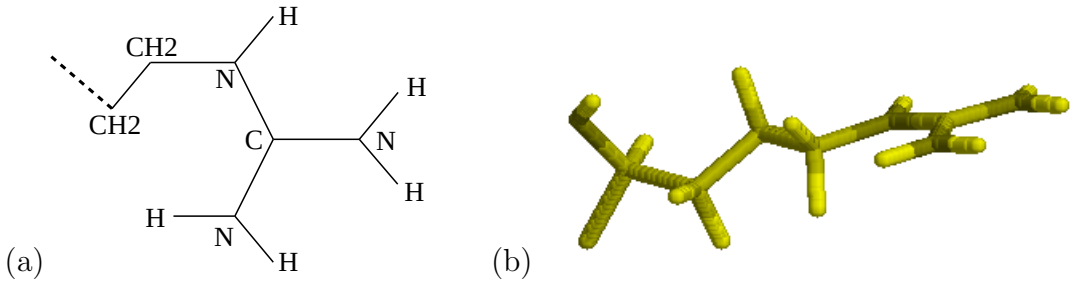


Figure 4.8: The planar structure of the terminal atoms of the Arginine sidechain: (a) cartoon in the plane of the terminal groups, (b) a rotated view of Arg220 in PDB file 1QM0 to show the planar structure from the side.

4.4.2 Proline

Proline is unique because it connects twice to the backbone, as depicted in Figure 4.7. This causes a special rigidity not found with other residues. Proline has substantial implact on the stabilization of loops [10, 172, 174, 185, 272, 326].

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 [146, 374].

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 4.9. Compounds of guanidinium (e.g., guanidinium hydrochloride and guanidinium thiocyanate [278, 365]) have the ability to **denature** proteins, that is, to cause them to unfold [105, 445]. Urea (CON_2H_4, a.k.a. carbamide) is also a protein denaturant [366, 361, 63] and resembles guanadine except that the NH group is replaced by an oxygen, cf. Figure 4.9. How the denaturing process occurs is not fully understood [295, 365, 361], 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 [366, 63]. Urea can both denature proteins and dissociate protein complexes [361].

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 [278]. The atoms beyond the C_δ carbon are all in a plane, cf. Figure 4.8. 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.

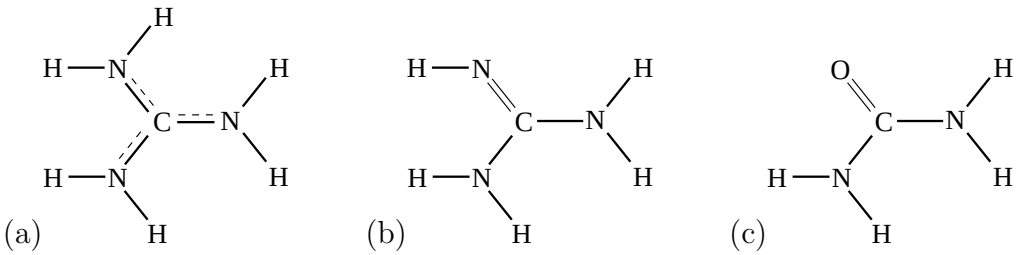


Figure 4.9: The molecules (a) guanidinium (ion), (b) guanidine and (c) urea. The net positive charge for the guanidinium ion (a) is distributed around the three nitrogen groups, as is indicated by the dashed lines.

One property of arginine is that polyarginine is the polypeptide most able to cross a cell membrane without the help of a transporter molecule [292], and compounds rich in Arg have similar behavior [452].

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 [101]. 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 [101]). 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 [305] 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 re-

[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.

member. 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 [428] 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 14.7.1). The number of water molecules in the simulation in [428] 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.

The ends of some sidechains can vary depending on the ionic composition of the solvent [82]. The pH value (Section 14.7.1) 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 [82, 404] 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 4.10. 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 4.10, 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.

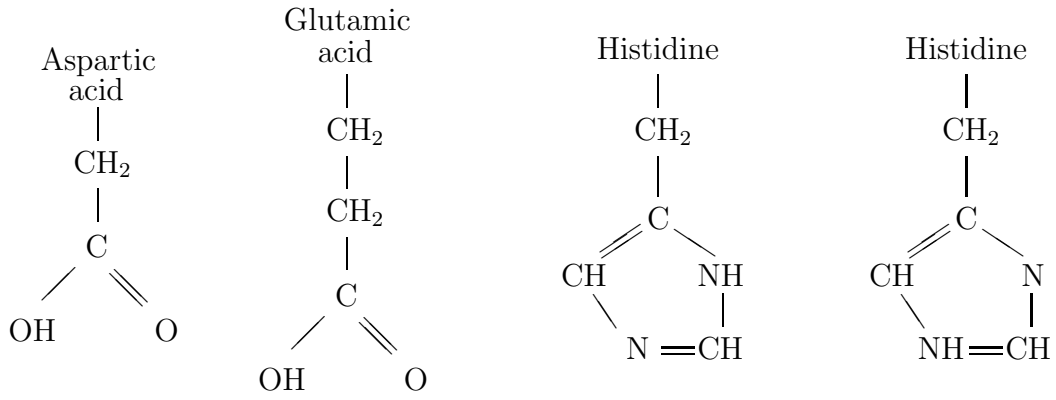


Figure 4.10: Neutralized sidechains. Either oxygen on Asp and Glu could be protonated. There are two forms of His as shown with either nitrogen being protonated [337].

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 [43]. There are two ways in which His can become deprotonated, as indicated in Figure 4.10. 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 [152, 85, 165, 249, 403, 395, 217, 203, 287] that is definitive for each sidechain, 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 [283].

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.

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 trivial lens: we simply 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 simply 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. We will formalize this later in (7.1) when we look at some more sophisticated metrics for examining the likelihood of finding things with a given property.

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 [52], listed in column two in Table 4.3. We do not claim that the small set of proteins used in [52] 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 [52] 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 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) were eliminated, and PDB files having a resolution of less 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 may be underrepresented in the PDB because they frequently cannot form any stable structure, and thus are not seen in the PDB [194].

We have a total of five datasets represented in Table 4.3, one for each column. We would expect that the values in the third column would lie between the 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 three of the outliers are Cys, His, Met, and Trp, the three 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.

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.

4.8 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.

4.9 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.*

Exercise 4.3 *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.*

Exercise 4.4 *In Figure 4.6, 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.*

Res.	[52]	All	Helix	Sheet	Loop
Ala	7.77	8.40	11.26	6.33	6.62
Arg	6.27	4.73	5.47	4.36	4.17
Asn	3.36	4.63	4.18	2.86	6.13
Asp	5.42	5.69	5.79	3.29	7.53
Cys	0.78	1.69	1.22	1.87	1.55
Gln	3.15	3.72	4.58	2.97	3.27
Glu	8.59	6.28	8.12	4.48	5.40
Gly	7.30	8.03	5.16	5.35	12.61
His	1.92	2.26	2.10	2.27	2.42
Ile	6.66	5.49	5.12	9.64	3.42
Leu	8.91	8.46	10.44	9.67	5.67
Lys	7.76	5.81	6.43	4.73	5.81
Met	2.41	2.00	2.32	1.92	1.72
Phe	3.61	3.98	3.88	5.66	3.09
Pro	4.35	4.62	2.65	1.90	8.28
Ser	4.66	6.24	5.91	5.17	7.23
Thr	4.87	5.77	4.77	6.95	6.13
Trp	1.02	1.55	1.57	1.99	1.26
Tyr	3.00	3.68	3.38	5.41	2.98
Val	8.17	6.97	5.64	13.19	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 [52]. 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.

PDB code	sidechain name	analog/comment
ACE	acetyl group	truncated residue
AIB	alpha-aminoisobutyric acid	truncated Val (C_β removed)
NGA	3(C8 H15 N O6)	N-acetyl-D-galactosamine
CCN	2(C2 H3 N)	Acetonitrile
DAL	D-Alanine	D conformation of Ala
DBU	3(C4 H7 N O2)	(2e)-2-aminobut-2-enoic acid
HYP	hydroxyproline	OH group attached to C_γ of Pro
MSE	Selenomethionine	Methionine (S replaced by Se)
PCA	C5 H7 N O3	pyroglutamic acid
PHL	L-Phenylalaninol	Phenylalanine [375]
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	ClO^-	chloride ion
GAL	beta-d-galactose	(C6 H12 O6)
HOH	H_2O	water
MOH	methanol	wood alchohol
NAG	n-acetyl-d-glucosamine	

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

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 4.2. 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 [143] 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 5.4.

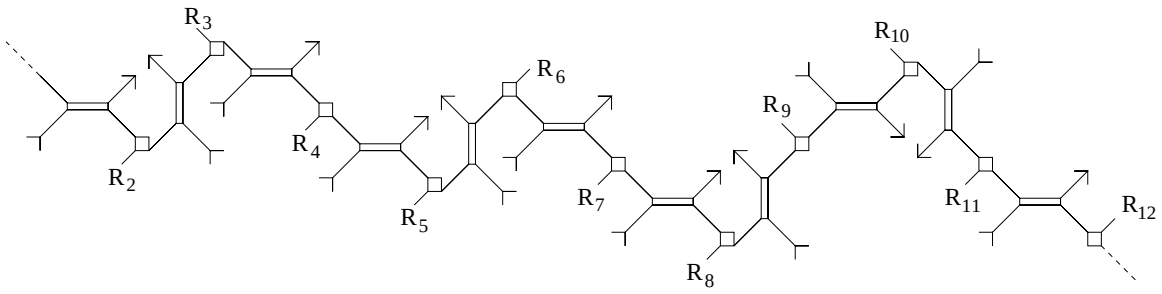


Figure 5.1: Cartoon of a peptide sequence where all of the peptides are in the trans form (cf. Figure 4.3). The small boxes represent the C-alpha carbons, the arrow heads represent the amide groups NH, the arrow tails represent the carbonyl groups CO, and the thin rectangular boxes are the double bond 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 [211, 309, 384].

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 [425]. 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 [256].

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 [109]. Not suprisingly, the frequency of alternation is approximately two in beta-sheets,

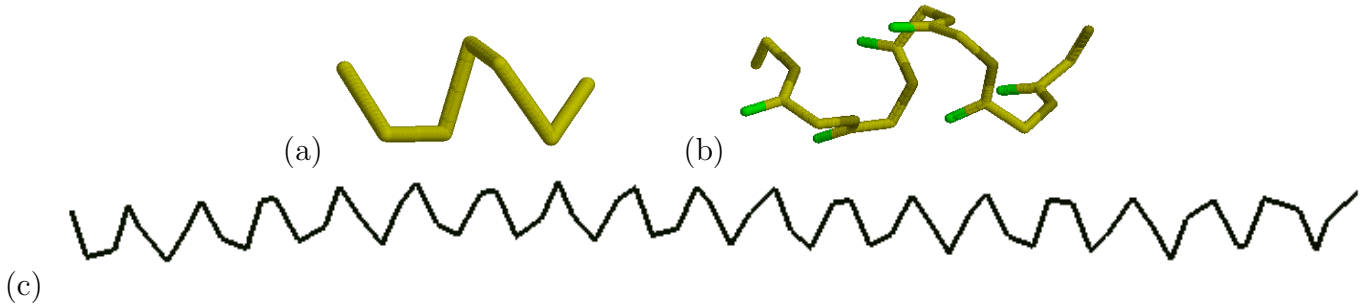


Figure 5.2: Two helices (one short, one long) in chain B in PDB file 1HTM, which depicts the Influenza Hæmagglutinin HA2 chain. (a) and (b) represent the helix formed by the seven residues B-Asn-146 through B-Ile-152: (a) shows only the backbone and (b) indicates all of the backbone heavy atoms, showing 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.

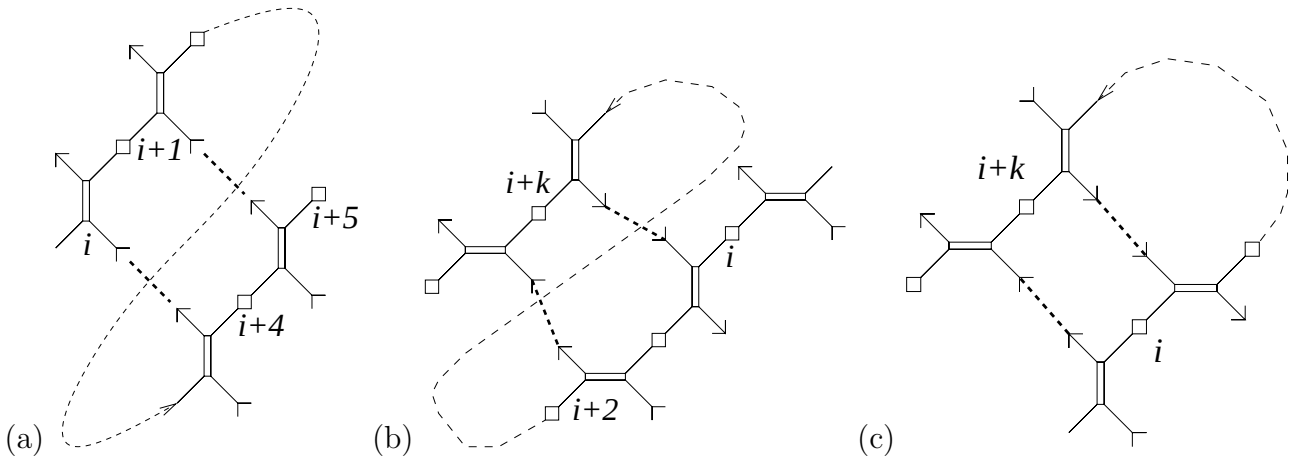


Figure 5.3: The hydrogen bond (dashed line) configuration in (a) α-helix, (b) parallel β-sheet, and (c) antiparallel β-sheet. The dotted lines indicate how the backbone is connected. The amide (N-H) groups are depicted by arrow heads and the carbonyl (O-C) groups are depicted by arrow tails, thus indicating the dipole of the backbone.

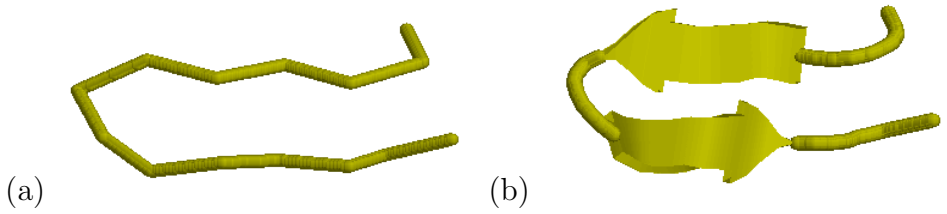


Figure 5.4: Two views of the peptide in PDB file 1K43 which forms an antiparallel β-sheet: (a) backbone, (b) ribbon cartoon.

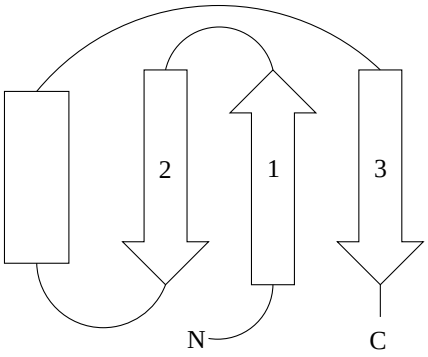


Figure 5.5: Example of a simple fold with three beta sheets and one alpha helix (in the primary sequence between sheets 2 and 3). The letters N and C indicate the N-terminal and C-terminal ends. Proximity of the edges of the beta sheets indicates hydrogen bonding.

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 [431]. The Structural Classification of Proteins (or SCOP) database [6] provides a classification of these folds [431]. 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 [11, 107, 64]. The I-sites library of a dozen or so structural units has been used [64] 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. [432]). 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 [254] 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 4.3. 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:

$$C_\alpha(x), C_\beta(x), C_\gamma(x), C_\delta(x), C_\epsilon(x), C_\zeta(x), C_\eta(x)$$

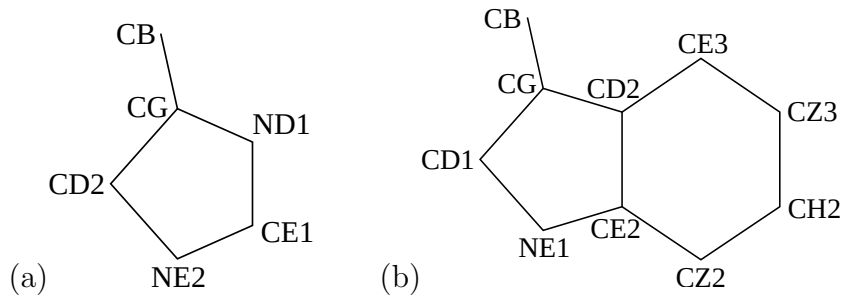


Figure 5.6: PDB notation for (a) histidine and (b) tryptophan.

are the $\alpha, \beta, \gamma, \delta, \epsilon, \eta$ 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_{γ^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 [256], 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 [256], in keeping with the planar nature of sheets.

The distribution of the θ angle peaks roughly at 44, 82 and 167 degrees [256]. The peptide bond makes it difficult for θ to be much less than 50 degrees, thus the smaller peak corresponds to a motif

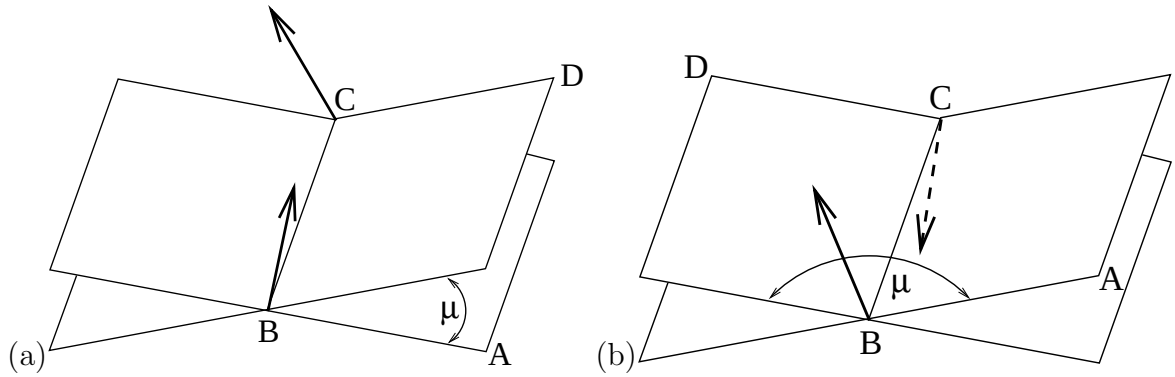


Figure 5.7: The geometry of the dihedral angle $[A, B, C, D]$ in two different configurations: (a) cis form [39], (b) trans form. The normals are defined using the left-hand rule.

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 [14, 256]. For the larger values of θ, hydrophobic residues are found in most pairs; β-sheets have alternating hydrophobic and hydrophilic pairs [256]. By contrast [256], 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 [14].

5.2.2 Dihedral angles

We also recall the standard main-chain **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 5.7. 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 5.7, 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. 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

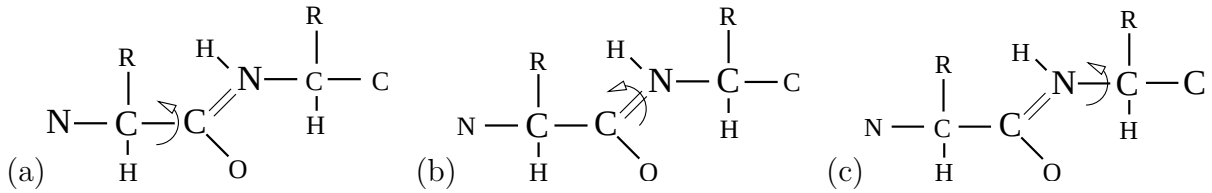


Figure 5.8: The primary dihedral angles: (a) ψ, (b) ω, (c) ϕ. The four points used in each definition are in a larger font. The arrows indicate the axes of rotation.

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 ω.

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 5.6). This correlation has recently been observed to have significant predictive power [149].

The conformations of ϕ_i, ψ_i are typical of different secondary structures, such as α-helix or β-sheet. The Ramachandran plot [191] depicts the distributions of angles that are commonly adopted (cf. Exercise **??**).

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.

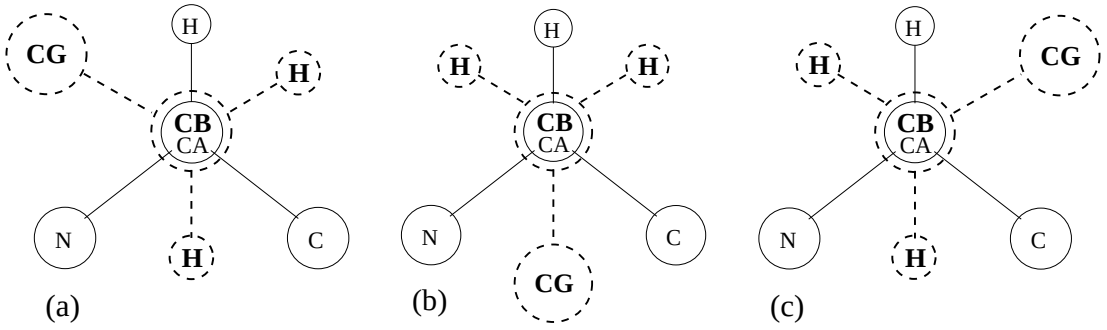


Figure 5.9: The primary sidechain rotamer conformations (a) gauche+, (b) gauche-, and (c) trans, corresponding to χ_1 values of (a) -60 degrees, (b) +60 degrees and (c) 180 degrees. The view is oriented so that the C_α and C_β atoms are aligned perpendicular to the plane of the page. The closer (and 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.

Libraries of angular orientations of the different segments have been developed [261, 339]. 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 5.9, 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 [257].

5.3 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 [182], and subsequent studies have further refined estimates of sidechain volume, including sizes of individual atom groups [408]. For reference, we have summarized in Tables 5.2–5.4 some estimates [408] 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 [408]. For reference, a water molecule occupies about 30Å^3 (see Section 14.7.3).

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 [182, 408] 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

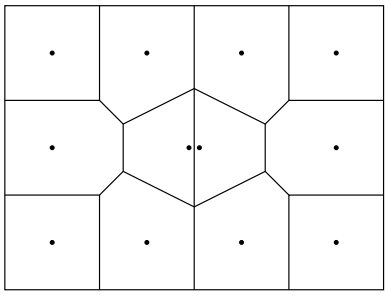


Figure 5.10: Voronoi diagram for a set of points in two dimensions.

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 [408], 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 5.10, 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 5.10 we show an extreme case in which the points in S can get arbitrarily close to the boundary of the enclosing P_s. In [182, 408], 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.

At first blush, it might seem that the volumes occupied e.g. by the carbonaceous groups in Table 5.2 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 5.5 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 5.2–5.4.

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(γ^1) in threonine, whole volume is about 10Å^3. There are some obvious trends, e.g., in Table 5.2 it is clear that the volume of the carbonaceous 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 is lysine is only 20.6-21.9Å^3, smaller than

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 5.2: Volume of carbonaceous atom groups CH_n in protein sidechains [408]. 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 5.3: Volume of nitrogen atom groups NH_n in protein sidechains [408]. t is the number of carbonaceous groups in the category and n is the number of hydrogens in these groups.

the NH_2 groups of argenine.

But there is also a more subtle size issue that is solvent dependent. One of the significant conclusions [182] 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 the 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 [182, 408], although the apportionment between backbone and sidechain differs systematically. This

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 5.4: Volume of protein backbone atom groups (Å^3) [408]; '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 5.5: Relation between volume and length in the range of volumes relevant for proteins.

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 residues would remain the same.

5.4 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 [415]. The distribution of multiple domains is known [431] to be exponential. More precisely, the probability $p(k)$ of having k domains was found [431] to be closely approximated by

$$p(k) \approx 0.85e^{0.41(k-1)}. \tag{5.7}$$

Such a distribution implies [431] "the evolution of multidomain proteins by random combination of domains."

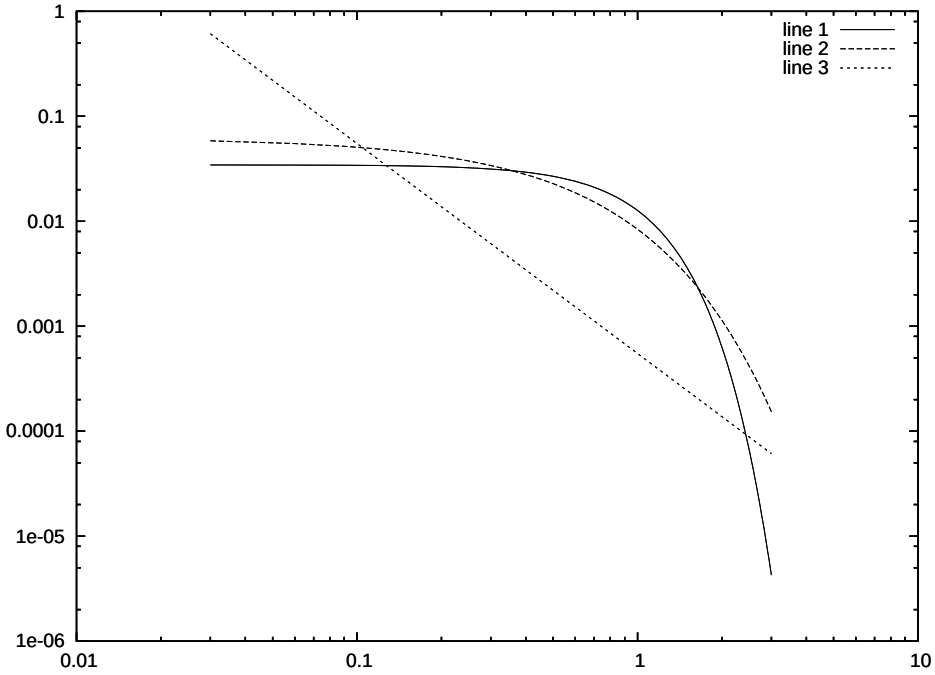


Figure 5.11: Comparisons of the probability distributions $c_1 e^{-x^2}$ (solid line), $c_2 e^{-2x}$ (dashed line), and $c_3 x^{-2}$ (dotted straight line) on the interval $[0.03, 3]$. The constants c_i are chosen to make each distribution integrate to one on $[0.03, 3]$.

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 [436]. 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 [345, 436]. More precisely, if we let $p(k)$ be the probability of having k neighbors in this graph, we find

$$p(k) \approx ck^{-\gamma}. \tag{5.8}$$

A network with a distribution (5.8) is often called a **scale free networ**k [35]. This provides a useful baseline regarding what to expect in terms of protein interactions. We will see just such a distribution in Figure 11.9.

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 [436], 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 [436] was Pfam [148]. The objective of "Pfam is a comprehensive collection of protein domains and families, represented as multiple sequence alignments and as profile hidden Markov models"

[148]. For a description of profile hidden Markov models, see [107].

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 [413]. On the other hand, sequence similarity does not always imply fold similarity [173].

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 [431] (twenty species were studied in [345]). 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 [345, 436] to follow a power-law with a distinctive decay rate for each species, with "those of smaller genomes displaying a steeper decay" [345], that is, a larger γ in (5.8). We will see a similar type of behavior in Figure 11.9 when we consider the relationship between wrapping and structural connectivity.

We have described two characteristic probability distributions here, the exponential (5.7) and the power-law (5.8). 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 5.11 presents a comparison of the Gaussian distribution with the power-law distribution (5.8) and the exponential distribution (5.7). On thing revealed by Figure 5.11 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 11.9.

5.5 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*

$$\sum_{k=0}^{\infty} p^k q = 1, \tag{5.9}$$

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" [431].

Exercise 5.2 *Use protein constituent volume data [182, 408] (cf. Tables 5.2–5.4) 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 5.11 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 Threonine and Valine by examining high-resolution structures.*

Exercise 5.5 *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 5.6 *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 5.7 *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*

$$\mu(p_i, p_{i+1}, \ldots, p_{i_\ell}) = \sum_{j=0}^{\ell} \mathcal{C}(p_{i+j}) \left(j - \tfrac{1}{2}\ell\right) \tag{5.10}$$

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 5.8 *Consider the definition of macrodipole introduced in Exercise 5.7. Explain why the α-helical polypeptide $Glu_{20}Ala_{20}$ would be more stable than $Ala_{20}Glu_{20}$.*

Exercise 5.9 *Determine the Ramachandran plot [191] 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, sheet or turn in the PDB file.*

Chapter 6

Hydrogen bonds

The concept of the hydrogen bond was established by 1920 [241], and possibly earlier [224, 78]. 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 [28, 400, 235], they are not yet fully understood and are still actively studied [201, 202]. 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 [151].

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.

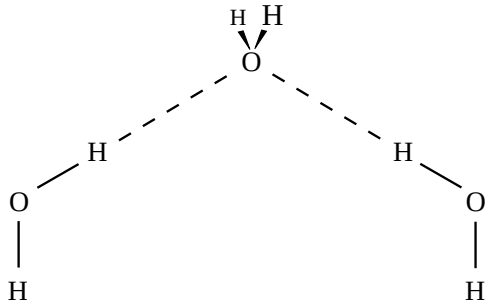


Figure 6.1: Cartoon of the configuration of three water molecules and the hydrogen bonds formed, depicted by the dashed lines. The oxygen atoms accept two hydrogen bonds, whereas the hydrogen atoms are involved in donating only a single bond.

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

Table 6.1: 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" [227].

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 [227]. 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 [227]. In some sources (including Wikipedia), the electronegativity of the constituents is given as the key factor. But according to [227], "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

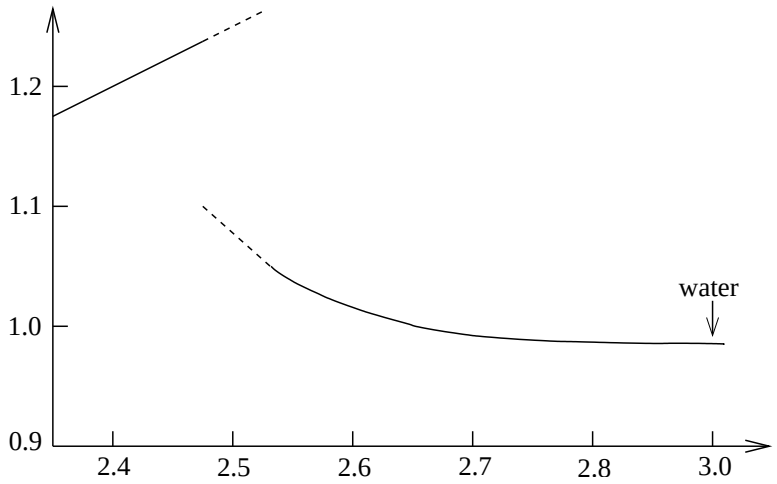


Figure 6.2: Cartoon of the bifurcation of O-H..O hydrogen bonds from a symmetric arrangement to an asymmetric arrangement, based on Figure 4 of [335]. The horizontal axis is the O-O distance and the vertical coordinate is the O-H distance (both in Ångstroms). The upper-left segment is the symmetric arrangement. 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.

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.

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 [387]. In Chapter 16, 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

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.

bridges [333, 437]. 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 [202]. In some cases, the network of hydrogen bonds can be complicated, as shown in Figure 6.3.

6.1.2 Hydrogen bond strength

Assessing the strength of hydrogen bonds remains a significant challenge [242, 389]. There is a strong angular dependence for the energy of the hydrogen bond [400, 306]. Moreover, the nature of the hydrogen bond can depend on the context: even backbone-backbone hydrogen bonds can be different in alpha helices and beta sheets [400, 226, 240].

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" [227], 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, 291]. The difficulty with models of

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.

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 optimal 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 [388, 390].

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 [306], 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.

The accurate simulation of one of the simplest hydrogen bonds, in the water dimer, has been of recent interest [223], even though this computation has been carried out for several decades [227]. 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 [215, 258, 171].

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 [400, 151]. However, in many cases, the positions of the missing hydrogens can be inferred according to simple rules. For example, 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

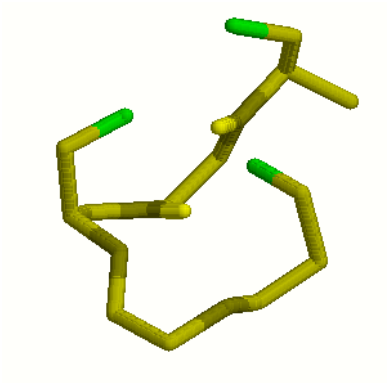


Figure 6.3: Three serines (52, 54 and 56) from chain H the PDB file 1DQM come together to form a complex of hydrogen bonds. The amide groups on the peptide base of H-Gly55 and H-Ser56 provide donors to the oxygens on the end of the sidechain of H-Ser52. The sidechain-sidechain hydrogen bonds among the terminal OH groups on the serines depend on the hydrogen placements.

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 [323]; also see the more recent corrections to estimates of bond lengths and angles in [200].

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 6.4. The angle between the C-N and the C-O vectors is very close to 120 degrees [297], 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 [297].

The position of most hydrogens can be modeled by the bond lengths and angles given in [297]. A program called HBPLUS [282] 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) [165, 151].

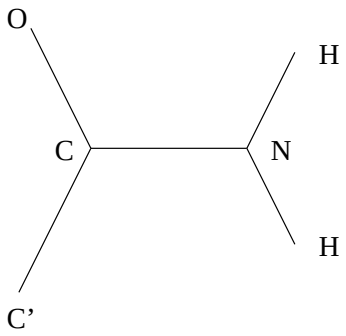


Figure 6.4: Hydrogen placement for Asn and Gln. Shown is the terminal group of atoms for the sidechains. The atom marked C' denotes the preceding carbon in the sidechain, viz., CB for Asn and CG for Gln.

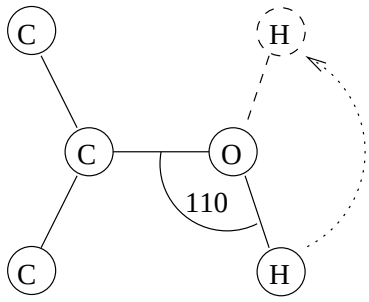


Figure 6.5: Hydrogen placement for Tyr. The hydrogen is in the plane of the aromatic ring, with the angle between C-O and O-H being 110 degrees. Both positions are possible for the terminal hydrogen.

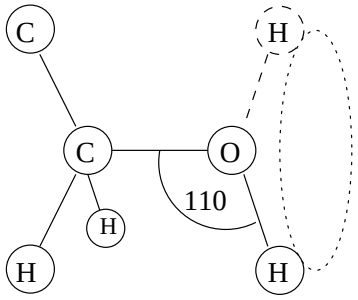


Figure 6.6: Hydrogen placement for Ser and Thr: anywhere on the dotted circle. The angle between C-O and O-H being 110 degrees, but the hydrogen is otherwise unconstrained. A Cys sidechain not in a disulfide bond would be similar, with O replaced by S.

Most hydrogens can be located uniquely. In particular, the Appendix in [297] depicts the locations of such hydrogens, as well as providing precise numerical coordinates for their locations. However, other hydrogens are not uniquely determined. For example, 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 6.5. The one which makes the stronger H-bond with an acceptor is presumably the one that is adopted.

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 6.6. A Cys sidechain that is not engaged in a disulfide bond would be similar, with the oxygen in Figure 6.6 replaced by a sulfur.

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. 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 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

One 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 6.8) based on those used in [28], 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

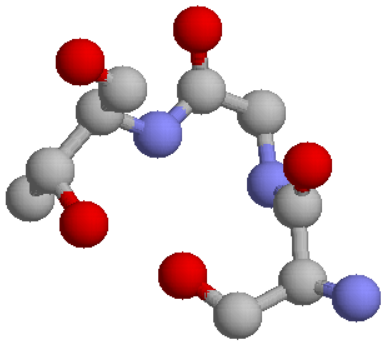


Figure 6.7: Ambiguous hydrogen placement for serine-28 (lower right)—threonine-30 (upper left) sidechain-sidechain hydrogen bond involving the terminal O-H groups; from the B chain in the PDB file 1C08. The sidechain of isoleucine-29 has been omitted but the backbone atoms are shown connecting the two residues. Only the oxygen atoms in the terminal O-H groups are shown.

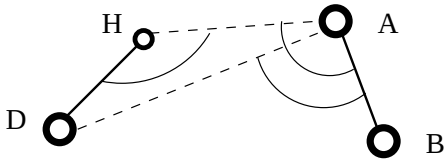


Figure 6.8: Geometric model for hydrogen bonds: D is the donor atom, H the hydrogen, A the acceptor, B acceptor antecedent (i.e. an atom one covalent bond away from the acceptor).

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.

6.4 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 [430]. These are referred to as carboxyl-carboxylate hydrogen bonds. Although these bonds would be expected in low pH environments [364], they have been found to be critical elements of ion channels [300]. 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.

6.5 Aromatic hydrogen bonds

In general, any negatively charged entity might form 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).

6.6 Carbonaceous hydrogen bonds

In some cases, carbonaceous groups (CH) can act as donors for hydrogen bonds [273].

6.7 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 [323] or the more recent corrections to estimates of bond lengths and angles in [200, 405].*

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 [297].*

Exercise 6.3 *Use the model for the energy of a hydrogen bond in [306] 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?*

Chapter 7

Determinants of protein-protein interfaces

We now turn to a key question: what factors are most influential in protein-ligand binding? Although we ultimately want to consider general ligands, we only work with ones that are themselves proteins in this chapter. 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 naıvely to get a sense of how the ideas developed.

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, 76, 156, 208, 260, 296, 351, 419]. A common belief has been that hydrophobic interactions are a dominant motif for protein-ligand binding. According to [437], "The prevailing view holds that the hydrophobic effect has a dominant role in stabilizing protein structures."

There have been several attempts to define a hydrophobicity scale for protein sidechains as a guide to protein-ligand binding [44, 109, 237, 287, 294, 310, 340]. A similar, but distinct, concept is that of **lipophilicity** [274] which measures the extent to which substances dissolve in a nonpolar solvent. Although the scales are designated as hydrophobicity measures, they are really intended to be proxies for the local dielectric environment [287]. 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. As we will see in Chapter 19, 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 [167] to discover hydrophobic interactions by examining protein-protein interfaces. The hypothesized form of the interactions in such studies determined the basic choices

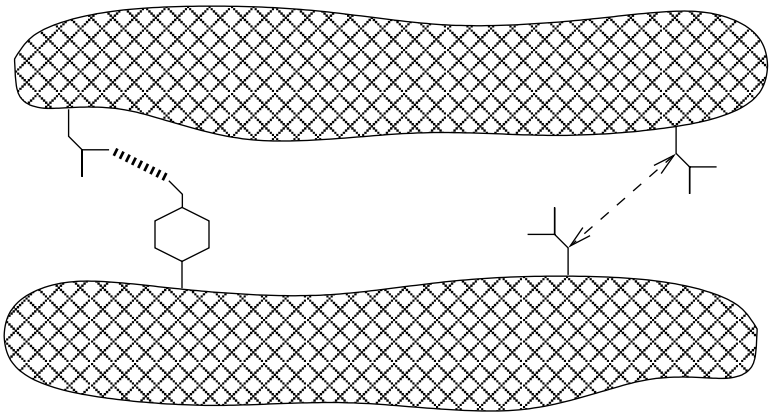


Figure 7.1: Cartoon showing a protein-protein interface and some possible ways to characterize interactions between residues. On the right, two residues (Asp and Asn fit the description, but Gln, Glu or Val are similar) have their C_β carbons close. On the left, a hydrogen bond between Tyr and Gln is depicted, but with the C_β carbons of each at a greater distance.

that guided the data mining. In particular, a definition of 'sidechain interaction' using only the proximity of C_β carbons was used [167] to quantify interactivity; it does not take into account individual sidechain features. This is depicted in Figure 7.1 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, 118, 125, 128, 145, 304, 419].

Water removal from hydrophobic patches on protein surfaces has a high thermodynamic benefit [47, 76, 156, 208, 260, 296, 351, 419], 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 ranging typically from 7:1 to 10:1 [139]. Moreover, hydrophobic patches involved in associations at an interface are often paired with polar groups [383]. We will ultimately explain how this can be energetically favorable, but for now let us switch to another point of view, namely the identification of inter-molecular bonds across protein-protein interfaces.

Whether or not hydrophobic effects are important for protein-ligand binding, one might also

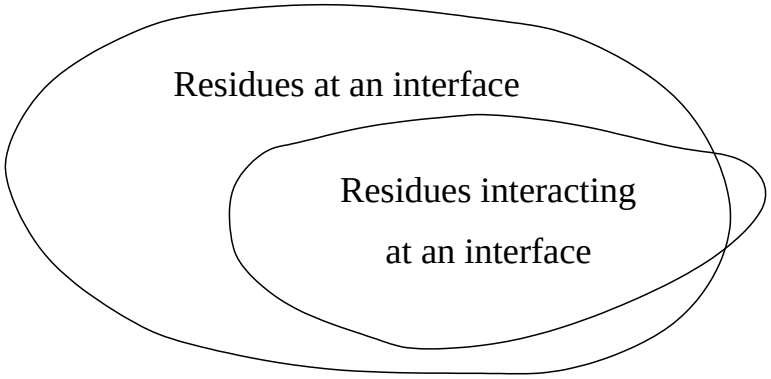


Figure 7.2: Cartoon showing possible relationship between two datasets.

expect that the sort of bonds that help proteins form their basic structure would also be involved in joining two different proteins together. Both hydrogen bonds and salt bridges play a significant role at protein interfaces [437]. The density of hydrogen bonds between two different proteins at an interface is about one per two square nanometers. If you think of a checkerboard with nanometer sized squares, then it is like having one hydrogen bond on each of the red squares. The average number of hydrogen bonds per interface is about ten. On the other hand, the average number of salt bridges per interface is only two [437]. 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 [437] 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 2.7.

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 [437], polar atom pairs bridged by water across the interface with hydrogen bonds were more numerous than direct hydrogen bond pairs, with each water molecule connecting 3.8 cross-chain atom pairs on average.

7.1 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, but we also review others in Section 7.5. The data here is drawn primarily from [52, 144, 167]. 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 [167] if the distance between their C_β carbons (C_α for Gly) was less than 6 Ångstroms. The contact area for a protein-protein interface may be estimated by comparing the solvent accessible 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. Since there is little agreement in how interfaces are defined, we will not attempt to give the details in each case.

The site specificity of protein-protein interactions has been widely studied due to its central biological significance [90, 167, 192, 208, 209, 210]. 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 [137, 381]. But it is also true that such residues are more abundant over-all (see Table 7.2).

The first question [144] 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 [52].

The second question [167] 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 [167] is a natural metric. Thus, two residues are defined [167] 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, but it is notable that the same sort of simple measure based on the initial segment is successful in other contexts [257]. 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.

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 7.2, 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 [52], 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.

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	1	+0.25	+0.24	2.5	-1.0
Val	V	3	+0.20	-0.31	4.2	-1.5
Met	M	3	+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	3	-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 (7.3). Dehydron Propensity is also presented as frequency $f - 5$; 5% is the average propensity to be at interface or engaged in a dehydron. The hydropathy scale of Kyte et al. [237] and the hydrophobicity values s of Levitt [246] are included for reference.

7.2 Interface propensity

The common belief is that hydrophobic residues on the surface of proteins are likely candidates to support interfaces in protein-protein association. In Section 7.3, we present evidence that supports this case with suitable clarifications. However, [144] 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 [144] to be at interfaces are, in decreasing order of frequency, Asn, Thr, Gly, Ser, Asp, Ala, and Cys, the group depicted in Figure 4.4(a). None of these residues is distinctively hydrophobic. This is quite a surprising result, and it demands an explanation.

To begin with, let us clarify the basic notions. If we have a dataset with N different types of characteristics (e.g., $N = 20$ and the characteristics are the different amino acids), then the **frequency** f_i of the i-th characteristic is defined by

$$f_i = \frac{o_i}{\sum_{j=1}^{N} o_j} \tag{7.1}$$

where o_j is the number of occurrences of the j-th characteristic in the dataset. In some cases, frequencies are represented as percentages, in which case we simply multiply by 100 in (7.1).

If we have two datasets with the same characteristics, with frequencies f_i and g_i, respectively, then one can define a **relative frequency**

$$r_i = f_i/g_i \tag{7.2}$$

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 compare two different comparisons. In [144], the following approach was taken.

Define a normalized **relative propensity** via

$$R_i = \frac{r_i}{\sum_{j=1}^{N} r_j}. \tag{7.3}$$

These relative propensities sum to one, so we can think of them like ordinary frequencies. Similarly, we multiply by 100 in (7.1) 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 [144]. 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 [144] 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 [144] 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 main-chain 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 [144], namely, the amount of wrapping a residue can provide.

As noted in [144], the seven residues in Figure 4.4(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 4.4(b)), with only one sidechain rotameric degree of freedom. Therefore, its conformational hindrance upon binding also entails minimal loss in conformational entropy.

Considering 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 [144] 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 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).

7.3 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 review the results in [167] 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 recall that the hypothesis of [167] 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 [167] 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. We present in Table 7.2 the residues and their relative propensities, as defined in (7.3), in decreasing order, as determined in [167].

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 [167]. 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' [346], so this is indeed a result in line with the basic hypothesis. But there

Res. Code	Pairing Rel. Prop.	Pairing Rel. Freq.	Pairing Freq. [167]	Total Abundance [52]	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 [167]. Primary data is taken from the indicated references. Relative Propensity is defined in (7.3) and Relative Frequency is defined in (7.2). Interface Relative Propensity from Table 7.1 is included for comparison.

are other sidechains which are purely hydrophobic, and have few other interesting features, such as Val, Leu, Ile and Phe. If there hypothesis 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 [167], 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 precisely, to report frequencies as a percentage, define

$$F_i = 100 \frac{\sum_{j=1}^{20} C_{ij}}{\sum_{i,j=1}^{20} C_{ij}} \tag{7.4}$$

to be the amino acid pairing frequency, shown in the column entitled 'Pairing Freq. [167]' 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 [52] which are reproduced in the column entitled 'Total Abundance [52]' 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 (7.4), to the abundances reported in [52].

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 [167].

7.4 Pair frequencies

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

$$O_{ij} = \frac{C_{ij}}{f_i f_j} \tag{7.5}$$

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 one implies that the pairing is more common than would be expected for a random pairing, and conversely if it is less than one.

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 [167]. Note that the numbers listed in that table are the Log-odds-ratio inflated by a factor of $A = 10$.

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. In [167], a quantity G_{ij} is defined by multiplying the log odds ratio by a numerical factor of ten.

It is noteworthy that the odds ratios indicated in Table III of [167], 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 [167] 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 [346].

The subsequent four pairs in [167] 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 [84, 159, 439] (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 show a similar mix of polar interactions, not the expected hydrophobic-hydrophobic interactions.

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 [167] are much smaller than found for alpha helices or beta sheets [257]. The top thirty values for the odds ratios for amino acid pairs with $\theta < 50$ (Section 5.2.1) are all greater than two, with the highest being 3.75 [257]. Moreover, the top fifteen values for the odds ratios for amino acid pairs with $\theta > 155$, that is pairs in β sheets, are all greater than two [257]. 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 [167] was flawed, rather that the hypothesis was proved to be incorrect. If the dominant interactions had been hydrophobic, these techniques would have discovered them. One the other hand, it could be possible to pre-filter the data to eliminate identifiable bonds, and this might provide interesting new data.

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. 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.

In [167], the typical configuration of Arg-Trp is pictured, and similar polar pairings are highlighted, such as Lys-Lys (odds ratio 0.81).

7.5 Comparisons and caveats

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 [167]. However, the methodology is quite similar to that of the earlier paper [407], 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 [167] is different in definition from exposed and interior residues in the interface in [407], although similar in spirit. Figure 3B of [407] shows how the residues that are interacting (proximate) in an interface are very similar in composition to ones in the interior of proteins.

To illustrate the sensitivity of results depending on the database chosen, we review the results in [24] 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 preferences in order for the data in [71] to facilitate comparison with the data in [24]. 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 [24].

The dissection trilogy is completed in [25] in which an attempt is made to determine aminoacid distributions for 'nonspecific' interactions. This is intended to be a proxy for any surfaces which

Res. Code	Rel. Prop.	Rel. Freq.	Pair Freq.[167]	Total Abundance [52]	Rim/Core freq. [71]	Homodimer Rim/Core [24]
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.

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 [25] for a comparison with the data in [71] and [24].

Protein-ligand interfaces differ in function, and interfaces with different function can have different composition. In [208], 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 [311].

In [46], 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. 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 change in geometry for these residues. However, such substitutions might be extremely rare. 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.

The aromatic sidechains do play a special role in protein interfaces through what is called a cation-π interaction [159] (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 [247]. The cation-π motifs play a special role in protein interfaces [84, 439]. The cation-π interaction also has a significant role in α-helix stabilization [373].

A study of the role of evolution on protein interface composition can be found in [66]. In [180, 264], 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 [307, 308].

Identification of individual sidechains that may play the role of 'anchors' in protein-ligand recognition is studied in [347] 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 [77] as a way to discriminate between correct structures and incorrect ones. In [77], 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 Conclusions

Two main conclusions were obtained. The first is that residue hydrophobicity is not the primary variable 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 do 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.

7.7 Exercises

Exercise 7.1 *Compare the data for the surfaces in [24, 25, 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 [25], the distributions for nonspecific interaction surfaces are compared with the distributions for other surfaces [71, 24]. The comparison metric is the L^2 norm. Consider the effect of using the KL-divergence, Jensen-Shannon metric, and the earth-moving metric Section* **??**.

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 *Re-do the analysis in the chapter using the abundance data in Table 4.3 instead of [52].*

Chapter 8

Wrapping electrostatic bonds

By 1959, the role of hydrophbicity in protein chemistry was firmly established [214]. Soon afterward [155, 225], 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 [225] 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 absorbtion 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: "It seems unlikely, therefore, that interpeptide hydrogen bonds contribute significantly to the stabilization of macromolecular configuration in aqueous solution." [225] However, the authors did confirm the oppositve view in less polar solvents, so we would now say that their study indicated the requirement of hydrophobic protection of hydrogen bonds in proteins.

The subsequent paper [155] 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 [29, 30], it now seems clear that the variation in hydrophobicity in proteins has a significant and quantifyable effect on the behavior of proteins [93]. According to [437], "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' in proteins, signalling sites of interest, is having significant success among diverse groups [57, 135].

Attempts to quantify the hydrophobicity of different sidechains has a long history [246]. 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** [26] and **shielding** [161, 263]. We prefer the term wrapping sinces 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 [26] of hydrogen exchange [27], 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 [27] 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 [411], 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 [26, 27] 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 [263] 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 [263] 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 [16, 17, 20, 21]. More recent, and more complex, experiments [160] confirm that hydrogen bond strength is enhanced by a nonpolar environment.

Based on the accumulated evidence, 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 [278], in addition to the fact that it contains significant carbonaceous groups. A computational study [161] 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 [246] (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, 246] essentially does that. But the scalar quantity of real interest with regard to electrostatic bonds is the dielectric, which is described in Section 8.6.

8.1 Defining hydrophobicity

There have been several attempts to define a hydrophobicity scale for protein sidechains as a guide to protein-ligand binding [109, 237, 287, 340]. 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 [144].

Although the scales are designated as hydrophobicity measures, they are really intended to be proxies for the local dielectric environment [287]. 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 [26, 137, 145, 263, 331, 411]. 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 [137]. Such desolvation enhances the electrostatic energy contribution and stabilizes backbone hydrogen bonds [31]. 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 [86], as depicted in Figure 2.3.

The thermodynamic benefit associated with water removal from pre-formed structure makes under-wrapped proteins adhesive [125, 138, 140]. As shown in [137], 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.

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.

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 8.6. In Section 8.6, 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.

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	?	?

Table 8.1: Electronegativity scale [241, 324, 360] of principal atoms in biology. The 'outer electrons' row lists the number of electrons needed to complete the outer shell.

8.2.1 Electronegativity scale

The key to understanding the polarity of certain molecules is the **electronegativity scale** [241, 324, 360], 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 [324].

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. 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 [324]. 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

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 sulfer in Met and (CB) adjacent to sulfer in Cys may be considered polar. The notation 'na' indicates there are no nonpolar carbonaceous groups.

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.

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

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.

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 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.

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. 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 [131], 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 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 8.1. 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.

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.

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.

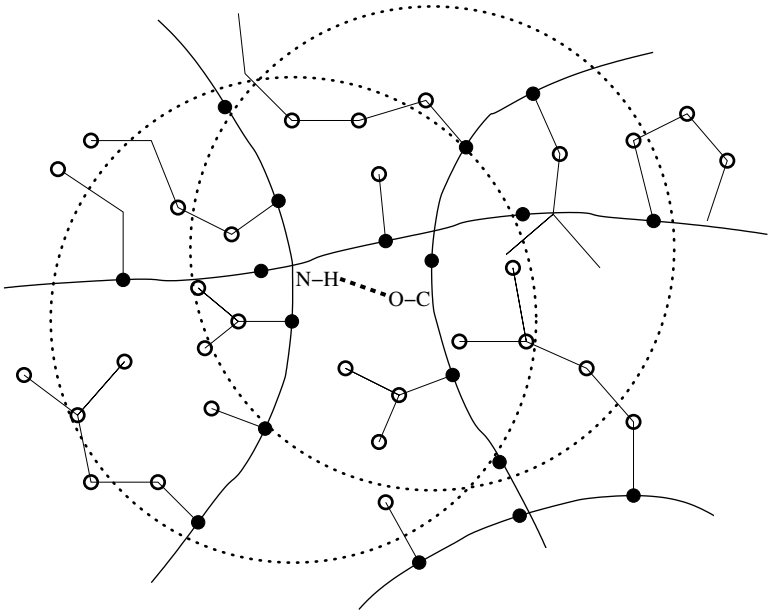


Figure 8.1: Caricature showing desolvation spheres with various side chains. The open circles denote the nonpolar carbonaceous groups, and the solid circles represent the C_α carbons. The hydrogen bond between the amide (N-H) and carbonyl (O-C) groups is shown with a dashed line. Glycines appear without 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 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 4.4(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 [131] 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 8.1. The desolvation circles in Figure 8.1 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 8.2. 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

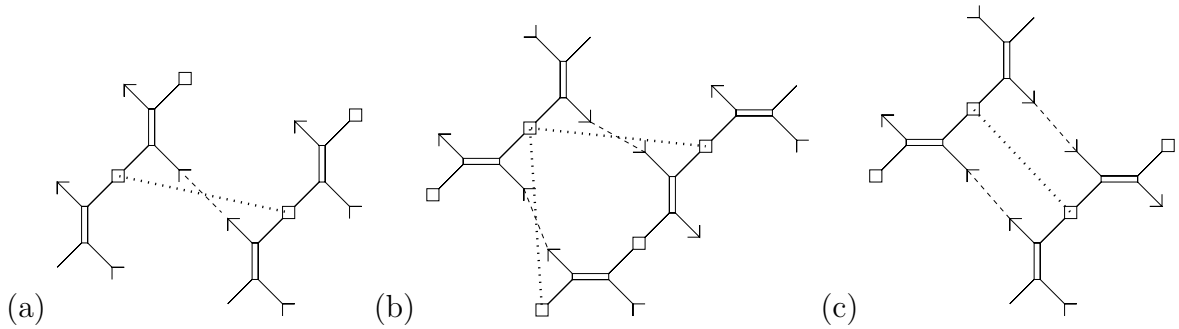


Figure 8.2: The hydrogen bond (dashed line) configuration in (a) α-helix, (b) antiparallel β-sheet, and (c) parallel β-sheet. A dotted line connects the C_α carbons (squares) that provide the centers of the spheres forming the desolvation domains in Figure 8.1. The amide (N-H) groups are depicted by arrow heads and the carbonyl (O-C) groups are depicted by arrow tails.

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 [131] 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 14.7.3).

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 [131], a nonredundant sample of 2811 PDB-structures was examined. The average ρ_R over the entire sample set was found to be 6.6 [131]. 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 $\sigma = 1.46$ [131]. 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.

The distribution of the selected proteins as a function of their average wrapping as measured by ρ_R is shown in Fig. 5 in [131]. 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 stabilized mostly by disulfide bonds.

To further assess the effectiveness of the residue-based assessment of wrapping, we review additional results and predictions of [131].

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 [189, 232, 284]. Figure 3a of [131] depicts the (many) UWHB's for the human cellular prion protein (PDB file 1QM0) [343, 350, 444]. 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 [343, 350, 444]. 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 [343]. 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, 92, 115, 175, 189, 284, 232] 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 [131]. A statistical analysis involving $\delta_{\max}$ [131] 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 [131] that six proteins might posses amyloidogenic propensity. Three of them,

- angiogenin (cf. PDB files 1B1E and 2ANG),
- meizothrombin (cf. PDB file 1A0H), and
- plasminogen (cf. PDB file 1B2I),

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

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 [245]. Meizothrombin is formed during prothrombin activation, and is known to be involved in blood clotting [213] and is able to bind to procoagulant phospholipid membranes [327]. Plasminogen has been identified as being a significant factor in wound healing [353].

8.4 Counting nonpolar groups

A more refined measure of hydrogen-bond protection has been proposed based on the number of vicinal nonpolar groups [125, 137]. 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 8.1. 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 16.1 for a radius of 6Å for the definition of the desolvation domain. In [138], 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 [137, 145]. In Figure 8.3 the distribution of wrapping is presented for a particular PDB file.

The 'group' definition of wrapping is similar to the definition of **buried** groups [249]. 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 5.3). Roughly speaking, when the local density of heavy atoms is greater than one per 39 Å^3 (corresponding to a box of side about 3.4 Å), that region is considered to be buried. For comparison, the average density of water is about one water per 30 Å^3 (corresponding to a box of side about 3.1 Å) (cf. Section 14.7.3), whereas we see in Tables 5.2–5.4 that most protein atom groups have a volume less than 39Å^3, and thus a density of greater than one per 39 Å^3.

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 8.3 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 8.3 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.

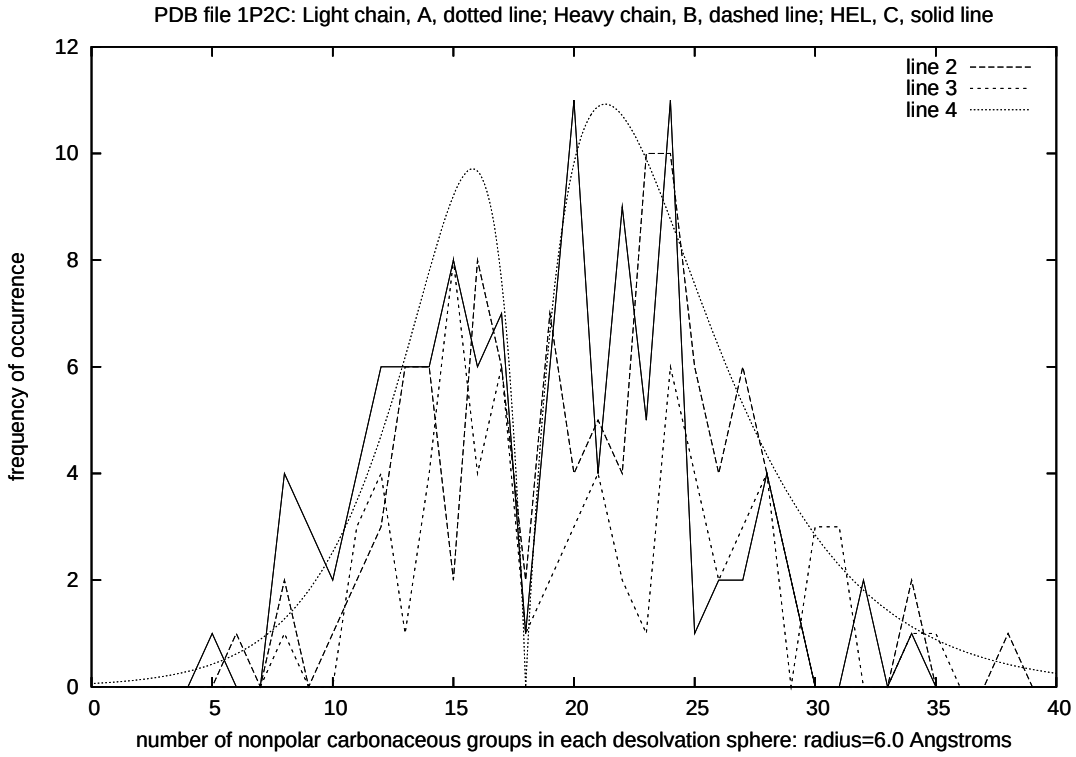


Figure 8.3: Distribution of wrapping for PDB file 1P2C. There are three chains: light, heavy chains of the antibody, and the antigen (HEL) chain. The desolvation radius is 6.0Å. Smooth curves (8.2) are added as a guide to the eye.

8.5 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 [131] 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 [131] 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.

8.6 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 [139] to quantify this effect using a continuous representation of the local solvent environment surrounding backbone hydrogen bonds [55, 125, 137, 145, 188, 313, 418]. 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 [125, 145]. 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 [125, 145].

We now describe an attempt to estimate ϵ as a function of the fixed positions $\{r_j \mid j = 1, \ldots, n_k\}$ 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 19) 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 [125], 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 [139] 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 8.1. In [139], 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

$$M_k = \max\left\{|\nabla_R(1/\epsilon(\{\mathbf{r}_j\}, R_o, r_H - r_O))| \;\middle|\; R_o \in C\right\}. \tag{8.5}$$

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 19, but for now we consider a heuristic model used in [139].

The model in [139] 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 [139], 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 groups. Further, a cut-off function

$$\Phi(\mathbf{r}) = (1 + |\mathbf{r}|/\xi)\, e^{-|\mathbf{r}|/\xi}, \tag{8.8}$$

where $\xi = 5$Å is a water dipole-dipole correlation length, approximates the effect of hydrogen bond length on its strength [139].

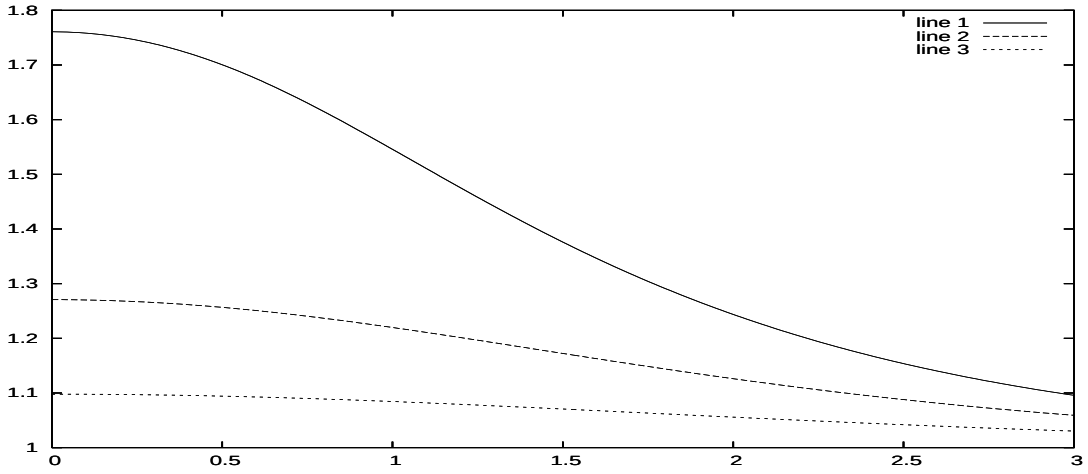


Figure 8.4: The function $\omega(x, y)$ plotted as a function of the distance along the x-axis connecting r_H and r_O, for three different values of the distance y from that axis: $y = 1$ (solid line), $y = 2$ (dashed line), $y = 3$ (dotted line). The coordinates have been scaled by Λ and the value of $|\mathbf{r}_O - \mathbf{r}_H| = 1$ was assumed.

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 8.4 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 [139] 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 [187], were found to be highly sensitive in the sense that

$$M_k > \lambda/10, \tag{8.12}$$

where λ was defined to be

$$\lambda = \frac{\epsilon_o^{-1} - \epsilon_w^{-1}}{2\text{Å}}. \tag{8.13}$$

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 [139] to the definition of a dehydron as a backbone hydrogen bond satisfying (8.12).

8.7 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 8.6, an attempt [139] 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 [130]. 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 [86, 87, 88, 129, 134, 290] or other site of interest. The residence times are distincly different for well wrapped and underwrapped hydrogen bonds [86]. Thus such residence times may be used as predictors of sites of interest [134, 290]. 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 [359].

8.8 Exercises

Exercise 8.1 *It was predicted [131] that the three proteins*

- *anti-oncogene A (PDB file 1A1U);*
- *RADR zinc finger peptide (PDB file 1A1K) and*
- *rubredoxin (PDB file 1B20).*

might have amyloidogenic tendencies. Investigate these three proteins to see why this might be the case.

Exercise 8.2 *In Section 5.2.4, we noted that sidechains have different conformations. Determine the number of different rotameric states possible for each sidechain (hint: read [261]). Compare the number of rotameric degrees of freedom for the seven residues listed in Figure 4.4(a) with the remaining group of thirteen sidechains.*

Exercise 8.3 *In Figure 8.2(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 *Explain whether you would expect methyl flouride to have a polar carbon, based on the electronegativity scale.*

Exercise 8.5 *Determine the extent to which the wrapping in a given protein is bimodal, as depicted in Figure 8.3. 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.*

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 [138]. A second presents data on the direct measurement of the dehydronic force using atomic force microscopy [122]. Another paper examines the effect of such a force on a deformable surface [140].

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 5.3. 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 [126] 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 [126] 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 [138]. 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 [138], 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 [138]. The particular surface was a Langmuir-Blodgett bilayer made of the lipid DLPC (1,2 dilauroyl-sn-glycero-3 phosphatidylcholine) [352]. We now review the model used in [138] to interpret the data.

9.2 A two-zone model

In [138], 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 [138]. 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

$$P(n_{UW}, n_W, T) = \frac{f e^{\Delta U/k_B T}}{(1-f)e^{\Delta V/k_B T} + f e^{\Delta U/k_B T}} \approx \frac{n_{UW} e^{\Delta U/k_B T}}{n_W e^{\Delta V/k_B T} + n_{UW} e^{\Delta U/k_B T}}, \tag{9.3}$$

with k_B = Boltzmann's constant. In [138], ΔV was assumed to be zero. In this case, (9.3) simplifies to

$$P(n_{UW}, n_W, T) = \frac{f e^{\Delta U/k_B T}}{(1-f) + f e^{\Delta U/k_B T}} \approx \frac{n_{UW} e^{\Delta U/k_B T}}{n_W + n_{UW} e^{\Delta U/k_B T}} \tag{9.4}$$

(cf. equation (2) of [138]). 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 [138] 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** [219] or **Stokes radius** [183] 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 [102, 233] is very accurate, since the variation in flow characteristics due to shape variation is quite small [219]. 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 [3]

$$\begin{aligned} \Pi(v,R,m) &= \int_\Lambda \int_{\Omega\setminus\Lambda} \int_{[0,\tau]} \frac{\alpha L^{-2}}{\pi\Gamma(t)} e^{-\alpha L^{-2}|\mathbf{r}-\mathbf{r}_0|^2/\Gamma(t)}\, dt d\mathbf{r}_0 d\mathbf{r} \\ &= \int_{\tilde\Lambda} \int_{\tilde\Omega\setminus\tilde\Lambda} \int_{[0,L/v]} \frac{\alpha}{\pi\Gamma(t)} e^{-\alpha|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}_0|^2/\Gamma(t)}\, dt d\tilde{\mathbf{r}}_0 d\tilde{\mathbf{r}} \end{aligned} \tag{9.7}$$

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 $6\text{Å}\times 10^8\text{Å}$ 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 [409].

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 [138]) 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 [138] 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 [122] 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 [122], the molecule SH-$(CH_2)_{11}$-Cl was chosen.

The force-displacement curve provided by the AFM have similarities for both monolayers [122]. 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 [122] 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 [138]. Part of the discrepancy could be explained by the fact that in [138] 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 [138] 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 [372] of the inner membranes of cellular or subcellular compartments containing soluble proteins [140]. These vary from simple bag-like membranes [110] (e.g., erythrocytes, a.k.a. red blood cells) to highly invaginated membranes [414] (e.g., mitochondrial inner membranes). This raises the question of what might be causing the difference in membrane structure [229, 251, 301, 417].

Some evidence [140] 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 [140].

9.4.1 Protein adsorption

Morphology induction was tested in fluid phospholipid (DLPC) bilayers (Section 9.1) coating an optical waveguide [140]. 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 [348, 394]. 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 [140]). 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 [140]) shows that protein aggregation is a competing effect in the protection of solvent-exposed hydrogen bonds ([137, 125, 126, 119, 145]): 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 [126]. 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 [140]). 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 [140] 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 [140]) 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.

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 force on a particular entity by all

charge groups	force name	power law	equation
charge-charge	salt bridge	r^{-1}	(3.1)
charge-dipole	Keesom force	$\cos\theta\, r^{-2}$	(10.23)
dipole-dipole	polar bond	r^{-3}	(3.11),(10.21)
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.

the others, and suppose this force is proportional to r^{-n} for some n. Summing over all space, we determine the total force. We can estimate this sum by computing sums over expanding spherical shell sets $\left\{\mathbf{r} \in \mathbb{R}^3 \;\middle|\; R-1 \le |\mathbf{r}| < R\right\}$ for $R = 1, 2, 3, \ldots$. In each spherical shell region, the sum of all forces, 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 force would be proportional to

$$\sum_{R=1}^{R_{\mathrm{mac}}} R^{2-n} \tag{10.1}$$

which is divergent (as R_{mac} increases) for $n \le 3$.

Note that R_{mac} is the size of a macroscopic system in microscopic units, so it is related to Avogadro's number, hence should be viewed as nearly infinite. The borderline case $n = 3$, for which the divergence is only logarithmic, corresponds to 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. 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

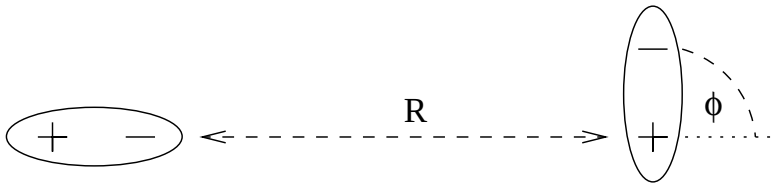


Figure 10.1: Dipole-dipole interaction: rotated in-line configuration. The positive (resp., negative) charge center of the dipole on the right is at $(0, 0)$ (resp., at $(\cos\phi, \sin\phi)$). The positive (resp., negative) charge center of the dipole on the left is at $(-R - 1, 0)$ (resp., at $(-R, 0)$).

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 10.1).

10.2.1 Dipole-dipole interactions

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.

In-line interaction configuration

Suppose we have two dipoles as indicated in Figure 10.1. 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

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R, 0)| &= \sqrt{(R + \cos\phi)^2 + \sin^2\phi} \\ &= \sqrt{1 + R^2 + 2R\cos\phi}, \end{aligned} \tag{10.2}$$

and the distance between the positive charge on the left and the negative charge on the right is

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R - 1, 0)| &= \sqrt{(1 + R + \cos\phi)^2 + \sin^2\phi} \\ &= \sqrt{1 + (R + 1)^2 + 2(R + 1)\cos\phi}. \end{aligned} \tag{10.3}$$

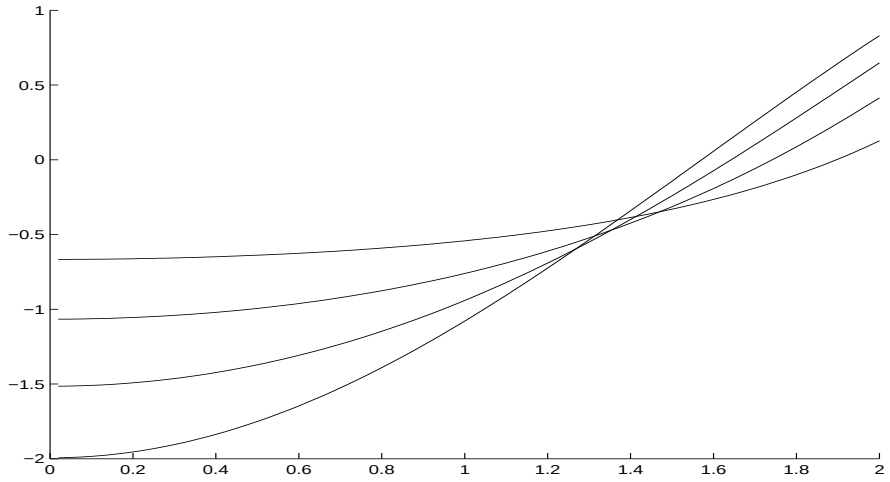


Figure 10.2: Dipole-dipole (in-line) interaction energy (vertical axis), scaled by R^3, for $R = 2, 4, 10, 1000$. The energies correspond to the dipole-dipole configuration in Figure 10.1. Horizontal ϕ-axis measured in radians. The flattest curve corresponds to $R = 2$.

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$\frac{1}{R+1} - \frac{1}{R} + \frac{1}{\sqrt{1+R^2+2R\cos\phi}} - \frac{1}{\sqrt{1+(R+1)^2+2(R+1)\cos\phi}}. \tag{10.4}$$

A plot of the interaction energy (10.4) is given in Figure 10.2 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 10.2 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

modeling a hydrogen bond using a simple dipole-dipole
interaction does not yield a very strong angular dependence.

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 10.3. 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).

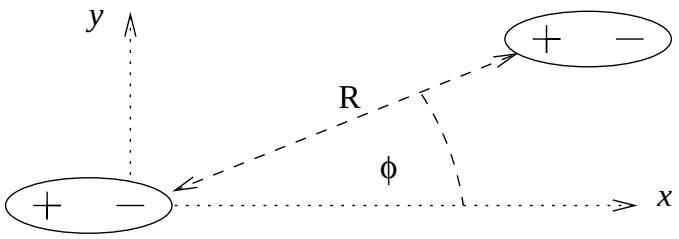


Figure 10.3: Dipole-dipole interaction: rotated parallel configuration. The negative (resp., positive) charge center of the dipole on the left is at $(0, 0)$ (resp., at $(-1, 0)$). The positive (resp., negative) charge center of the dipole on the right is at $R(\cos\phi, \sin\phi)$ (resp., at $(1 + R\cos\phi, \sin\phi)$).

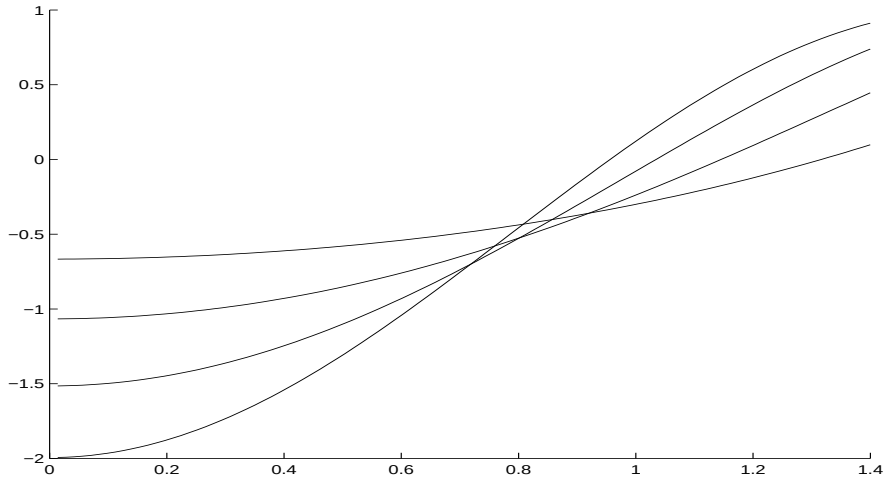


Figure 10.4: Dipole-dipole (parallel) interaction energy (vertical axis), scaled by R^3, for $R = 2, 4, 10, 1000$. The energies correspond to the dipole-dipole configuration in Figure 10.3. Horizontal ϕ-axis measured in radians.

The distance between the positive charges is the same as the distance between the negative charges because the dipoles are parallel:

$$|(1 + R\cos\phi, R\sin\phi)| = \sqrt{1 + R^2 + 2R\cos\phi}. \tag{10.5}$$

Similarly, the distance between the positive charge on the left and the negative charge on the right is

$$|(2 + R\cos\phi, R\sin\phi)| = \sqrt{4 + R^2 + 4R\cos\phi}. \tag{10.6}$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$-\frac{1}{R} + \frac{2}{\sqrt{1 + R^2 + 2R\cos\phi}} - \frac{1}{\sqrt{4 + R^2 + 4R\cos\phi}}. \tag{10.7}$$

A plot of the interaction energy (10.7) 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 10.2 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 10.5. 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

$$|(1 + R\cos\theta, R\sin\theta)| = \sqrt{1 + R^2 + 2R\cos\theta}. \tag{10.8}$$

The separation between the positive charge on the right and the negative charge on the left is

$$\begin{aligned} |R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi)| &= \sqrt{(R\cos\theta + \cos\phi)^2 + (R\sin\theta + \sin\phi)^2} \\ &= \sqrt{(R^2 + 1 + 2R(\cos\theta\cos\phi + \sin\theta\sin\phi)}. \end{aligned} \tag{10.9}$$

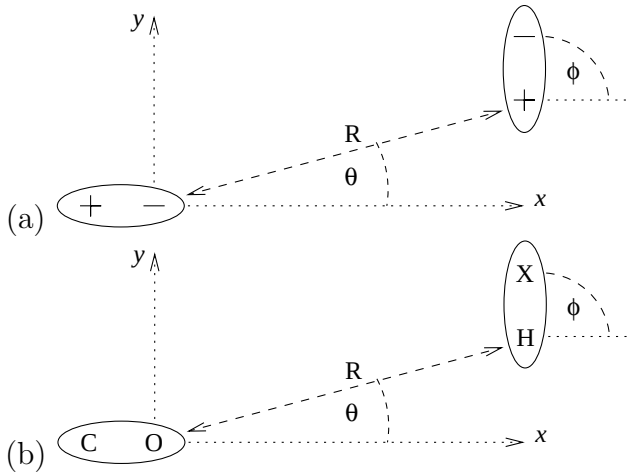


Figure 10.5: Dipole-dipole (two-angle) interaction configuration. (a) The abstract case. The negative (resp., positive) charge center of the dipole on the left is at $(0, 0)$ (resp., at $(-1, 0)$). The positive (resp., negative) charge center of the dipole on the right is at $R(\cos\theta, \sin\theta)$ (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).

Finally, the distance (squared) between the positive charge on the left and the negative charge on the right is

$$\begin{aligned} |R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi) - (-1, 0)|^2 &= |(1 + R\cos\theta + \cos\phi, R\sin\theta + \sin\phi)|^2 \\ &= (1 + \cos\phi)^2 + 2R\cos\theta(1 + \cos\phi) + R^2 + 2R\sin\theta\sin\phi + \sin^2\phi \\ &= 2(1 + \cos\phi) + 2R\cos\theta(1 + \cos\phi) + R^2 + 2R\sin\theta\sin\phi \\ &= 2(1 + \cos\phi)(1 + R\cos\theta) + R^2 + 2R\sin\theta\sin\phi. \end{aligned} \tag{10.10}$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$\begin{aligned} &-\frac{1}{R} + \frac{1}{\sqrt{1 + R^2 + 2R\cos\theta}} + \frac{1}{\sqrt{1 + R^2 + 2R(\cos\theta\cos\phi + \sin\theta\sin\phi)}} \\ &-\frac{1}{\sqrt{R^2 + 2(1 + \cos\phi)(1 + R\cos\theta) + 2R\sin\theta\sin\phi}}. \end{aligned} \tag{10.11}$$

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

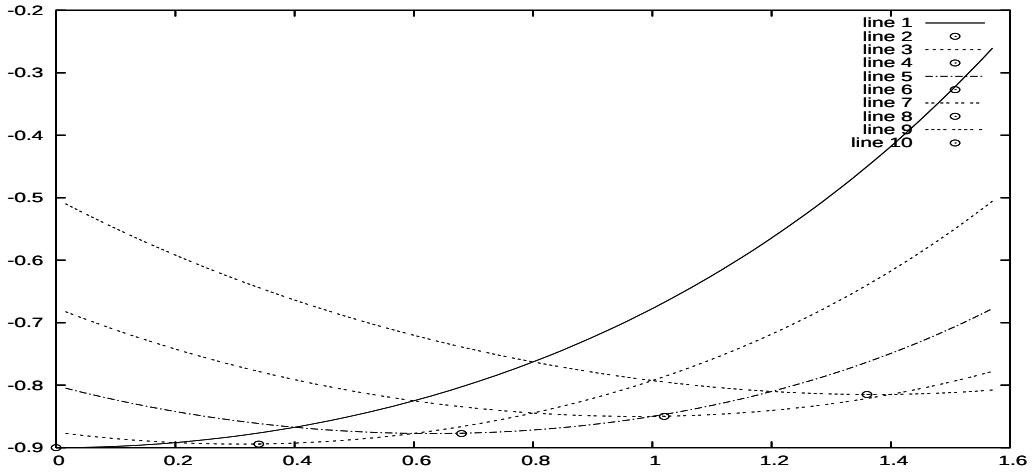


Figure 10.6: Dipole-dipole (two-angle) interaction energy (vertical axis), scaled by R^3, for $R = 3$, as a function of ϕ for various fixed values of $\theta = 0, 0.2, 0.4, 0.6, 0.8$. Approximate minimum values of the energy are indicated by circles at the points $\phi = 1.7\,\theta$. The horizontal ϕ-axis is measured in radians.

a function of ϕ for various values of θ in Figure 10.6. As an aid to the eye, we plot a circle at a 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 complex.

In the case that $\theta = \pi/2$, the expression (10.11) simplifies to

$$-\frac{1}{R} + \frac{1}{\sqrt{1+R^2}} - \frac{1}{\sqrt{R^2 + 2(1+\cos\phi) + 2R\sin\phi}} + \frac{1}{\sqrt{1+R^2+2R\sin\phi}}. \tag{10.12}$$

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 (10.12) does not occur at $\phi = \pi$, due to the asymmetry of the expression around this value. We leave as Exercise 10.4 to plot (10.12) 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 10.7, 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 $R \to \infty$.

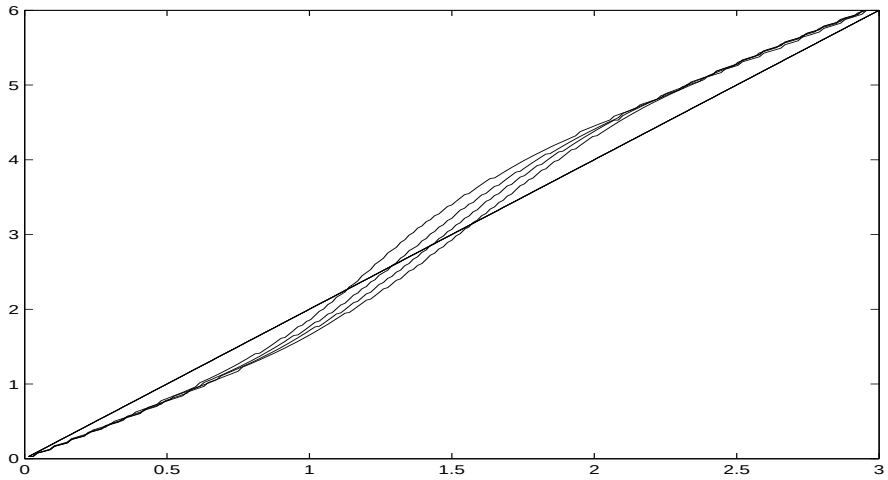


Figure 10.7: Optimal ϕ angle (vertical axis) corresponding to the minimum interaction energy as a function of θ (horizontal axis) for the dipole-dipole (two-angle) interaction, for $R = 3, 5, 10, 1000$ (the left-most curve corresponds to $R = 3$, and they move to the right with increasing R). Both the 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 10.8. 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).

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 $(r, 0, 0)$. Thus assume that there is are charges ± 1 at $\pm\mathbf{u}$, and correspondingly charges ± 1 at points $\mathbf{r} = (r, 0, 0) \pm \mathbf{v}$ 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

$$V(r, \mathbf{u}, \mathbf{v}) = \frac{1}{|\mathbf{u} - \mathbf{r} - \mathbf{v}|} + \frac{1}{|-\mathbf{u} - \mathbf{r} + \mathbf{v}|} - \frac{1}{|\mathbf{u} - \mathbf{r} + \mathbf{v}|} - \frac{1}{|-\mathbf{u} - \mathbf{r} - \mathbf{v}|}. \tag{10.13}$$

We use the relation

$$|\mathbf{w} + \mathbf{x} + \mathbf{y}|^2 = |\mathbf{w}|^2 + |\mathbf{x}|^2 + |\mathbf{y}|^2 + 2(\mathbf{w} \cdot \mathbf{x} + \mathbf{w} \cdot \mathbf{y} + \mathbf{y} \cdot \mathbf{x}) \tag{10.14}$$

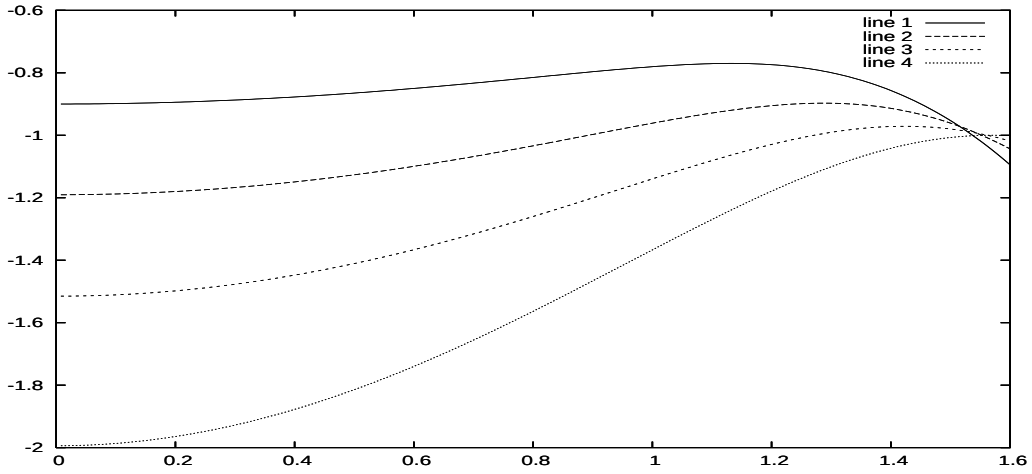


Figure 10.8: Dipole-dipole (two-angle) interaction energy minimum, scaled by R^3, for $R = 3, 5, 10, 1000$ (top to bottom), as a function of θ. Plotted is the energy (vertical axis) at the optimal value of ϕ that minimizes the energy as a function of ϕ for fixed θ. Horizontal θ-axis measured in radians.

to get

$$\begin{aligned} V(r,\mathbf{u},\mathbf{v}) &= \frac{1}{\sqrt{2+r^2-2\mathbf{u}\cdot\mathbf{v}-2r(u_1-v_1)}} + \frac{1}{\sqrt{2+r^2-2\mathbf{u}\cdot\mathbf{v}+2r(u_1-v_1)}} \\ &\quad - \frac{1}{\sqrt{2+r^2+2\mathbf{u}\cdot\mathbf{v}-2r(u_1+v_1)}} - \frac{1}{\sqrt{2+r^2+2\mathbf{u}\cdot\mathbf{v}+2r(u_1+v_1)}} \\ &= \frac{1}{\sqrt{2+r^2}}\left(\frac{1}{\sqrt{1-a-b}} + \frac{1}{\sqrt{1-a+b}} - \frac{1}{\sqrt{1+a-c}} - \frac{1}{\sqrt{1+a+c}}\right) \end{aligned} \tag{10.15}$$

where

$$a = \frac{2\mathbf{u}\cdot\mathbf{v}}{2+r^2}, \quad b = \frac{2r(u_1-v_1)}{2+r^2} = \frac{2\mathbf{r}\cdot(\mathbf{u}-\mathbf{v})}{2+r^2}, \quad \text{and} \quad c = \frac{2r(u_1+v_1)}{2+r^2} = \frac{2\mathbf{r}\cdot(\mathbf{u}+\mathbf{v})}{2+r^2}. \tag{10.16}$$

Expanding using the expression

$$(1-\epsilon)^{-\frac{1}{2}} = 1 + \tfrac{1}{2}\epsilon + \tfrac{3}{8}\epsilon^2 + \tfrac{5}{16}\epsilon^3 + \mathcal{O}\left(\epsilon^4\right), \tag{10.17}$$

we find for large r that

$$\begin{aligned} V(r,\mathbf{u},\mathbf{v})\sqrt{2+r^2} &= \frac{1}{\sqrt{1-a-b}} + \frac{1}{\sqrt{1-a+b}} - \frac{1}{\sqrt{1+a-c}} - \frac{1}{\sqrt{1+a+c}} \\ &\approx \tfrac{1}{2}\left(a+b+a-b+a-c+a+c\right) \\ &\quad + \tfrac{3}{8}\left((a+b)^2+(a-b)^2-(a-c)^2-(a+c)^2\right) \\ &\quad + \tfrac{5}{16}\left((a+b)^3+(a-b)^3+(a-c)^3+(a+c)^3\right) + \mathcal{O}\left(r^{-4}\right) \\ &= 2a + \tfrac{3}{4}\left(b^2-c^2\right) + \tfrac{5}{16}\left(4a^3+2a(b^2+c^2)\right) + \mathcal{O}\left(r^{-4}\right) \\ &= 2a + \tfrac{3}{4}\left(b^2-c^2\right) + \mathcal{O}\left(r^{-4}\right) \end{aligned} \tag{10.18}$$

because $a = \mathcal{O}(r^{-2})$, $b = \mathcal{O}(r^{-1})$, and $c = \mathcal{O}(r^{-1})$. But

$$b^2 - c^2 = \frac{4r^2}{(2+r^2)^2}\left((u_1 - v_1)^2 - (u_1 + v_1)^2\right) = \frac{-16u_1v_1r^2}{(2+r^2)^2} = \frac{-16(\mathbf{u}\cdot\mathbf{r})(\mathbf{v}\cdot\mathbf{r})}{(2+r^2)^2} \tag{10.19}$$

Therefore

$$V(r,\mathbf{u},\mathbf{v})\sqrt{2+r^2} = \frac{4\mathbf{u}\cdot\mathbf{v}}{2+r^2} - \frac{12(\mathbf{r}\cdot\mathbf{u})(\mathbf{r}\cdot\mathbf{v})}{(2+r^2)^2} + \mathcal{O}\left(r^{-4}\right). \tag{10.20}$$

Thus

$$V(r,\mathbf{u},\mathbf{v}) \approx \frac{4\mathbf{u}\cdot\mathbf{v} - 12(\mathbf{r}\cdot\mathbf{u})(\mathbf{r}\cdot\mathbf{v})}{r^3}. \tag{10.21}$$

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.

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 10.9. 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

$$V(r,\theta) = -\frac{a}{\sqrt{(1 - r\cos\theta)^2 + r^2\sin^2\theta}} + \frac{a}{\sqrt{(1 + r\cos\theta)^2 + r^2\sin^2\theta}}. \tag{10.22}$$

We leave as Exercise 10.8 to show that

$$V(r,\theta) \approx -\frac{2a\cos\theta}{r^2} \tag{10.23}$$

for large r and fixed θ.

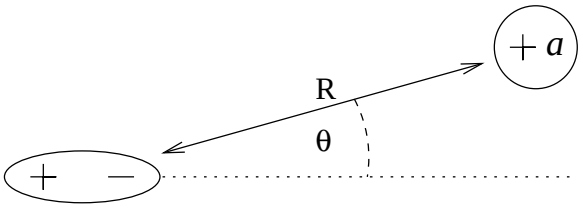


Figure 10.9: Charge-dipole interaction configuration. The dipole on the left has its negative charge at the origin and its positive charge center at $(-1, 0, 0)$. The positive charge $(+a)$ center on the right is at $R(\cos\theta, \sin\theta, 0)$.

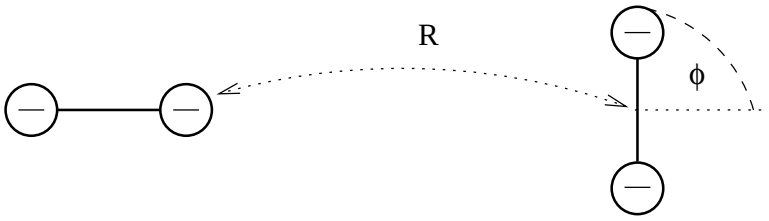


Figure 10.10: Charge-charge interaction configuration similar to what is found in an interaction between Asp and Glu. The pair of negative charges on the right is at $\pm(\cos\phi, \sin\phi)$ and the pair of negative charges on the left is at $(R \pm 1, 0)$.

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 10.10. 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

$$\begin{aligned} r_{++} &= |+(\cos\phi, \sin\phi) - (R+1, 0)|, \\ r_{-+} &= |-(\cos\phi, \sin\phi) - (R+1, 0)|, \\ r_{+-} &= |+(\cos\phi, \sin\phi) - (R-1, 0)|, \quad \text{and} \\ r_{--} &= |-(\cos\phi, \sin\phi) - (R-1, 0)|. \end{aligned} \tag{10.24}$$

With the denotations $C = \cos\phi, S = \sin\phi$, we find

$$\begin{aligned} r_{++}^2 &= (C - R - 1)^2 + S^2 = 2 + R^2 - 2\cos\phi(R+1) + 2R, \\ r_{-+}^2 &= (C + R + 1)^2 + S^2 = 2 + R^2 + 2\cos\phi(R+1) + 2R, \\ r_{+-}^2 &= (C - R + 1)^2 + S^2 = 2 + R^2 - 2\cos\phi(R-1) - 2R, \quad \text{and} \\ r_{--}^2 &= (C + R - 1)^2 + S^2 = 2 + R^2 + 2\cos\phi(R-1) - 2R. \end{aligned} \tag{10.25}$$

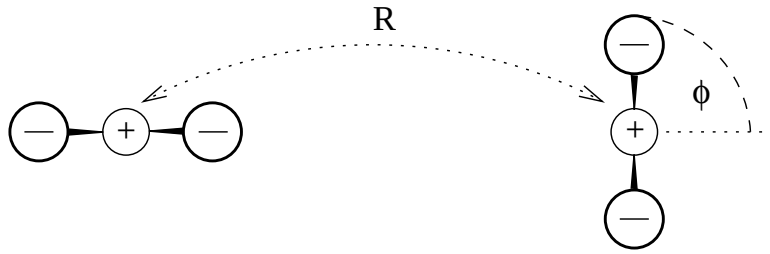


Figure 10.11: Charge-charge interaction configuration similar to what is found in an interaction between Asp and Glu, but with a more refined model. The pair of negative charges on the right are at $(\cos\phi, \sin\phi, 0)$ and $(-\cos\phi, -\sin\phi, 0)$ and the associated positive charge is at $(0, 0, -z)$. The pair of negative charges on the left are at $(R \pm 1, 0, 0)$, and the associated positive charge is at $(R, 0, -z)$.

We can write these expressions succinctly as

$$r_{\pm_1\pm_2} = \sqrt{2 + R^2 \pm_1 2(\cos\phi(R \pm_2 1)) \pm_2 2R}\,. \tag{10.26}$$

The energy (of repulsion) for the charge groups is

$$\frac{1}{r_{++}} + \frac{1}{r_{-+}} + \frac{1}{r_{+-}} + \frac{1}{r_{--}}, \tag{10.27}$$

and we seek to find the value of ϕ that minimizes it. We leave as Exercise 10.5 to plot the expression in (10.27) 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 10.11. 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 10.12, 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

$$V(\mathbf{r}) = \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r} - \mathbf{r}_k|}, \tag{10.28}$$

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

$$W(\mathbf{r}) = 1/|\mathbf{r}| \tag{10.29}$$

as follows. Define a translation operator $T_{\mathbf{x}}$ by

$$(T_{\mathbf{x}} f)(\mathbf{r}) = f(\mathbf{r} - \mathbf{x}) \tag{10.30}$$

for any function f. Then we can interpret the expression (10.28) as

$$V = \sum_{k=1}^{K} q_k T_{\mathbf{r}_k} W, \tag{10.31}$$

where W is defined in (10.29). In view of (10.31), we define the operator

$$\mathcal{D} = \sum_{k=1}^{K} q_k T_{\mathbf{r}_k}. \tag{10.32}$$

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 such as in Section 18.5.1 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

$$V_\epsilon(\mathbf{r}) = \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r} - \epsilon \mathbf{r}_k|}, \tag{10.33}$$

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:

$$V_\epsilon(\mathbf{r}) = |\mathbf{r}|^{-1} V_{\epsilon/|\mathbf{r}|}\left(|\mathbf{r}|^{-1}\mathbf{r}\right). \tag{10.34}$$

The proof just requires changing variables in (10.33):

$$V_\epsilon(\mathbf{r}) = \frac{1}{|\mathbf{r}|} \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r}|^{-1}|\mathbf{r} - \epsilon \mathbf{r}_k|} = \frac{1}{|\mathbf{r}|} \sum_{k=1}^{K} \frac{q_k}{|(|\mathbf{r}|^{-1}\mathbf{r}) - (\epsilon/|\mathbf{r}|)\mathbf{r}_k|} = |\mathbf{r}|^{-1} V_{\epsilon/|\mathbf{r}|}\left(|\mathbf{r}|^{-1}\mathbf{r}\right). \tag{10.35}$$

Given a general potential V of the form (10.28), we can think of this as having $\epsilon = 1$, that is, $V = V_1$. Using (10.35), we can derive the asymptotic form

$$V(\mathbf{r}) = V_1(\mathbf{r}) = |\mathbf{r}|^{-1} V_{|\mathbf{r}|^{-1}}\left(|\mathbf{r}|^{-1}\mathbf{r}\right) = \epsilon V_\epsilon(\omega), \tag{10.36}$$

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 (10.32), we define

$$\mathcal{D}_\epsilon = \sum_{k=1}^{K} q_k T_{\epsilon \mathbf{r}_k}, \tag{10.37}$$

and observe from (10.28) and (10.29) that

$$V_\epsilon = \mathcal{D}_\epsilon W. \tag{10.38}$$

We will see in typical cases that, for some $k \geq 0$,

$$\lim_{\epsilon \to 0} \epsilon^{-k} \mathcal{D}_\epsilon = \mathcal{D}_0, \tag{10.39}$$

where $\mathcal{D}_0$ is a differential operator of order k. The convergence in (10.39) is (at least) weak convergence, in the sense that for any smooth function f in a region $\Omega \subset \mathbb{R}^3$,

$$\lim_{\epsilon \to 0} \epsilon^{-k} \mathcal{D}_\epsilon f(\mathbf{x}) = \mathcal{D}_0 f(\mathbf{x}) \tag{10.40}$$

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

$$\lim_{\epsilon \to 0} \epsilon^{-k} V_\epsilon(\omega) = \mathcal{V}(\omega), \qquad |\omega| = 1, \tag{10.41}$$

where the limiting potential is defined by

$$\mathcal{V}(\mathbf{r}) = \mathcal{D}_0 W(\mathbf{r}). \tag{10.42}$$

Applying (10.36), (10.41), and (10.42), we conclude that

$$V(\mathbf{r}) \approx \frac{1}{|\mathbf{r}|^{k+1}} \mathcal{D}_0 W(|\mathbf{r}|^{-1}\mathbf{r}), \tag{10.43}$$

for large $\mathbf{r}$. More precisely, we will typically show that

$$\epsilon^{-k} \mathcal{D}_\epsilon \phi(\mathbf{r}) = \mathcal{D}_0 \phi(\mathbf{r}) + \mathcal{O}(\epsilon) \tag{10.44}$$

in which case we can assert that

$$V(\mathbf{r}) = \frac{1}{|\mathbf{r}|^{k+1}} \mathcal{D}_0 W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-k-2}). \tag{10.45}$$

10.4.2 Application of (10.45)

Let us show how (10.45) 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

$$\mathcal{D}_\epsilon \phi(x, y, z) = \phi(x + \tfrac{1}{2}\epsilon, y, z) - \phi(x - \tfrac{1}{2}\epsilon, y, z). \tag{10.46}$$

By Taylor's theorem, we can expand a function ψ to show that

$$\psi(x + \xi) - \psi(x - \xi) = 2\xi\psi'(x) + \tfrac{1}{3}\xi^3\psi^{(3)}(x) + \mathcal{O}(\xi^5). \tag{10.47}$$

Applying (10.47) to $\psi(x) = \phi(x, y, z)$, we have

$$\mathcal{D}_\epsilon \phi(x, y, z) = \epsilon \frac{\partial}{\partial x}\phi(x, y, z) + \mathcal{O}(\epsilon^3). \tag{10.48}$$

Taking limits, we see that

$$\epsilon^{-1}\mathcal{D}_\epsilon \to \frac{\partial}{\partial x} \tag{10.49}$$

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 (10.45) we have

$$V(\mathbf{r}) = |\mathbf{r}|^{-2} \left(\frac{\partial}{\partial x} W\right) \left(|\mathbf{r}|^{-1}\mathbf{r}\right) + \mathcal{O}(|\mathbf{r}|^{-3}). \tag{10.50}$$

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.

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.:

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

$$S = 2a + 1. \tag{10.51}$$

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.

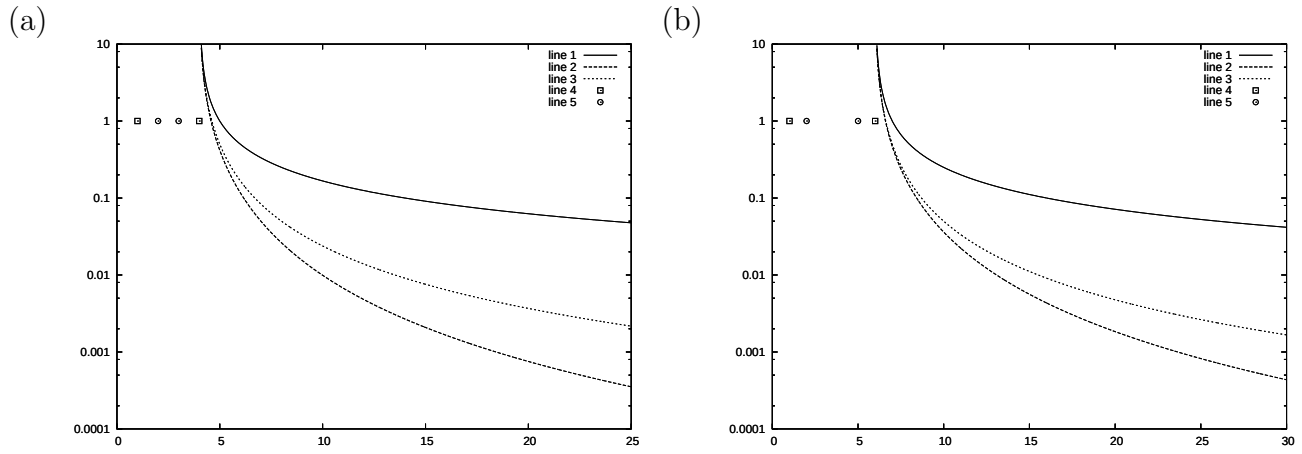


Figure 10.12: Comparison of single charge, dipole and quadrupole potentials. Dipole separation (a) two units and (b) four units. The locations of the negative charges are indicated by circles and the locations of the positive charges are indicated by squares. The upper solid line is the potential for 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.

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

$$\mathcal{D}_\epsilon = T_{\epsilon(a+1,0,0)} - T_{\epsilon(a,0,0)} + T_{\epsilon(-a-1,0,0)} - T_{\epsilon(-a,0,0)}. \tag{10.52}$$

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 (10.47) to ϕ and ϕ' we find that

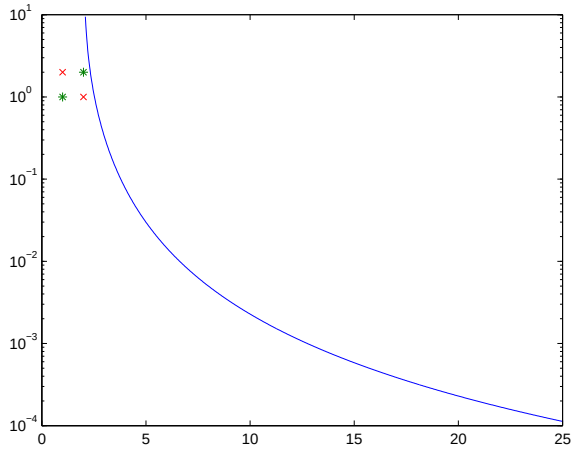
$$\begin{aligned}\mathcal{D}_\epsilon\phi(x) =&\phi(x-\epsilon(a+1)) - \phi(x-\epsilon a) + \phi(x+\epsilon(a+1)) - \phi(x+\epsilon a)\\ =&\epsilon\phi'(x-\epsilon(a+\tfrac{1}{2})) - \epsilon\phi'(x+\epsilon(a+\tfrac{1}{2})) + \mathcal{O}(\epsilon^3)\\ =&\epsilon^2(2a+1)\phi''(x) + \mathcal{O}(\epsilon^3)\\ =&\epsilon^2 S\phi''(x) + \mathcal{O}(\epsilon^3),\end{aligned} \tag{10.53}$$

where S is the separation distance between the dipoles. Thus

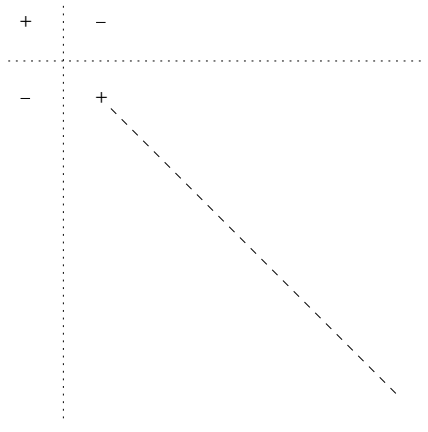
$$\lim_{\epsilon\to 0}\epsilon^{-2}\mathcal{D}_\epsilon = (2a+1)\frac{\partial^2}{\partial x^2} = S\frac{\partial^2}{\partial x^2}, \tag{10.54}$$

where $S = 2a+1$ is the separation distance (10.51) between the dipoles. Applying (10.45), we find

$$V(\mathbf{r}) = |\mathbf{r}|^{-3} S\frac{\partial^2}{\partial x^2} W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-4}) \tag{10.55}$$



(a)



(b)

Figure 10.13: 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.

for large $\mathbf{r}$, where W $(= 1/r)$ is defined in (10.29).

The potential for opposing dipoles is depicted in Figure 10.12 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:

\+ –

– +

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 10.13. Note that it dies off a bit more rapidly than

the potential for the opposing dipoles (cf. Figure 10.12). Defining

$$\mathcal{D}_\epsilon = \sum_{k=1}^{K} q_k T_{\epsilon r_k} \tag{10.56}$$

and applying (10.47) twice, we see that

$$\begin{aligned} \mathcal{D}_\epsilon \phi(\mathbf{r}) =& \sum_{k=1}^{K} q_k \phi(\mathbf{r} - \epsilon \mathbf{r}_k) \\ =& 4 \frac{\partial}{\partial x} \frac{\partial}{\partial y} \phi(0) \epsilon^2 + \mathcal{O}(\epsilon^3). \end{aligned} \tag{10.57}$$

Thus

$$V(\mathbf{r}) = |\mathbf{r}|^{-3} 4 \frac{\partial}{\partial x} \frac{\partial}{\partial y} W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-4}). \tag{10.58}$$

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 [143]. This is shown in Figure 10.14. 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 10.14(b) shows the skewness of the two 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.

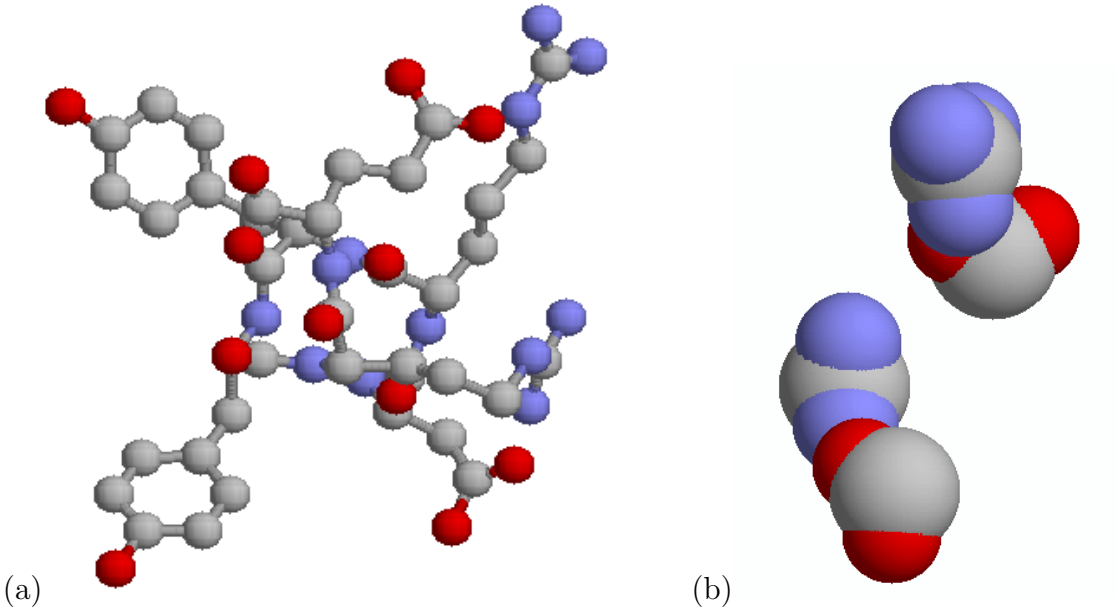
We can approximate the electronic structure of water as system with positive charges

$$q_1 = q_2 = a \;\text{ at }\; \mathbf{r}_{1,2} = (\pm c, -1, 0) \tag{10.59}$$

and negative charges

$$q_3 = q_4 = -b \;\text{ at }\; \mathbf{r}_{3.4} = (0, y^0, \pm d), \tag{10.60}$$

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 [270], 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.



(a) (b)

Figure 10.14: 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.

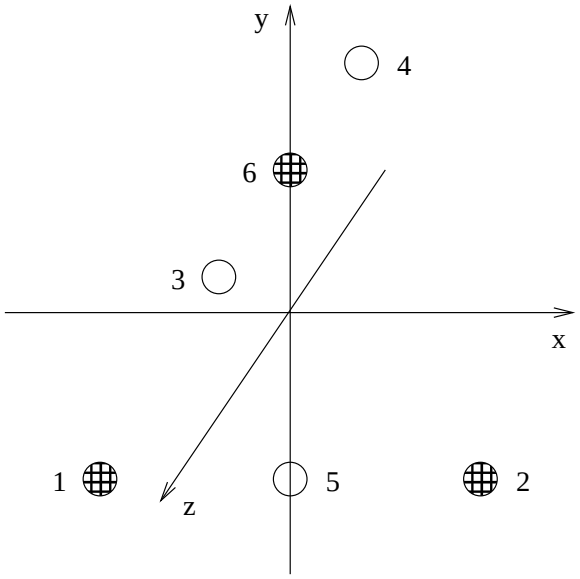


Figure 10.15: Configuration of charges in water model. Open circles indicate negative charge locations; shaded circles indicate locations of positive charge.

To discover the exact multipole nature of the water model encoded in (10.59) and (10.60), we modify it to form a quadrupole. We extend the system (10.59–10.60) to involve two more charges:

$$\begin{aligned} q_5 = -2a \ \text{ at } \ \mathbf{r}_5 = (0,-1,0) \ \text{ and} \\ q_6 = 2b \ \text{ at } \ \mathbf{r}_6 = (0,y_0,0). \end{aligned} \tag{10.61}$$

The configuration of charges is depicted in Figure 10.15.

The extended system (10.59),(10.60),(10.61) is a quadrupole due to the cancellations leading to an expression such as (10.57). More precisely, note that the charges at locations 1, 2 and 5 correspond to a second difference stencil centered at point 5 for approximating

$$\frac{\partial^2 \phi}{\partial x^2}(0,-1,0) \tag{10.62}$$

(with suitable scaling). Similarly, the charges at locations 3, 4 and 6 correspond to a second difference stencil centered at point 6 for approximating

$$\frac{\partial^2 \phi}{\partial z^2}(0,y^0,0) \tag{10.63}$$

(with suitable scaling). Therefore

$$\begin{aligned} \mathcal{D}_\epsilon \phi(0) =& \sum_{k=1}^{6} q_k \phi(\epsilon \mathbf{r}_k) \\ =& ac^2\epsilon^2 \frac{\partial^2 \phi}{\partial x^2}(0,-1,0) - bd^2\epsilon^2 \frac{\partial^2 \phi}{\partial z^2}(0,y^0,0)\epsilon^2 + \mathcal{O}(\epsilon^4), \end{aligned} \tag{10.64}$$

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 (10.61). 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 (10.64), and let V^W be the water potential using the model (10.59–10.60). Thus we have written the water potential as

$$V^W = V^D + V^Q \tag{10.65}$$

for an explicit dipole potential V^D, with charges at $\mathbf{r}_5$ and $\mathbf{r}_6$, and a quadrupole. Thus the water model (10.59–10.60) 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 (10.45) that the order of decay of the potential can be determined by the arrangement of the charges (10.37). When the net charge is non-zero, we have $k = 0$, but when the net charge is zero

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.

then $k \geq 1$. The dipole is the case $k = 1$, and $k = 2$ is called a quadrupole. Similarly, $k = 3$ is 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 14.7.2), and the polarizability $\alpha \approx 1.2$Å^3. Thus an electric field strength of only one-tenth of an electron per square Ångstrom (0.1e-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

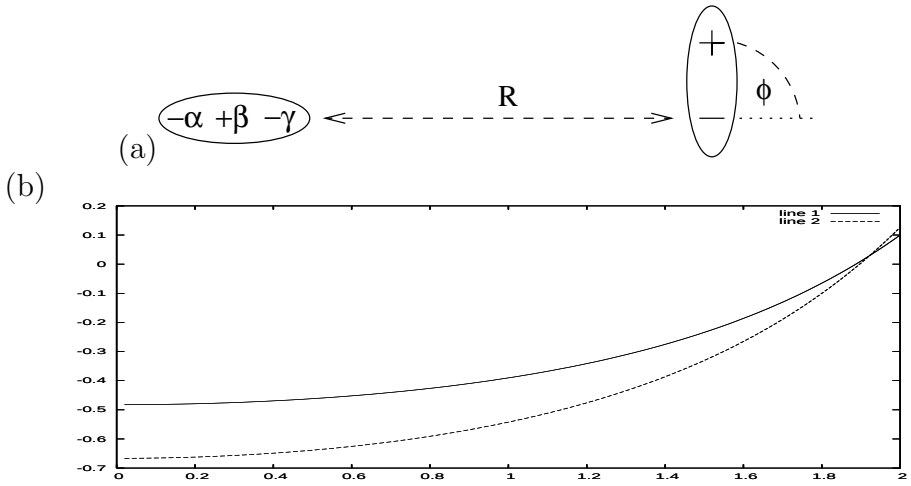


Figure 10.16: (a) A modified model for dipole-dipole interaction. The negative (resp., positive) charge center of the dipole on the right is at $(0,0)$ (resp., at $(\cos\phi, \sin\phi)$). The charge centers of the multipole on the left are at $(-R-1,0)$ (negative charge $-\alpha$), $(-R,0)$ (positive 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 10.1. 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 10.1.

position of the negative charge on the right we take as the origin, and we assume the separation 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 10.1 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

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R+\delta, 0)| &= \sqrt{(R-\delta+\cos\phi)^2 + \sin^2\phi} \\ &= \sqrt{1 + (R-\delta)^2 + 2(R-\delta)\cos\phi}. \end{aligned} \tag{10.66}$$

The distance between the two positive charges is

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R, 0)| &= \sqrt{(R+\cos\phi)^2 + \sin^2\phi} \\ &= \sqrt{1 + R^2 + 2R\cos\phi}, \end{aligned} \tag{10.67}$$

and the distance between the main (α) negative charge on the left and the positive charge on the right is

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R-1, 0)| =& \sqrt{(1+R+\cos\phi)^2 + \sin^2\phi} \\ =& \sqrt{1+(R+1)^2 + 2(R+1)\cos\phi}. \end{aligned} \tag{10.68}$$

Thus the interaction energy for the dipole pair (assuming unit charges) is

$$\begin{aligned} &\frac{\alpha}{R+1} - \frac{\beta}{R} + \frac{\gamma}{R-\delta} - \frac{\gamma}{\sqrt{1+(R-\delta)^2 + 2(R-\delta)\cos\phi}} \\ &+\frac{\beta}{\sqrt{1+R^2+2R\cos\phi}} - \frac{\alpha}{\sqrt{1+(R+1)^2+2(R+1)\cos\phi}}. \end{aligned} \tag{10.69}$$

A plot of the interaction energy (10.69) is given in Figure 10.16 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 6.6. 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 10.17. 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 10.17(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 6.6) 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 10.17, 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

$$\frac{-1}{\sqrt{x_0^2 + (y_0 - \cos\theta)^2 + \sin^2\theta}} + \frac{1}{\sqrt{x_1^2 + (y_1 - \cos\theta)^2 + (z_1 - \sin\theta)^2}} \tag{10.70}$$

For given x_0, y_0, x_1, y_1, z_1, this expression can be minimized to find the optimal θ.

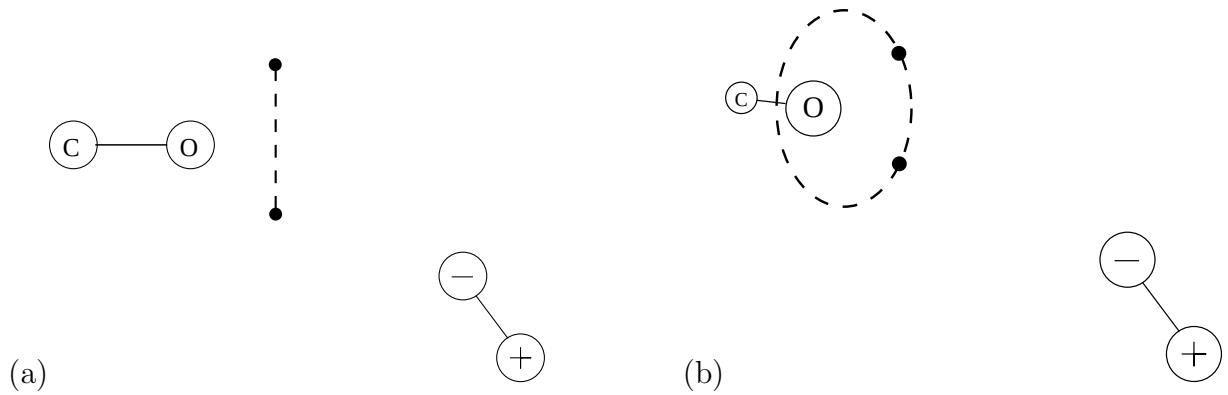


Figure 10.17: Configuration for the placement of hydrogen at the end of the sidechain of serine or threonine in response to a nearby dipole. The dashed line indicates the circle of possible hydrogen placements. (a) The plane of the circle is orthogonal to the plane of the page. (b) The plane of the circle is skew to the plane of the page.

We can also use the expression (**??**) to find the optimum θ. In coordinates determined so that the hydrogen and the dipole lie in a plane, the interaction field (**??**) 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

$$\nabla V(x(\phi), y(\phi), z(\phi)) \cdot (x'(\phi), y'(\phi), z'(\phi)) = 0. \tag{10.71}$$

10.8 Exercises

Exercise 10.1 *Show that the interaction energy (10.4) tends to the asymptotic form*

$$\frac{-2\cos\phi}{R^3}. \tag{10.72}$$

Exercise 10.2 *Show that the interaction energy (10.7) tends to the asymptotic form*

$$\frac{-\frac{1}{2} - \frac{3}{2}\cos\phi}{R^3}. \tag{10.73}$$

Exercise 10.3 *Verify that the second term in the energy expression in (10.11) is indeed the same as (10.5). Also verify that the fourth term in the energy expression in (10.11) is correct.*

Exercise 10.4 *Plot the expression in (10.12) 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 (10.27) 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 (10.27) in the case that the charge group has a positive charge as well as the negative charges, as shown in Figure 10.11. 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 and angle ϕ. Determine the value of ϕ at the minimum.*

Exercise 10.8 *Prove that the asymptotic expression (10.23) is valid for fixed θ and large r. (Hint: show that*

$$V(r,\theta) = \frac{a}{r}\left(\frac{\sqrt{1-2r^{-1}\cos\theta+r^{-2}}-\sqrt{1+2r^{-1}\cos\theta+r^{-2}}}{\sqrt{1-2r^{-1}\cos\theta+r^{-2}}\sqrt{1+2r^{-1}\cos\theta+r^{-2}}}\right) \tag{10.74}$$

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 (10.23) when $\theta = \pi/4$ and $r = 3$.*

Exercise 10.10 *Show that a charge system with only the charges as indicated in (10.61) 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 10.12. 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 10.12 are chosen to be Ångstroms, and that the unit of charge corresponds to one electron. Determine the units of the vertical axis (see Chapter 14).*

Exercise 10.14 *Compare the asymptotic expression (10.21) for the general dipole-dipole interaction with the special cases considered earlier in the text, e.g., (3.11), (3.15), (10.72) and (10.73).*

Exercise 10.15 *Determine whether the force between two neutral groups as depicted in Figure 10.18 is attractive or repulsive.*

Exercise 10.16 *Show that the interaction energy for two quadrupoles as depicted in Figure 10.18 is*

$$\frac{1}{r-2} - \frac{4}{r-1} + \frac{6}{r} - \frac{4}{r+1} - \frac{1}{r+2} \approx \left(\frac{\partial^4}{\partial r^4}\right)\left(\frac{1}{r}\right) = \frac{24}{r^5} \tag{10.75}$$

as $r \to \infty$ by considering the error in the difference operator represented by the left-hand side, cf. Exercise 3.9.

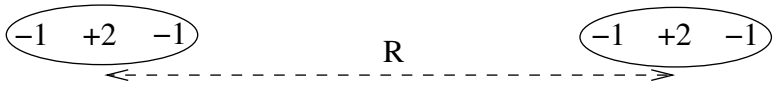


Figure 10.18: Interaction between two quadrupoles.

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 [139]. We review them here briefly. They illustrate the way that dehydrons play a role in signaling and in helping to form protein structures.

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 [139], as depicted in the PDB

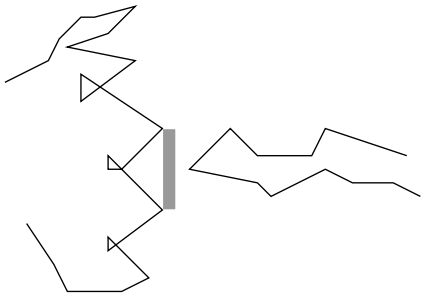


Figure 11.1: Cartoon of two fragments of proteins binding at a dehydron on one of the proteins. The dehydron is depicted as a grey strip.

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 [139]. 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 [139] 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 [139]. 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.

11.1.2 Forming structures

Dehydrons also appear more broadly in structural roles [139]. 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 [133]. 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 [448, 449, 450]. The formation of the capsid in picornaviruses [158, 234] was shown to be essentially determined by the distribution of dehydrons in the individual virus-peptide (VP) subunits [139]. 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) [158]. 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 [451, 448, 450]. 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 [139]). 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 [158] 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 [158]. 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 [158]).

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

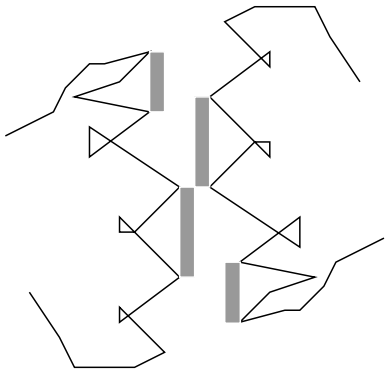


Figure 11.2: Cartoon of two fragments of proteins binding along interfaces with multiple dehydrons on both of the proteins. The dehydrons are depicted as grey strips.

target [139].

Dehydrons in the human rhinovirus (PDB code 1R1A [218]) can be seen to have similar properties [139]. 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 [218].

The Mengo encephalomyocarditis virus (PDB code 2MEV) [234] has a large number of **crystal contacts**. These are points at which each individual unit makes contact with another in the formation of a crystal used in the process of X-ray imaging. Thus, they may be thought of as a type of artifact of the protein-protein interactions related to the imaging process. The pattern of dehydrons in the virus is shown in Fig. 7 of [139]. 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 11.1 are not involved in the organization of the capsid. Of the 60 residues known to be engaged in crystal contacts for this virus [234], 54 of them have sidechain carbonaceous groups in the desolvation domains of the 15 dehydrons marked in Table 11.1. Thus, the dehydons not associated with the structural organization of the capsid can be seen to correlate with the crystal packing.

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 [428, 429], but the 'active site' in which catalysis takes place is localized. There are thousands of enzymes that play a role in metabolic pathways [367].

A84-LEU-(H)–A81-TYR-(O)
A98-GLU-(H)–A96-GLY-(O)
A96-GLY-(H)–A102-GLU-(O)
B162-LYS-(H)–A96-GLY-(O)
A101-SER-(H)–A98-GLU-(O)
A205-HIS-(H)–B208-GLN-(O)
A210-ASN-(H)–A207-ARG-(O)
A212-GLY-(H)–A210-ASN-(O)
B252-VAL-(H)–B101-ARG-(O)
B161-ARG-(H)–B159-THR-(O)
B228-ALA-(H)–B225-SER-(O)
C72-VAL-(H)–C70-THR-(O)
C141-GLN-(H)–C138-SER-(O)
C180-ILE-(H)–C178-ALA-(O)
C204-SER-(H)–C202-PRO-(O)

Table 11.1: List of dehydrons for the protein unit of Mengo encephalomyocarditis virus (PDB code 2MEV) [234] engaged in known crystal contacts (their wrapping hydrophobic groups belong to the side chains of residues known to form crystal contacts [218]). The proton donor residue is marked as (H) and the electron-donor residue supplying the carbonyl group is marked (O).

Many drugs are designed to be inhibitors of enzymes [133], 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 [133, 428]. The process of water removal can be enhanced energetically by the presence of dehydrons near the active site [133]. 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 [133, 141]. Moreover, drugs can be re-designed to enhance the interaction between hydrophobic groups on the drug and dehydrons on the target protein [91, 135, 141].

11.3 Neurophysin/vasopressin binding

In [51], 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 [51] and exhibit some additional candidate interactions to explain the particular details of the interaction zone.

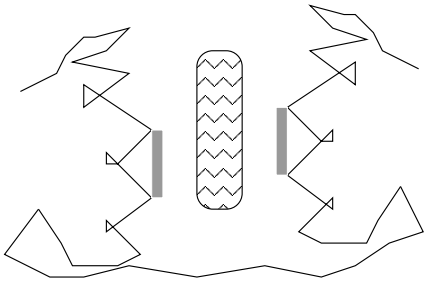


Figure 11.3: Cartoon of an enzyme with a small molecule (herringbone shading) in the active site. The dehydrons are depicted as grey strips.

Part of the reason for the suggestion [51] of a polar/aromatic interaction is the large number of disulfide bonds in neurophysin [363]. 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 [51]. 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 of neurophysin.

One corollary of the large number of disulfide bonds in neurophysin is the expectation of a large number of underwrapped hydrogen bonds [124]. 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

Re-examination of the PDB file 2BN2 shows that B-Phe98 and B-Tyr99 wrap several under-wrapped mainchain-mainchain hydrogen bonds listed in Table 11.2. 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 [51] 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 stability that this ligand displays compared to the B-Phe98,B-Phe99 variant [51], 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 [51] 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 hydropho-

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 11.2: 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.

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 11.3: 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.

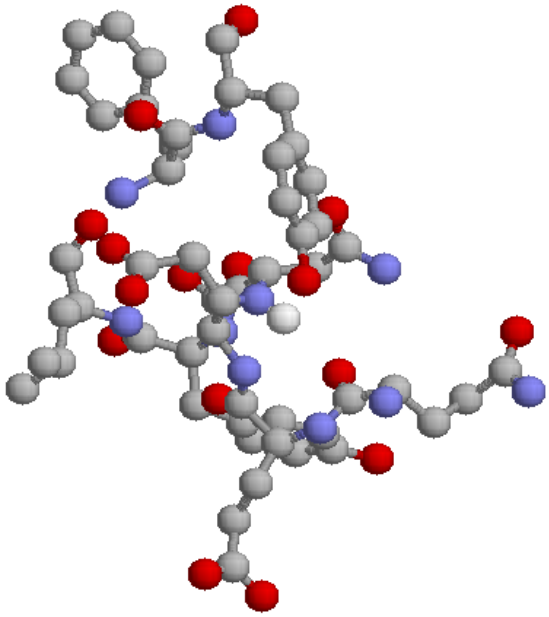


Figure 11.4: Wrapping of hydrogen bond Asn48N—Gln45O in PDB file 2BN2. The hydrogen (in white) attached to Asn48N has been added to indicate the bond, which is in the middle of the figure. The ligand pair B-Tyr99 and B-Phe98 is in the upper left of the figure. Residues 45 to 50 in the A chain are shown to indicate the bulk of the wrapping.

bic groups.

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 It also only wraps the mainchain-mainchain bond A-Asn48—A-Gln45. The mutation Leu98 binds with about half the affinity of Phe98 [51]. 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 [247], 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

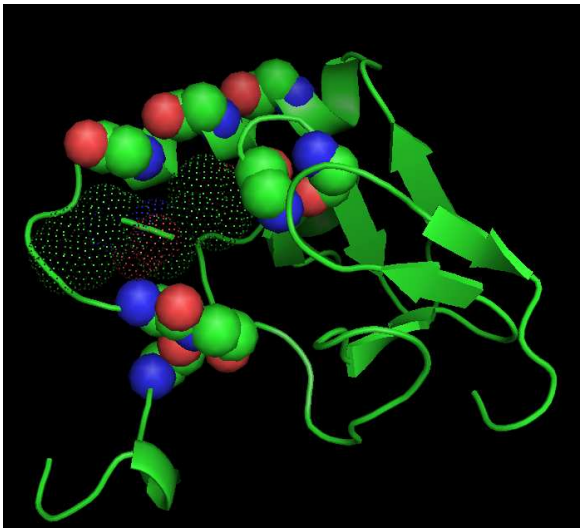


Figure 11.5: The active site of the bovine protein neurophysin bound to a fragment of the hormone vasopressin, in the PDB file 2BN2. The fragment B-Phe 98, B-Tyr 99 is shown as a cloud, and several of the hydrogen bonds in Table 11.2 are indicated by showing their backbone atoms as spheres.

hydrogen bond at the end with Cys44.

The role of residue 48 in chain A was explored in [51] by a natural 'mutation' arising due to the fact that bovine (Asn48) and ovine (Ile48) differ at this location. The ovine affinity is slightly higher, but of the same order. The 'mutation' Asn48→Ile48 is isosteric, so it is plausible that the mainchain-mainchain hydrogen bond to Gln45-O is maintained.

11.3.3 2BN2 versus 1JK4

The PDB structure 1JK4 also represents neurophysin bound to a larger fragment of vasopressin [434], 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 11.3, 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 11.3.

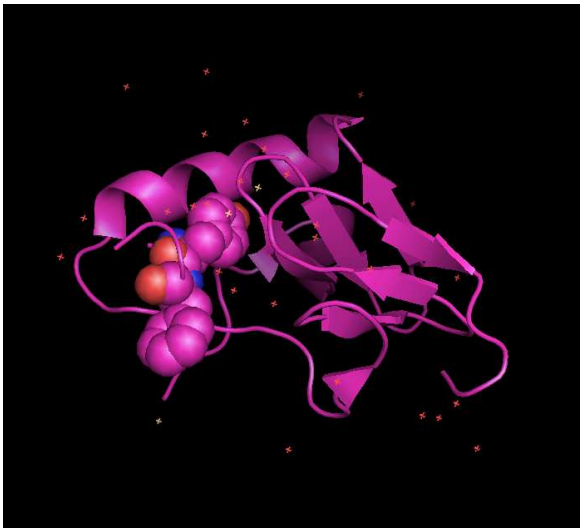


Figure 11.6: The active site of the bovine protein neurophysin bound to a fragment of the hormone vasopressin, in the PDB file 1JK4. The two residues B-Tyr 2, B-Phe 3 of the vasopressin fragment are shown as spheres.

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 inlcude the type of graph-theory analysis introduced in Section 5.4. 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 [228, 412, 436]. 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, 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 [142]. Within a conserved domain fold, the number of dehydrons in higher

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 11.4: 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.

eukaryotes is consistently greater than in species (e.g., bacteria) of lower complexity, as indicated in Figure 2.1 and Table 11.4.

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 [139], 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 [139].

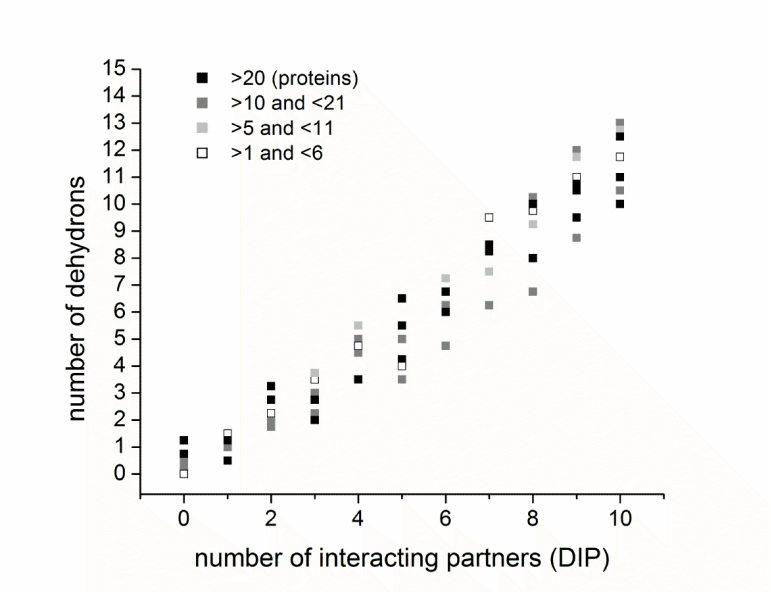


Figure 11.7: Reproduction of Fig. 1 in [139]. The vertical axis is the number of dehydrons of a given domain (fold) averaged over all proteins with that domain in the PDB. The horizontal axis is number of interactive partners for a given domain as reported in DIP. For any fixed number of partners, there 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.

11.4.1 Interactivity correlation

A study of how dehydrons provide a novel indicator of proteomic interactivity was presented in [142]. The number of dehydrons of a given protein in different organisms can vary substantially [139], as we have depicted in Figure 2.1 for myoglobin. Six other groups of proteins were analyzed in [139], and these variations are evident in Table 11.4 which reproduces the data from [139]. 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 5.4 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 [358]. The objective of DIP is "to integrate the diverse body of experimental evidence on protein-protein interactions" [358].

In [142], a subset of DIP was selected containing domains from the yeast proteome whose

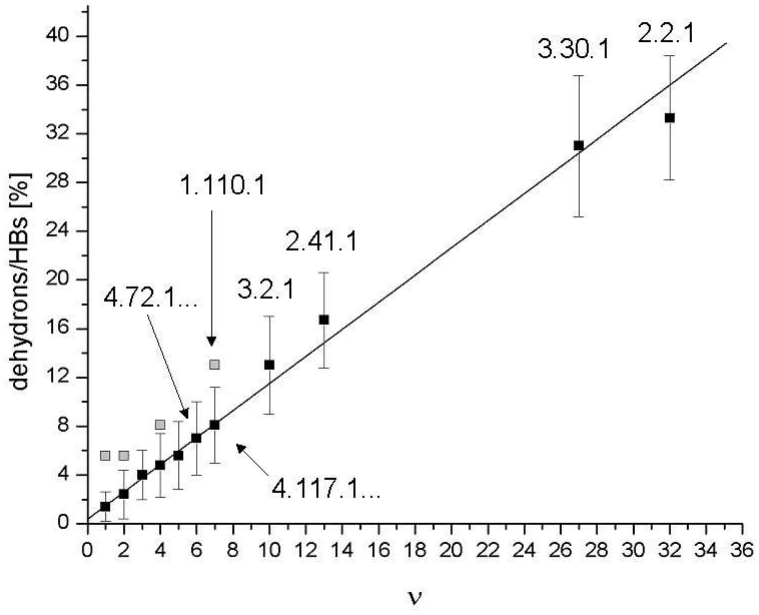


Figure 11.8: SCOP families distributed according to their average value r of the quantity $r_{d/HB}$ defined in (11.1) (vertical axis). The horizontal axis represents the interactivity ν of the families.

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 strict requirement of $\rho_G < 12$ (cf. Section 8.4) imposed in the definition of dehydron [142]. A significant correlation was demonstrated in Fig. 1 in [142] 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 [142], reproduced in Figure 11.7, 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 5.4, 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 (5.8) [436]. 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 [142], in Figure 11.8. Furthermore, the dehydron patterns associated with structural domains for a given species define scale-free interactive networks as in (5.8), shown in Figure 11.9. 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 [7, 34, 142].

A analysis similar to Section 5.4 was carried out in [142] based on the Structural Classification of Proteins (or SCOP) superfamilies [259, 302] as vertices in a graph. Interactivity among the superfamilies was defined using a variant of method described in Section 5.4 [436] utilizing protein domains with identified interactions (complexation or intramolecular interaction) in PDB files from different superfamilies [4, 103, 199, 207, 319] to define graph edges. That is, two superfamilies ϕ_1 and ϕ_2 are deemed to be interacting if they each contain folds f_1 and f_2 that are found in a single PDB file, either within a given protein (in the case of a multidomain protein) or in a protein complex represented in a single PDB file.

To account for differences in protein length, an average $r_{d/HB}$ of dehydrons per hundred hydrogen bonds was determined [142], defined as

$$r_{d/HB} = 100 n_{DH}/n_{HB}, \tag{11.1}$$

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 11.4. 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 11.8 (which reproduces Fig. 3a in [142]). 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 [148, 107] (which does not require PDB structures) when it is not. (See Section 5.4 for more information on Pfam.) The dark squares in Figure 11.8 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 11.8 indicate the SCOP superfamily classification. The value of the vertical axis in Figure 11.8 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 5.4, we consider data values (r_ϕ, f_ϕ separately for different species. More precisely, both the calculation of r_ϕ 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 11.9 (reproducing Fig. 3b in [142]), the set of values (r_ϕ, f_ϕ) are plotted on a log-log scale, and a power-law behavior

$$f(r) \approx c r^{-\gamma} \tag{11.2}$$

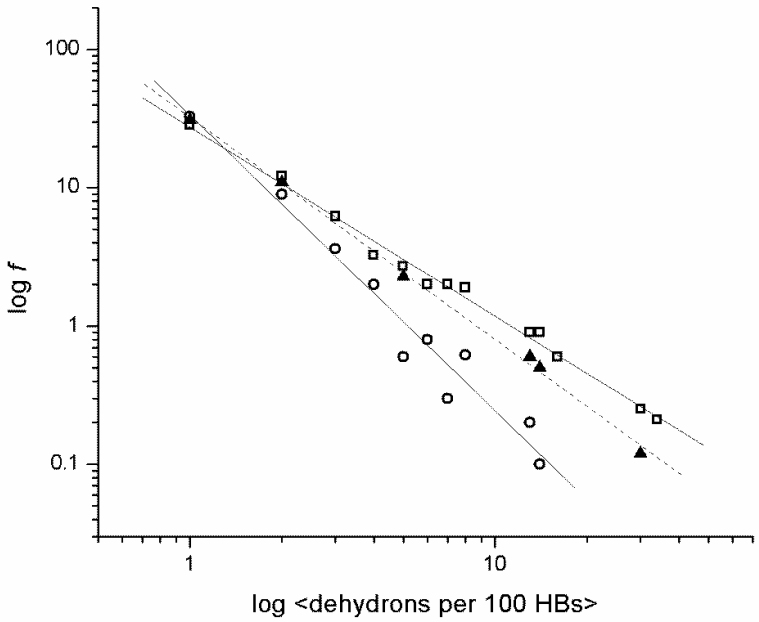


Figure 11.9: SCOP families distributed according to their average value r of the quantity $r_{d/HB}$ defined in (11.1) (horizontal axis). The vertical axis represents the percentage f of total number of families with approximatly this value of r. The data represents three different species: H. sapiens (open squares), M. musculus (solid triangles), E. coli (open circles).

is seen for each species (E. coli, mouse and human). Not surprisingly, in view of Table 11.4, 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 11.4 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 (11.2). 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 [123, 303].

11.6 Exercises

Exercise 11.1 *Determine the residues most likely to be involved in catalytic activity in the active site of an enzyme.*

Exercise 11.2 *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 16) 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). 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 [101, 265]. 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.

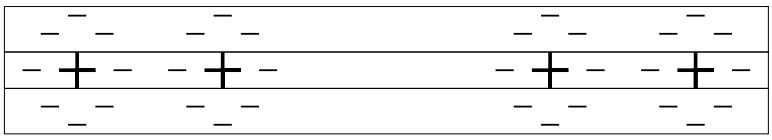


Figure 12.1: Side view of an aromatic ring showing three slabs containing alternating negative and net positive charges, providing a directional quadrupole perpendicular to the face of the ring. The + charges correspond to eight times as much charge as each - charge.

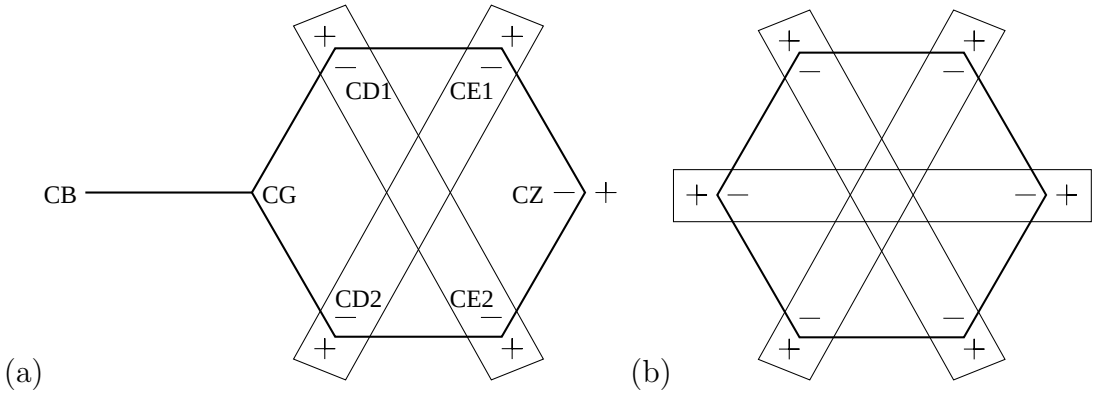


Figure 12.2: Arrangement of partial charges from Table 12.1 for (a) the phenylalanine sidechain (only carbons are shown). Corresponding partial charges are depicted for (b) a benzene ring. The boxes indicate the counterbalancing dipoles which combine to form quadrupoles.

12.1 Partial charge model

Partial charges are frequently used to model aromatics [67, 247] 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 [270].

12.2 Cation-π interactions

Among pair interactions at interfaces (Section 7.3), 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 [84, 159, 286, 329, 338, 354, 427, 439, 446]. 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 [84, 439]. The cation-π interaction also has a significant role in α-helix stabilization [373].

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 [247].

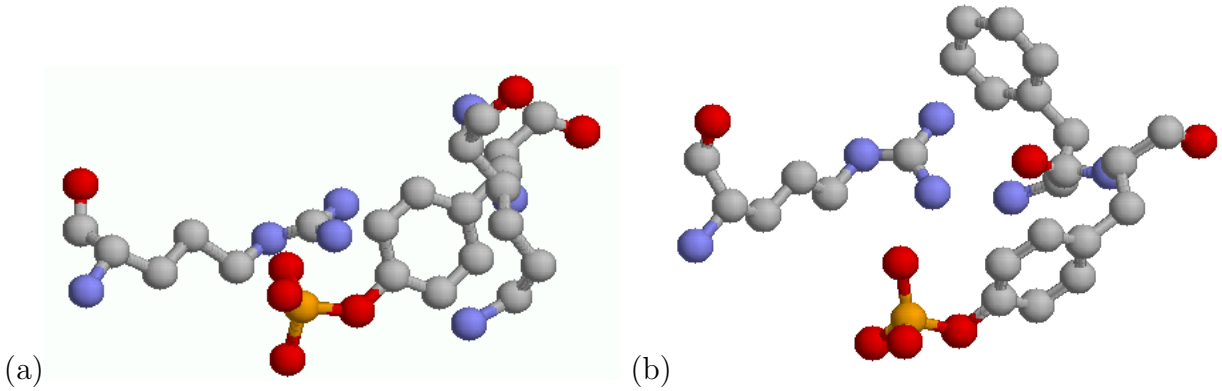
Cation-π interactions can take place with other resides, such as a phosphorylated tyrosine, cf. Figure 12.3. The aromatic ring still provides a distributed negative charge, unaffected by the addition of the phosphate group. And the two-sided nature of the polarity of an aromatic ring means that it can 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 [255] 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, 247]. 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. However, there are also other types of interactions that appear to be possible.



(a) (b)

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).

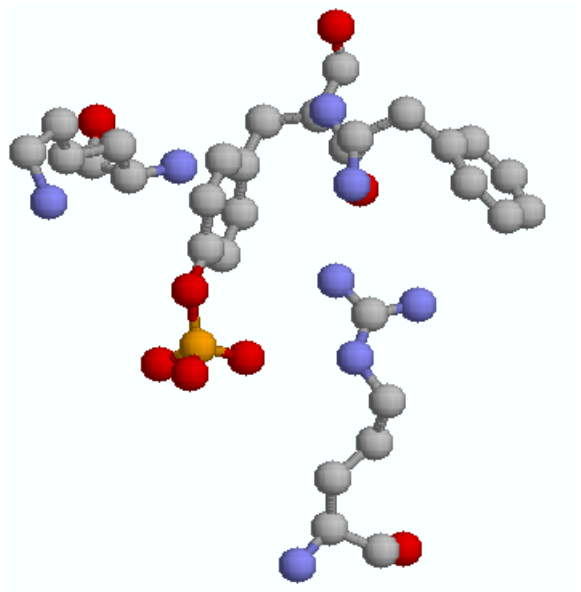
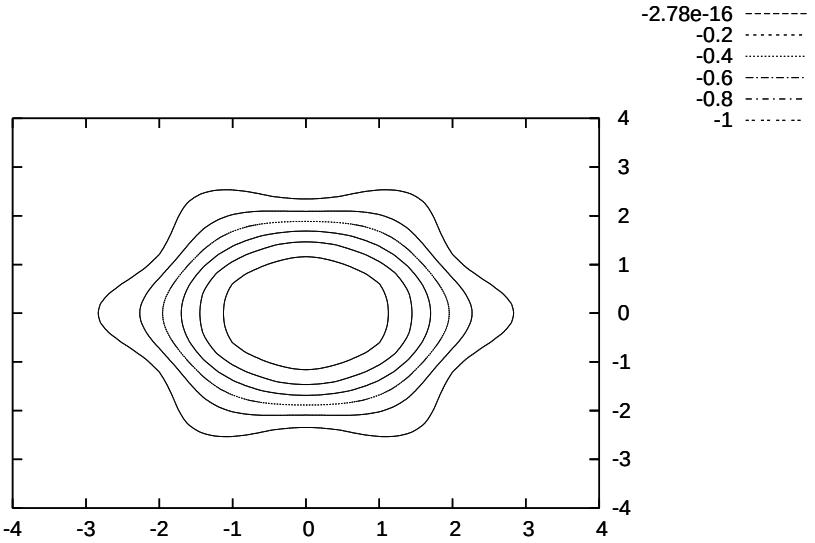


Figure 12.4: Cation-π group found in the SH2 domain in the PDB file 1TZE. Shown are a phosphorylated tyrosine (PTR-I4, middle left) and a phenylalanine (Phe-I3, upper right) together with Arg-E67 (lower right) and Lys-E109 (upper left).

(a)



(b)

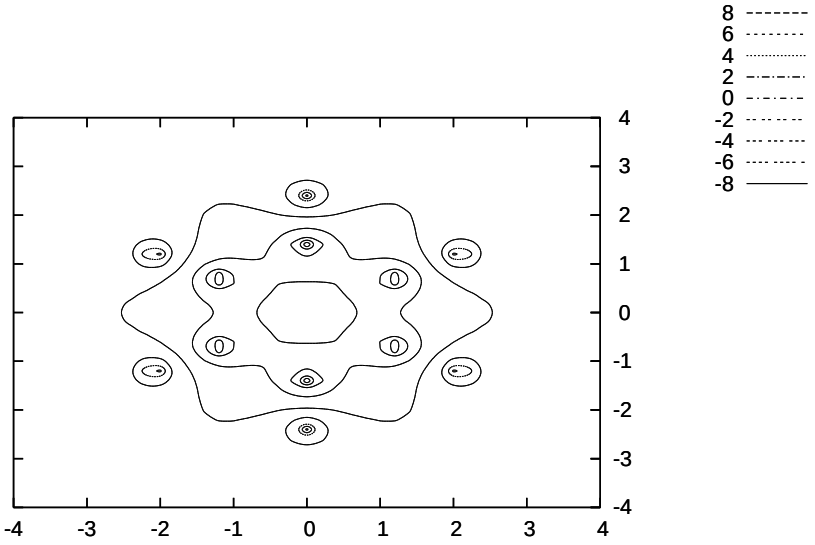


Figure 12.5: Electric potential of the partial-charge model of a benzene ring [247] in parallel planes above the ring, using the partial charges in Table 12.1: (a) 1.0 Ångstrom above and (b) 0.1 Ångstrom above.

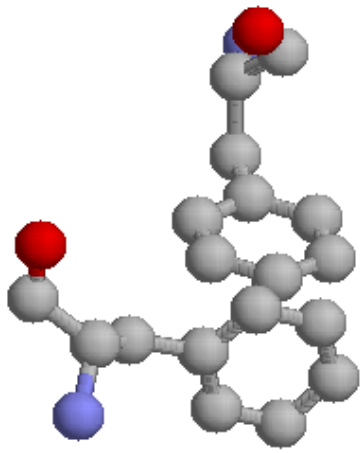


Figure 12.6: Pairwise interactions of two phenylalanines (117 and 153) in a mutant (Ser117 → Phe) of T4 lysozyme detailed in the PDB file 1TLA [5].

12.3.1 Aromatics as acceptors of hydrogen bonds

In [202], 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 [65, 247, 330, 385, 386, 393]. 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 [247]).

12.4 Aromatic-aromatic interactions

It is possible for two aromatics to interact via more complex interactions, including edge-to-face interactions [243]. In Figure 12.6 the relative orientations of two phenylalanines is shown [5] 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.

12.5 Exercises

Exercise 12.1 *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 12.2 *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 12.3 *Catalog 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.*

Chapter 13

Peptide bond rotation

We now consider an application [120] 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 4.10. 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 [323] 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. 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, 113, 200, 266]. It has been known [323] even longer that the planar state is not the preferred vacuum state. What determines the variation in ω is the local electronic environment [120], as we will review here.

The peptide bond is what is known as a **resonance** [323] 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 [323]. 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 [323].

We have already seen several examples of atomic groups whose electronic structure is a resonance 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. Similary, the

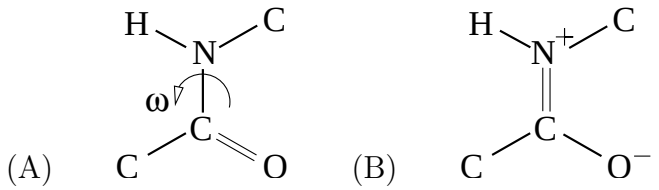


Figure 13.1: Two resonance states of the peptide bond. The double bond between the central carbon and nitrogen keeps the peptide bond planar in the right state (B). In the left state (A), the single bond can rotate, cf. Figure 5.8(b).

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) (see Chapter 18) 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.

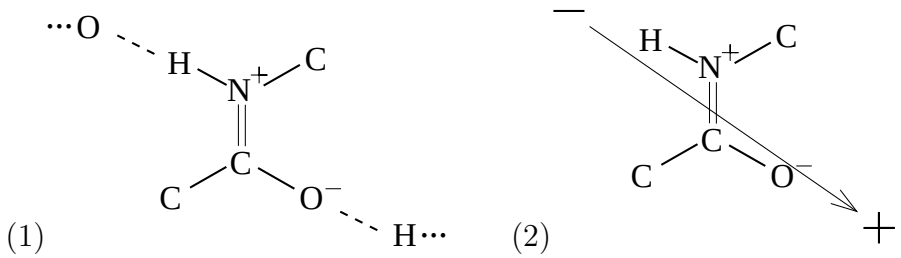


Figure 13.2: The resonance state (B) of the peptide bond shown, on the left (1), with hydrogen bonds (dashed lines) to an acceptor (indicated by the oxygen preceded by dots) and a donor (indicated by the nitrogen followed by dots). These hydrogen bonds induce a polar field that reduces the preference 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 [120] based on the observation that the value $C_B = 0.4$ [323] 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 19) 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 [120].

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

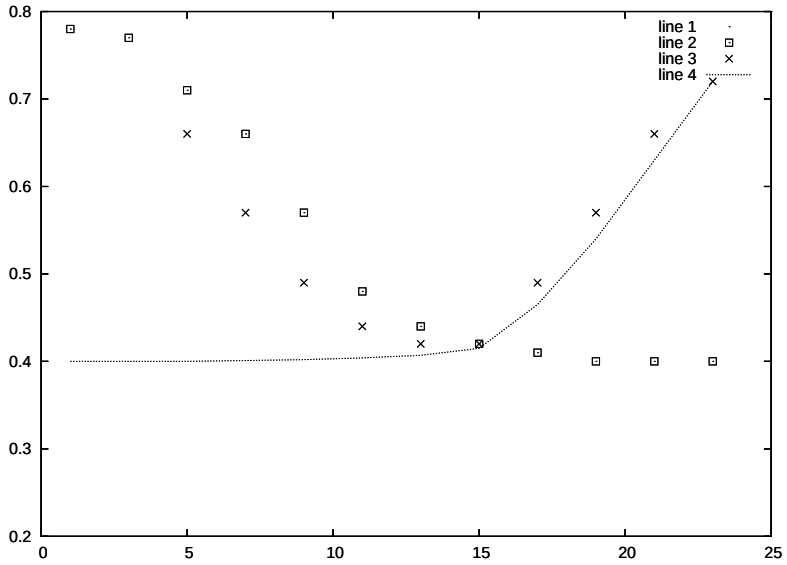


Figure 13.3: Portion of the double-bond (planar) state in the resonance for residues in two different classes, adapted from [120]. The horizontal axis is the number of nonpolar carbonaceous groups inside a sphere of radius 6 Ångstroms centered at the midpoint of the peptide bond, cf. Section 8.4. The 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 solid line represents a hypothetical response due only to the dehydration of the mainchain bonds.

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 [120], 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, that is, ones not involved in either α-helices or β-sheets. Group II consisted of peptides forming at least one backbone hydrogen bond. In each major group, subsets were defined based on the level ρ of wrapping in the vicinity of backbone.

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 $C_B = 0.4$ [323] 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). 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.

An alternate way of viewing the data is as follows. We 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.

The left side of Figure 13.3 corresponds to Group Ia and Group IIa (panels a and b, respectively), 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. On the other hand, there is a difference between the Ib and IIb groups, as indicated on the right side of Figure 13.3 (panels a and b, respectively). Group Ib prefers the vacuum state (A), whereas IIb tends to the (B) state. Figure 13.3 presents an even more refined analysis. It involves groups Iρ and IIρ for different values of ρ.

13.4 Implications for protein folding

After the "hydrophobic collapse" [447] 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 [349]. 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. (19.9). 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 19), 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

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.

14.1 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 14.1, 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 14.1: An example of basic *versus* derived units.

basic units	meter, second, kilogram, ...
derived units	joule (energy), newton (force), ...

Table 14.2: 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):

$$E = mv^2 = m(\ell/t)^2. \tag{14.1}$$

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 14.2.3.

14.1.1 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.

14.2 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).

14.2.1 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. It would be much more reasonable to pick scales more appropriate for the atomic scale [397] such as the Bohr radius $a_0 = 0.529189$Å, which is based on properties of the electron distribution for the hydrogen atom. 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 = 1.6605310^{-27} kilograms.

14.2.2 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 micometers. 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 = 1.66×10^{-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 14.3: Some units relevant for biochemistry.

14.2.3 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:

$$\text{permittivity} = \frac{\text{charge}^2}{\text{energy-length}} = \frac{\text{charge}^2\text{time}^2}{\text{mass-length}^3}. \tag{14.2}$$

Thus it is possible to have the permittivity, charge and energy be one in any units by varying the spatial (length) unit:

$$\text{length} = \frac{\text{charge}^2}{\text{energy-permittivity}}. \tag{14.3}$$

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

$$\begin{aligned}
\varepsilon_0 =& 8.8542 \times 10^{-12} \text{C}^2\text{N}^{-1}\text{m}^{-2} \\
=& 3.450 \times 10^{26} q_e^2 \text{N}^{-1}\text{m}^{-2} = 3.450 \times 10^{26} q_e^2 \text{J}^{-1}\text{m}^{-1} \\
=& 1.444 \times 10^{27} q_e^2 \text{cal}^{-1}\text{m}^{-1} = 1.444 \times 10^{30} q_e^2 \text{kcal}^{-1}\text{m}^{-1} \\
=& 2.40 \times 10^{6} q_e^2 (\text{kcal/mole})^{-1}\text{m}^{-1} = 2.40 \times q_e^2 (\text{kcal/mole})^{-1}\mu\text{m}^{-1} \\
=& 0.72 q_e^2 (\text{kcal/mole})^{-1}\text{lfs}^{-1}\,,
\end{aligned} \tag{14.4}$$

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 [89] 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 [184]. If this is the chosen spatial unit, then $\varepsilon_0 = 1$ in these units. For reference, very large viruses [451] 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 14.4: 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 14.5: Quantum chemistry units.

14.2.4 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

$$\begin{aligned} k_B =& 1.380 \times 10^{-23} \text{J/K} \\ =& \frac{1.380 \times 6.022}{4.1868} \text{cal/mole-K} \\ =& 1.984 \, \text{cal/mole-K} \end{aligned} \tag{14.5}$$

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,

$$h = 6.626068 \times 10^{-34} \, \text{m}^2\text{-kg/s} = 39.90165 \, \text{Ångstroms}^2\text{-dalton/picosecond}, \tag{14.6}$$

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.

14.3 Quantum chemistry units

The Schrödinger equation has three terms which must have the same units in order to be dimensionally correct. If we divide (18.1) 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 (18.1) 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 [397] would be to use the Bohr radius $a_0 = 0.529189$Å. This scale only differes 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

$$h = 39.90165\,\text{Ångstroms}^2\text{-dalton/picosecond} = 137.45\,a_0^2\text{-}m_e\text{/femtosecond}. \tag{14.7}$$

If we also adopt the hartree[1] [376] 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

$$E_h = m_e c^2 \alpha^2 = 4.356 \times 10^{-18}\text{joules} = 1.040 \times 10^{-21}\text{kcal} = 626.5\,\text{kcal/mole}, \tag{14.8}$$

(cf. Table 3.1) where $\alpha \approx .007$ is a dimensionless number known as the **fine structure constant**, cf. Exercise 14.7. Moreover, we also have the coefficient of the potential in (18.1) equal to E_h; that is, $e^2/(4\pi\varepsilon_0 a_0) = E_h$.

The time-derivative term in (18.1) 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 (18.1) 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.

14.4 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 term 'micromolar' is used to express the fractional concentration of one substance in another, e.g., water.

[1] Douglas Hartree (1897-1958) pioneered approximation methods for quantum chemistry calculations.

basic units	speed of light, Planck's constant, Boltzmann's constant
derived units	time unit=0.7 yoctoseconds, length unit=0.2 femtometers

Table 14.6: Some units to simplify mathematical equations.

14.5 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 14.4). 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. If we want $\hbar = 1$ and $c = 1$ [320], 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 14.5 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-$(\mathrm{L/t})^2$. Similarly, in these units $k_B = 9.2468 \times 10^{-14} K^{-1}$.

14.6 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 [2], 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 [2].

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.

14.7 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.

14.7.1 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.

14.7.2 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 [171, 80].

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 $\alpha \approx 1.2 \text{Å}^3$.

14.7.3 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, 6.022×10^{23} water molecules, weighs 18.0153 grams. At 4 degress Centigrade, where water has its maximum density, one gram of water occupies one cubic centimeter, or 10^{24}Å^3. Thus a mole of water occupies $180.153 \times 10^{23} \text{Å}^3$ (at 4° C), so a single molecule of water occupies about 29.92Å^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).

14.7.4 Fluid viscosity and diffusion

Fluids display an aggregate behavior known as **viscosity**. Fluid dynamicists [38] call the viscosity μ and physicists [239] 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.

14.7.5 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,

[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.

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.

14.7.6 Diffusion

14.8 Exercises

Exercise 14.1 *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 14.2 *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 14.3 *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 14.4 *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 14.5 *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 14.6 *Suppose temperature is in degrees Kelvin and mass is in daltons. Determine a velocity scale such that $k_B = 1$.*

Exercise 14.7 *The fine structure constant is*

$$\alpha = \frac{q_e^2}{2\,h\,\varepsilon_0\,c}, \tag{14.9}$$

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 14.8 *The Rydberg constant R_∞ is*

$$R_\infty = \frac{m_e\, q_e^4}{8\, \varepsilon_0^2\, h^3\, c}, \tag{14.10}$$

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 14.9 *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 15

More electrostatic details

15.1 Multipole expansions

We will be interested in the potential due to a charge distribution

$$V(\mathbf{x}) = \int_{\mathbb{R}^3} \frac{\rho(\mathbf{y})\, d\mathbf{y}}{|\mathbf{x}-\mathbf{y}|} \tag{15.1}$$

where ρ is the charge distribution. In some models, we want to consider point charges, so that we should really write $d\rho$ instead of $\rho(\mathbf{y})\, d\mathbf{y}$ to allow for the possibility that there are singularities in the charge distribution. More precisely, we will only consider cases where ρ has a only finite number of point singularities (a weighted finite sum of Dirac δ-functions). With this understanding, we continue to use the notation in (15.1). More complex distributions could also be considered, e.g., models of electrons distributed on a surface.

The integral in (15.1) is well defined for all $\mathbf{x}$ provided ρ is bounded and integrable, and it is also finite for $\mathbf{x}$ not in the singular set of ρ in the case that it has point singularities. We are mainly interested in the asymptotic behavior for large $|\mathbf{x}|$.

By Taylor's theorem, for any function u we can write

$$u(\mathbf{y}) = \sum_{\alpha} \frac{\mathbf{y}^\alpha}{\alpha!} D^\alpha u(0), \tag{15.2}$$

assuming that the series converges. Let us apply this to $u(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}|^{-1}$, where we think of $\mathbf{x}$ as a fixed parameter. Defining $w(\mathbf{x}) = |\mathbf{x}|^{-1}$, we find

$$D^\alpha_{\mathbf{y}} u(\mathbf{x},\mathbf{y}) = (-1)^{|\alpha|} D^\alpha_{\mathbf{x}} u(\mathbf{x},\mathbf{y}), \tag{15.3}$$

so that

$$D^\alpha_{\mathbf{y}} u(\mathbf{x},0) = (-1)^{|\alpha|} D^\alpha_{\mathbf{x}} w(\mathbf{x}), \tag{15.4}$$

since $u(\mathbf{x},0) = w(\mathbf{x})$ for all $\mathbf{x}$. Applying this in (15.2) yields

$$\frac{1}{|\mathbf{x}-\mathbf{y}|} = \sum_{\alpha} \frac{\mathbf{y}^\alpha}{\alpha!} (-1)^{|\alpha|} D^\alpha_{\mathbf{x}} w(\mathbf{x}) = \sum_{\alpha} \frac{\mathbf{y}^\alpha}{\alpha!} R_\alpha(\mathbf{x}), \tag{15.5}$$

where

$$R_\alpha(\mathbf{x}) := (-1)^{|\alpha|} D^\alpha \left(|\mathbf{x}|^{-1}\right). \tag{15.6}$$

Differentiating under the integral sign in (15.1), we find

$$V(\mathbf{x}) = \sum_\alpha \int_{\mathbb{R}^3} \frac{\mathbf{y}^\alpha}{\alpha!} \rho(\mathbf{y})\, d\mathbf{y}\ (-1)^{|\alpha|} D^\alpha \left(|\mathbf{x}|^{-1}\right) = \sum_\alpha \rho_\alpha R_\alpha(\mathbf{x}), \tag{15.7}$$

where

$$\rho_\alpha := \int_{\mathbb{R}^3} \frac{\mathbf{y}^\alpha}{\alpha!} \rho(\mathbf{y})\, d\mathbf{y}. \tag{15.8}$$

The coefficient ρ_0 is the **monopole** coefficient. The coefficients ρ_α for $|\alpha| = 1$ form the **dipole vector**. The coefficients ρ_α for $|\alpha| = 2$ are called the **quadrapole tensor** (or matrix). The coefficients ρ_α for $|\alpha| = 3$ are called the **octapole tensor**.

We can write

$$R_\alpha(\mathbf{x}) = (-1)^{|\alpha|} D^\alpha \left(|\mathbf{x}|^{-1}\right) = \frac{p_\alpha(\mathbf{x})}{|\mathbf{x}|^{2|\alpha|+1}}, \tag{15.9}$$

where p_α is a homogeneous polynomial of degree $|\alpha|$ (proof by induction). By direct calculation we see that $p_0(\mathbf{x}) \equiv 1$ and

$$p_\alpha(\mathbf{x}) = \mathbf{x}^\alpha \quad \text{for } |\alpha| = 1. \tag{15.10}$$

By induction, we can then see that

$$p_{\alpha+e^i}(\mathbf{x}) = (2|\alpha| + 1) x_i p_\alpha(\mathbf{x}) - |\mathbf{x}|^2 p_{\alpha,i}(\mathbf{x}), \tag{15.11}$$

where $p_{\alpha,i} = D^{e^i} p_\alpha = \frac{\partial}{\partial x_i} p_\alpha$. Thus $R_\alpha(\mathbf{x})$ is homogeneous of degree $-1 - |\alpha|$ and

$$|R_\alpha(\mathbf{x})| = \left|D^\alpha \left(|\mathbf{x}|^{-1}\right)\right| = \mathcal{O}(|\mathbf{x}|^{-|\alpha|-1}) \tag{15.12}$$

for large $|\mathbf{x}|$.

Note that (15.10) and (15.11) are compatible (recall that $p_0(\mathbf{x}) \equiv 1$). To evaluate p_α for $|\alpha| = 2$, write

$$p_{e^j+e^i}(\mathbf{x}) = 3x_i p_{e^j}(\mathbf{x}) - |\mathbf{x}|^2 p_{e^j,i}(\mathbf{x}) = 3x_i x_j - |\mathbf{x}|^2 \delta_{ji}, \tag{15.13}$$

where δ_{ji} is the Kronecker delta. Therefore

$$\left(p_{e^j+e^i}(\mathbf{x})\right) = \begin{pmatrix} 2x_1^2 - x_2^2 - x_3^2 & 3x_1x_2 & 3x_1x_3 \\ 3x_1x_2 & -x_1^2 + 2x_2^2 - x_3^2 & 3x_2x_3 \\ 3x_1x_3 & 3x_2x_3 & -x_1^2 - x_2^2 + 2x_3^2 \end{pmatrix}. \tag{15.14}$$

Certain combinations of the terms $R_\alpha(\mathbf{x}) = (-1)^{|\alpha|} D^\alpha \left(|\mathbf{x}|^{-1}\right)$ behave differently. Let $\mathbf{e}^i$ denote the i-th coordinate vector. Then

$$\sum_{i=1}^3 R_{2\mathbf{e}^i}(\mathbf{x}) = \sum_{i=1}^3 D^{2\mathbf{e}^i} \left(|\mathbf{x}|^{-1}\right) = \Delta \left(|\mathbf{x}|^{-1}\right) = 0, \tag{15.15}$$

for $\mathbf{x} \neq 0$, where Δ is the Laplacean. Written out in coordinates, this reads

$$R_{(2,0,0)} + R_{(0,2,0)} + R_{(0,0,2)} = 0 \tag{15.16}$$

Higher-order relations hold as well. For example

$$\sum_{i=1}^{3} R_{2\mathbf{e}^i+e^j}(\mathbf{x}) = \sum_{i=1}^{3} D^{2\mathbf{e}^i+e^j}\left(|\mathbf{x}|^{-1}\right) = \frac{\partial}{\partial x_j}\Delta\left(|\mathbf{x}|^{-1}\right) = 0, \tag{15.17}$$

Written out in coordinates, this reads (for $j = 1$)

$$R_{(3,0,0)} + R_{(1,2,0)} + R_{(1,0,2)} = 0 \tag{15.18}$$

15.1.1 Hydrogen potential

As an example, let us take ρ corresponding to the hydrogen atom. Thus

$$\rho = \delta_0 - \tilde{\rho}, \tag{15.19}$$

where $\tilde{\rho}$ denotes the electron density given by

$$\tilde{\rho}(\mathbf{x}) = \frac{1}{8\pi} e^{-|\mathbf{x}|}. \tag{15.20}$$

Since

$$\int_{\mathbb{R}^3} \tilde{\rho}(\mathbf{x}) = \frac{1}{8\pi}\int_{\mathbb{R}^3} e^{-|\mathbf{x}|}\, d\mathbf{x} = \frac{1}{2}\int_0^\infty r^2 e^{-r}\, dr = 1, \tag{15.21}$$

the monopole coefficient $\rho_0 = 0$. Similarly, due to the symmetry of ρ, $\rho_\alpha = 0$ for $|\alpha| = 1$. However, for $\alpha = 2\mathbf{e}^i$, where $\mathbf{e}^i$ denotes the i-th coordinate vector, we find

$$\rho_{2\mathbf{e}^i} = \int_{\mathbb{R}^3} \frac{x_i^2}{2}\tilde{\rho}(\mathbf{x})\, d\mathbf{x} = \frac{1}{16\pi}\int_{\mathbb{R}^3} x_i^2 e^{-|\mathbf{x}|}\, d\mathbf{x} = c > 0, \tag{15.22}$$

where c is the same for all i, by symmetry. On the other hand, for $\alpha = \mathbf{e}^i + \mathbf{e}^j$ for $i \neq j$, we have $\rho_\alpha = 0$, again by symmetry. Thus the quadrapole matrix for the hydrogen atom is a constant times the identity matrix. In view of (15.12), this means that there is no term asymptotic to $|\mathbf{x}|^{-3}$ in the hydrogen potential.

To verify this result, we can derive it in a different way. We can work with cylindrical coordinates $(x, y, z) = (r\cos\theta, r\sin\theta, z)$. By symmetry, it is sufficient to compute V only on the z-axis. In these coordinates

$$V(0,0,R) = \frac{1}{R} - \frac{1}{8\pi}\int_{\mathbb{R}^3} \frac{e^{-\sqrt{r^2+z^2}}\, r\, dr\, d\theta\, dz}{\sqrt{r^2+(z-R)^2}} = \frac{1}{R} - \frac{1}{4}\int_0^\infty\int_\infty^\infty \frac{e^{-\sqrt{r^2+z^2}}\, r\, dr\, dz}{\sqrt{r^2+(z-R)^2}}. \tag{15.23}$$

Writing $\sqrt{r^2+(z-R)^2} = R\sqrt{(\epsilon r)^2+(\epsilon z-1)^2}$, with $\epsilon = 1/R$, and expanding using the expression

$$(1-\delta)^{-\frac{1}{2}} = 1 + \tfrac{1}{2}\delta + \tfrac{3}{8}\delta^2 + \mathcal{O}\left(\delta^3\right), \tag{15.24}$$

we find that

$$\begin{aligned}\int_0^\infty \int_\infty^\infty \frac{e^{-\sqrt{r^2+z^2}}\, r\, dr\, dz}{\sqrt{r^2+(z-R)^2}} &= \frac{1}{R}\int_0^\infty \int_\infty^\infty \frac{e^{-\sqrt{r^2+z^2}}\, r\, dr\, dz}{\sqrt{\epsilon^2(r^2+z^2)-2\epsilon z+1}} \\ &= \frac{1}{R}\int_0^\infty \int_\infty^\infty e^{-\sqrt{r^2+z^2}} \left(1+\epsilon z-\tfrac{1}{2}\epsilon^2(r^2+z^2)+\tfrac{3}{8}(2\epsilon z-\epsilon^2(r^2+z^2))^2\right)\, r\, dr\, dz + \mathcal{O}\left(\epsilon^4\right) \\ &= \frac{1}{R}\int_0^\infty \int_\infty^\infty e^{-\sqrt{r^2+z^2}} \left(1+\epsilon z-\tfrac{1}{2}\epsilon^2 r^2+\epsilon^2 z^2\right)\, r\, dr\, dz + \mathcal{O}\left(\epsilon^4\right) \\ &= \frac{4}{R}+\frac{1}{R^3}\int_0^\infty \int_\infty^\infty e^{-\sqrt{r^2+z^2}} \left(-\tfrac{1}{2}r^2+z^2\right)\, r\, dr\, dz + \mathcal{O}\left(R^{-4}\right),\end{aligned} \tag{15.25}$$

since the term involving the factor z in the integrand vanishes by symmetry. The rather daunting looking remaining integral can be reconfigured back into Cartesian coordinates as

$$\int_0^\infty \int_\infty^\infty e^{-\sqrt{r^2+z^2}} \left(-\tfrac{1}{2}r^2+z^2\right)\, r\, dr\, dz = \frac{1}{2\pi}\int_{\mathbb{R}^3} e^{-|\mathbf{x}|} \left(-\tfrac{1}{2}|\mathbf{x}|^2+\tfrac{3}{2}z^2\right)\, d\mathbf{x}. \tag{15.26}$$

By symmetry,

$$\int_{\mathbb{R}^3} e^{-|\mathbf{x}|}|\mathbf{x}|^2\, d\mathbf{x} = \int_{\mathbb{R}^3} e^{-|\mathbf{x}|}\left(x^2+y^2+z^2\right)\, d\mathbf{x} = 3\int_{\mathbb{R}^3} e^{-|\mathbf{x}|}z^2\, d\mathbf{x}, \tag{15.27}$$

so we get a confirmation that the $\mathcal{O}\left(R^{-3}\right)$ term vanishes.

15.1.2 Charge interactions

The interaction energy between two charge distributions ρ and $\tilde{\rho}$ is given by

$$\int_{\mathbb{R}^3} V(\tilde{\mathbf{x}})\tilde{\rho}(\tilde{\mathbf{x}})\, d\tilde{\mathbf{x}} = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\tilde{\rho}(\mathbf{x})\rho(\mathbf{x})\, d\mathbf{x}\, d\tilde{\mathbf{x}}}{|\tilde{\mathbf{x}}-\mathbf{x}|}. \tag{15.28}$$

It is often useful to think of the charge distributions ρ and $\tilde{\rho}$ as separated by a distance $\mathbf{r}$, so that the location of the origin of the coordinates for both can be taken as the charge centers. With this modification, (15.28) becomes

$$E(\mathbf{r}) = \int_{\mathbb{R}^3} V(\tilde{\mathbf{x}}-\mathbf{r})\tilde{\rho}(\tilde{\mathbf{x}})\, d\tilde{\mathbf{x}} = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\tilde{\rho}(\mathbf{x})\rho(\mathbf{x})\, d\mathbf{x}\, d\tilde{\mathbf{x}}}{|\tilde{\mathbf{x}}-\mathbf{x}-\mathbf{r}|}. \tag{15.29}$$

Applying (15.5) yields

$$\frac{1}{|\mathbf{r}-\tilde{\mathbf{x}}+\mathbf{x}|} = \sum_\alpha \frac{(\tilde{\mathbf{x}}-\mathbf{x})^\alpha}{\alpha!} R_\alpha(\mathbf{r}), \tag{15.30}$$

Therefore

$$E(\mathbf{r}) = \sum_\alpha \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{(\tilde{\mathbf{x}}-\mathbf{x})^\alpha \tilde{\rho}(\tilde{\mathbf{x}})\rho(\mathbf{x})\, d\mathbf{x}\, d\tilde{\mathbf{x}}}{\alpha!} R_\alpha(\mathbf{r}). \tag{15.31}$$

The expressions

$$C(\rho,\tilde{\rho},\alpha)=\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{(\tilde{\mathbf{x}}-\mathbf{x})^\alpha\tilde{\rho}(\tilde{\mathbf{x}})\rho(\mathbf{x})\,d\mathbf{x}\,d\tilde{\mathbf{x}}}{\alpha!} \tag{15.32}$$

can be simplified, as follows. First of all,

$$C(\rho,\tilde{\rho},0)=\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\tilde{\rho}(\tilde{\mathbf{x}})\rho(\mathbf{x})\,d\mathbf{x}\,d\tilde{\mathbf{x}}=\int_{\mathbb{R}^3}\tilde{\rho}(\tilde{\mathbf{x}})\,d\tilde{\mathbf{x}}\int_{\mathbb{R}^3}\rho(\mathbf{x})\,d\mathbf{x}=\tilde{\rho}_0\rho_0. \tag{15.33}$$

This constitutes the charge-charge interaction. Similarly, for $|\alpha|=1$, $(\tilde{\mathbf{x}}-\mathbf{x})^\alpha=\tilde{\mathbf{x}}^\alpha-\mathbf{x}^\alpha$, so

$$\begin{aligned}C(\rho,\tilde{\rho},\alpha)=&\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\tilde{\mathbf{x}}^\alpha\tilde{\rho}(\tilde{\mathbf{x}})\rho(\mathbf{x})\,d\mathbf{x}\,d\tilde{\mathbf{x}}-\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\mathbf{x}^\alpha\tilde{\rho}(\tilde{\mathbf{x}})\rho(\mathbf{x})\,d\mathbf{x}\,d\tilde{\mathbf{x}}\\ =&\int_{\mathbb{R}^3}\tilde{\mathbf{x}}^\alpha\tilde{\rho}(\tilde{\mathbf{x}})\,d\tilde{\mathbf{x}}\int_{\mathbb{R}^3}\rho(\mathbf{x})\,d\mathbf{x}-\int_{\mathbb{R}^3}\mathbf{x}^\alpha\rho(\mathbf{x})\,d\mathbf{x}\int_{\mathbb{R}^3}\tilde{\rho}(\tilde{\mathbf{x}})\,d\tilde{\mathbf{x}}\\ =&\tilde{\rho}_\alpha\rho_0-\tilde{\rho}_0\rho_\alpha.\end{aligned} \tag{15.34}$$

This constitutes the charge-dipole interaction.

If there is no net charge in either distibution, then $C(\rho,\tilde{\rho},\alpha)=0$ for $|\alpha|\le 1$.

When $|\alpha|=2$, there are two cases to consider. First suppose that $\alpha=2e^i$ for some $i=1,2,3$. Then

$$(\tilde{\mathbf{x}}-\mathbf{x})^\alpha=(\tilde{x}_i-x_i)^2=\tilde{x}_i^2-2\tilde{x}_ix_i+x_i^2. \tag{15.35}$$

Therefore

$$C(\rho,\tilde{\rho},\alpha)=\tilde{\rho}_\alpha\rho_0-2\tilde{\rho}_{e^i}\rho_{e^i}+\tilde{\rho}_0\rho_\alpha. \tag{15.36}$$

Similarly, if $\alpha=e^i+e^j$ for $i\ne j$, then

$$(\tilde{\mathbf{x}}-\mathbf{x})^\alpha=(\tilde{x}_i-x_i)(\tilde{x}_j-x_j)=\tilde{x}_i\tilde{x}_j-x_i\tilde{x}_j-\tilde{x}_ix_j+x_ix_j, \tag{15.37}$$

and

$$C(\rho,\tilde{\rho},\alpha)=\tilde{\rho}_\alpha\rho_0-\tilde{\rho}_{e^j}\rho_{e^i}-\tilde{\rho}_{e^i}\rho_{e^j}+\tilde{\rho}_0\rho_\alpha. \tag{15.38}$$

In view of (15.36), we can say that (15.38) holds in general for $\alpha=e^i+e^j$ for any $i,j=1,2,3$:

$$C(\rho,\tilde{\rho},e^i+e^j)=\tilde{\rho}_{e^i+e^j}\rho_0-\tilde{\rho}_{e^j}\rho_{e^i}-\tilde{\rho}_{e^i}\rho_{e^j}+\tilde{\rho}_0\rho_{e^i+e^j}. \tag{15.39}$$

When both distributions have net charge zero, this simplifies to

$$C(\rho,\tilde{\rho},e^i+e^j)=-\tilde{\rho}_{e^j}\rho_{e^i}-\tilde{\rho}_{e^i}\rho_{e^j}. \tag{15.40}$$

This constitutes the dipole-dipole interaction.

When $|\alpha|=3$, we write $\alpha=e^i+e^j+e^k$ for $i,j,k=1,2,3$, and

$$\begin{aligned}(\tilde{\mathbf{x}}-\mathbf{x})^\alpha=&(\tilde{x}_i-x_i)(\tilde{x}_j-x_j)(\tilde{x}_k-x_k)\\ =&\tilde{x}_i\tilde{x}_j\tilde{x}_k-\tilde{x}_i\tilde{x}_jx_k+\tilde{x}_ix_jx_k-x_ix_jx_k\\ &+x_ix_j\tilde{x}_k-x_i\tilde{x}_j\tilde{x}_k+x_i\tilde{x}_jx_k-\tilde{x}_ix_j\tilde{x}_k.\end{aligned} \tag{15.41}$$

Therefore

$$\begin{aligned} C(\rho,\tilde{\rho},e^i+e^j+e^k) =& \tilde{\rho}_{e^i+e^j+e^k}\rho_0 - \tilde{\rho}_{e^i+e^j}\rho_{e^k} + \tilde{\rho}_{e^i}\rho_{e^j+e^k} - \tilde{\rho}_0\rho_{e^i+e^j+e^k} \\ &+\tilde{\rho}_{e^k}\rho_{e^i+e^j} - \tilde{\rho}_{e^j+e^k}\rho_{e^i} + \tilde{\rho}_{e^j}\rho_{e^i+e^k} - \tilde{\rho}_{e^i+e^k}\rho_{e^j}\,. \end{aligned} \tag{15.42}$$

For distributions with net charge zero, this represents the dipole-quadrapole interaction. For distributions with non-zero net charge, this includes a charge-octapole interaction.

Suppose in addition that the dipole term for $\tilde{\rho}$ is also zero, that is, $\tilde{\rho}_\alpha = 0$ for $|\alpha| = 1$. Then (15.42) simplifies to

$$C(\rho,\tilde{\rho},e^i+e^j+e^k) = -\tilde{\rho}_{e^i+e^j}\rho_{e^k} - \tilde{\rho}_{e^j+e^k}\rho_{e^i} - \tilde{\rho}_{e^i+e^k}\rho_{e^j}\,. \tag{15.43}$$

Suppose we consider a simple dipole where $\rho_{(1,0,0)} = 1$ and the other dipole coefficients are zero. If we take the example of the hydrogen atom for $\tilde{\rho}$ interacting with this dipole we find

$$\begin{aligned} C_{111} =& -3c \\ C_{122} =& C_{212} = C_{221} = -c \\ C_{133} =& C_{313} = C_{331} = -c, \end{aligned} \tag{15.44}$$

with the rest of the (twenty) coefficients zero, where we have written C_{ijk} for $C(\rho,\tilde{\rho},e^i+e^j+e^k)$ and c is the constant in (15.22). Using the multi-index notation, this simplifes to

$$C(\rho,\tilde{\rho},(3,0,0)) = -3c, \quad C(\rho,\tilde{\rho},(1,2,0)) = C(\rho,\tilde{\rho},(1,0,2)) = -c \tag{15.45}$$

From (15.18), we conclude that

$$E(\mathbf{r}) = -2cR_{(3,0,0)}(\mathbf{r}) + \mathcal{O}\left(|\mathbf{r}|^{-5}\right) = \mathcal{O}\left(|\mathbf{r}|^{-4}\right). \tag{15.46}$$

This result is surprising since the potential corresponding to the quadrapole moment of hydrogen has does not contribute an $\mathcal{O}\left(|\mathbf{r}|^{-3}\right)$ term.

Chapter 16

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 5.8) 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.

16.1 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 16.2.

16.2 Proline-like configurations

Proline is the unique residue that turns back and forms a second covalent bond to the backbone, as shown in Figure 4.7. 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 16.1. 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

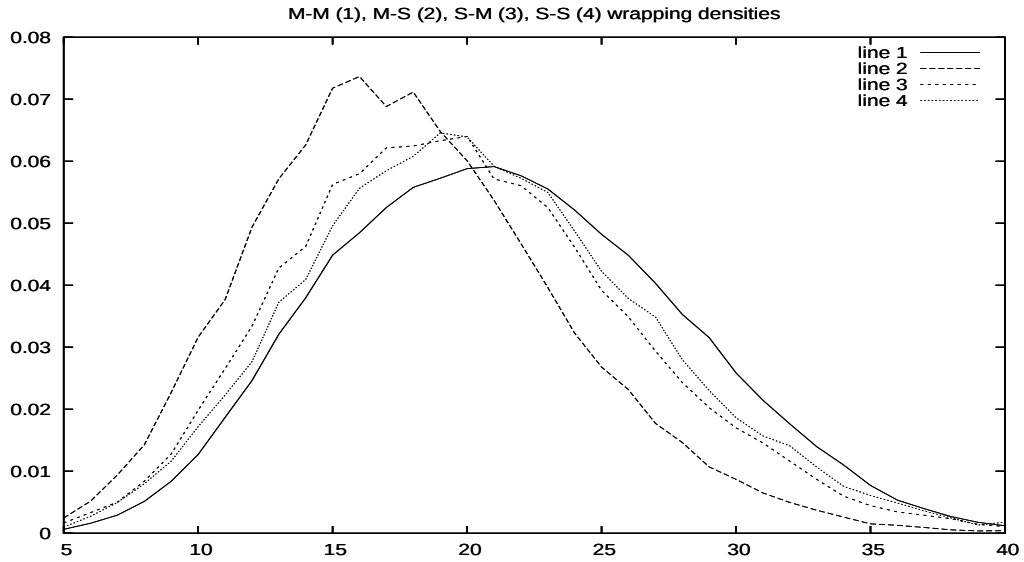


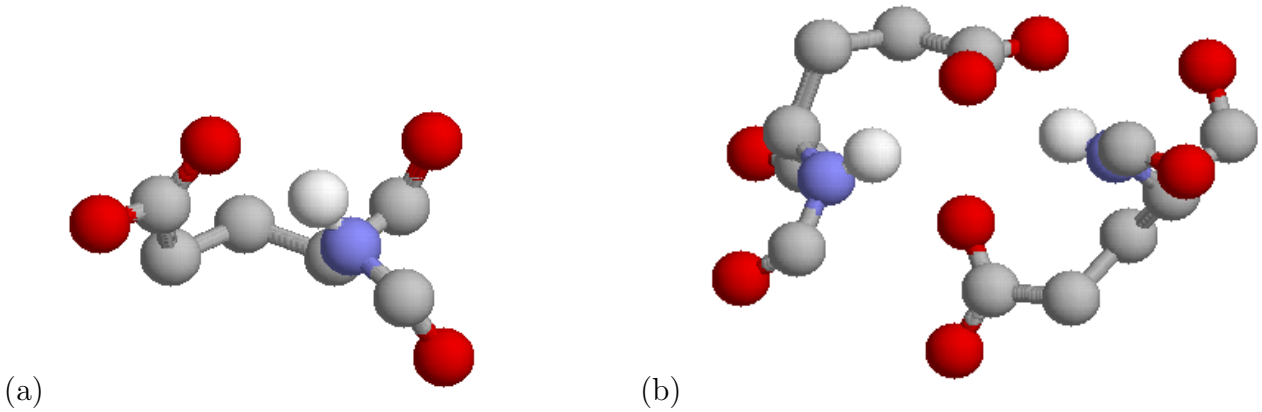
Figure 16.1: Distribution of wrapping for hydrogen bonds in a subset of the PDB Select structures. Bonds involving self-attachment, which have a significantly smaller desolvation domain, have been excluded. Solid line: M-M; large-dash line: M-S; small-dash line: S-M; dotted line: S-S. Desolvation radius 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.

16.2.1 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 16.2(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 16.2(b).



(a) (b)

Figure 16.2: 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 16.1, 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 16.1 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 16.1: 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 16.3 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 16.1 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 16.1), 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 16.1. 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

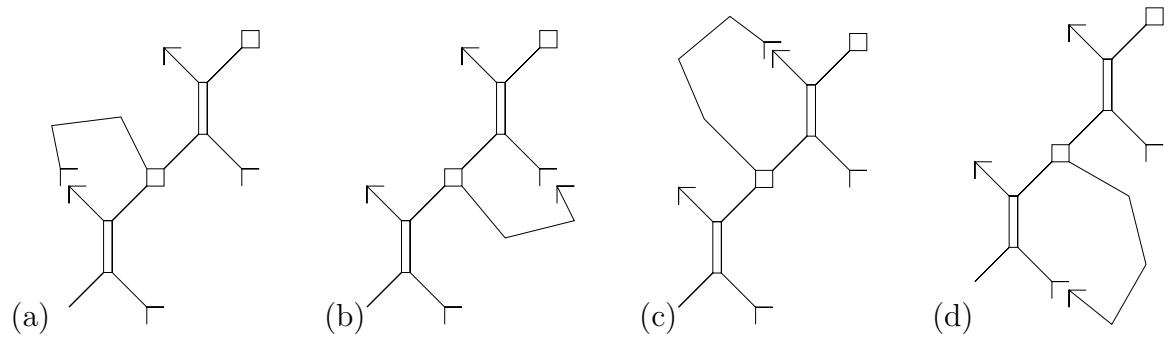


Figure 16.3: Hydrogen bond-forming sidechain configurations in the case of (a) zero-offset, mainchain-sidechain (M-S) bonds; (b) zero-offset, sidechain-mainchain (S-M) bonds; (c) +1 offset, mainchain-sidechain (M-S) bonds; and (d) -1 offset, sidechain-mainchain (S-M) bonds. See Table 16.1 for (a) and (b) and Table 16.2 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 16.1 is depicted in Figure 16.3(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 16.3(c-d). In Table 16.2, 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 16.2 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 16.2: 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 16.3 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 16.1.

The next closest positions for a residue to make hydrogen bonds with the mainchain of its sequence neighbor are depicted in Figure 16.4. In Table 16.3, 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 16.2 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 16.4.

16.2.2 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:

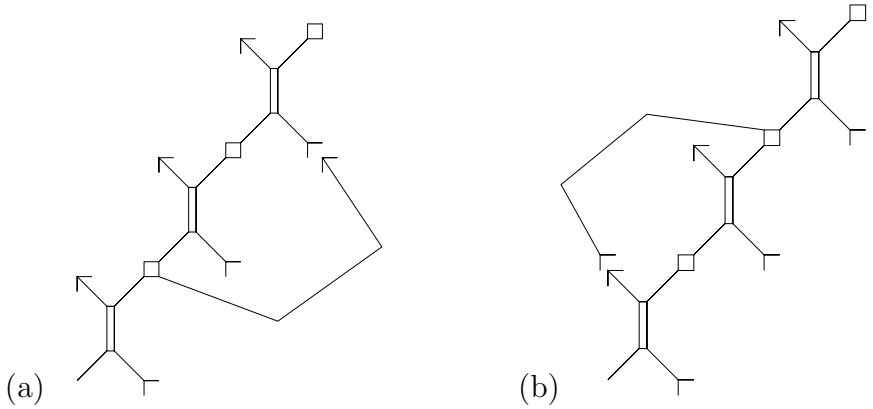


Figure 16.4: Configuration of the (a) +1 offset sidechains for the sidechain-mainchain (S-M) configuration, and the (b) -1 offset sidechains for the mainchain-sidechain (M-S) configuration. See Table 16.3.

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 16.3: 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 16.4.

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 16.4: 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 16.5: 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 16.1.

14.2 for Asp and 14.9 for Asn. In Table 16.5, 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 $k - 2$.

16.3 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 16.6, 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 16.6. 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 16.6: 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 $|j| > 10$;

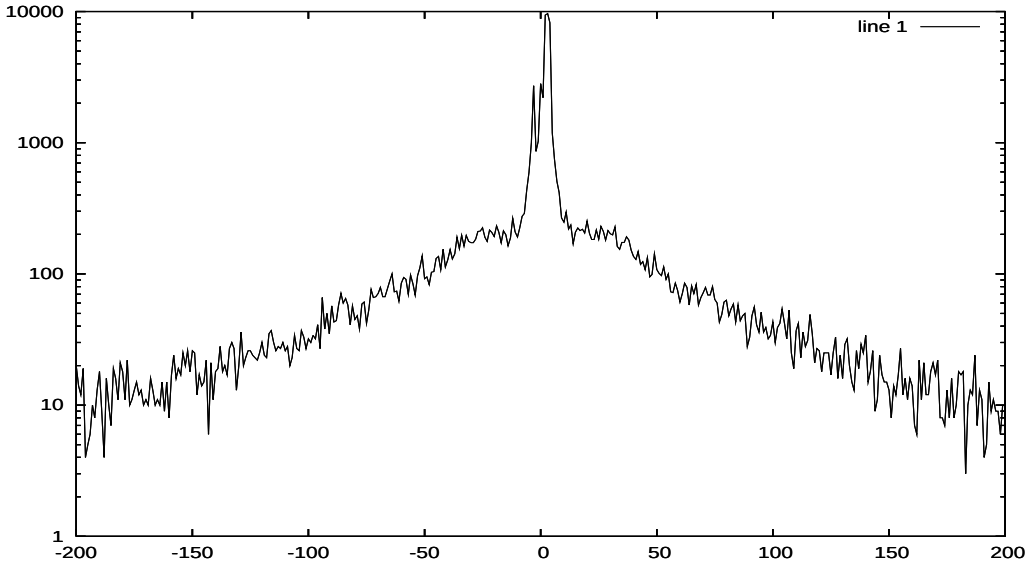


Figure 16.5: Distribution of sequence distances between donors and acceptors in hydrogen bonds between mainchains and sidechains. Both M-S and S-M bonds are included.

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 16.7), 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 16.5. We see that it is heavily concentrated on small distances, and Figure 16.6 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 16.6 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 16.7 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

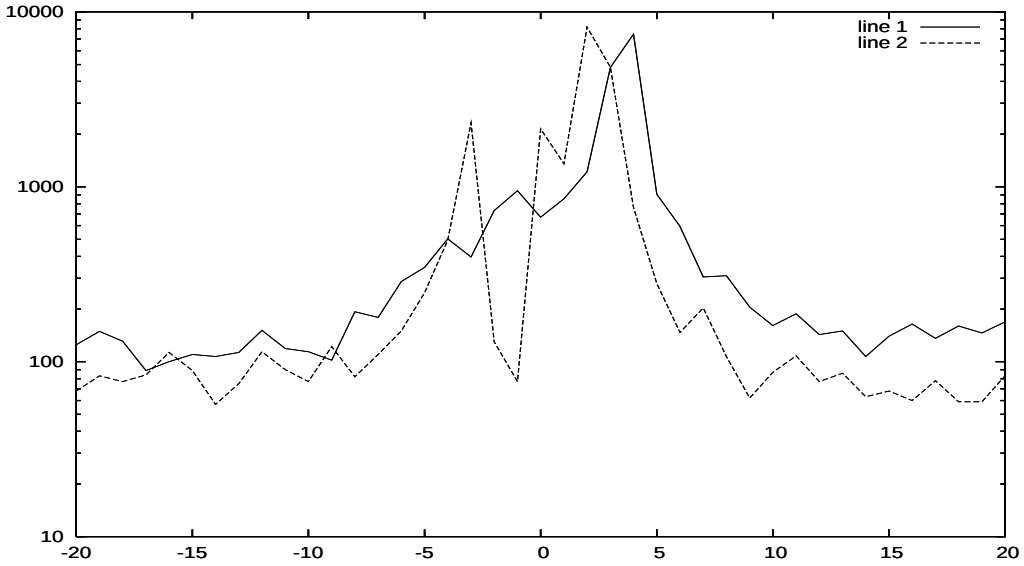


Figure 16.6: Distribution of sequence distances between donors and acceptors in hydrogen bonds between mainchains and sidechains. Solid line: S-M bonds; dashed line: M-S bonds.

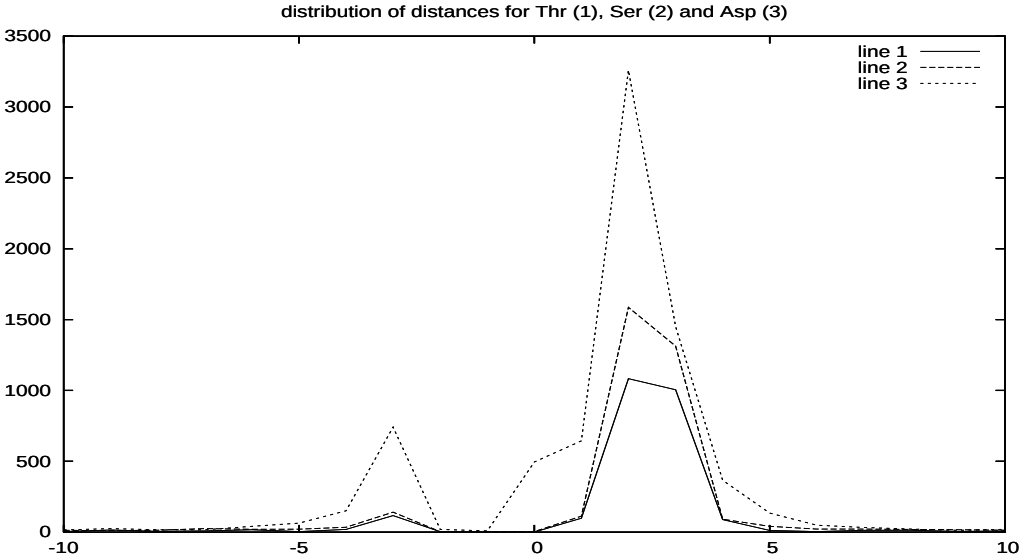


Figure 16.7: Distribution of sequence distances for M-S hydrogen bonds for Thr (solid line), Ser (dashed line) and Asp (dotted line).

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 16.1).

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 16.6 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 16.8. However, it is spread more broadly over -10 to +10, rather than being concentrated in a region ± 2 around its mode ($M = 5$).

16.4 Unusual hydrogen bonds

16.4.1 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

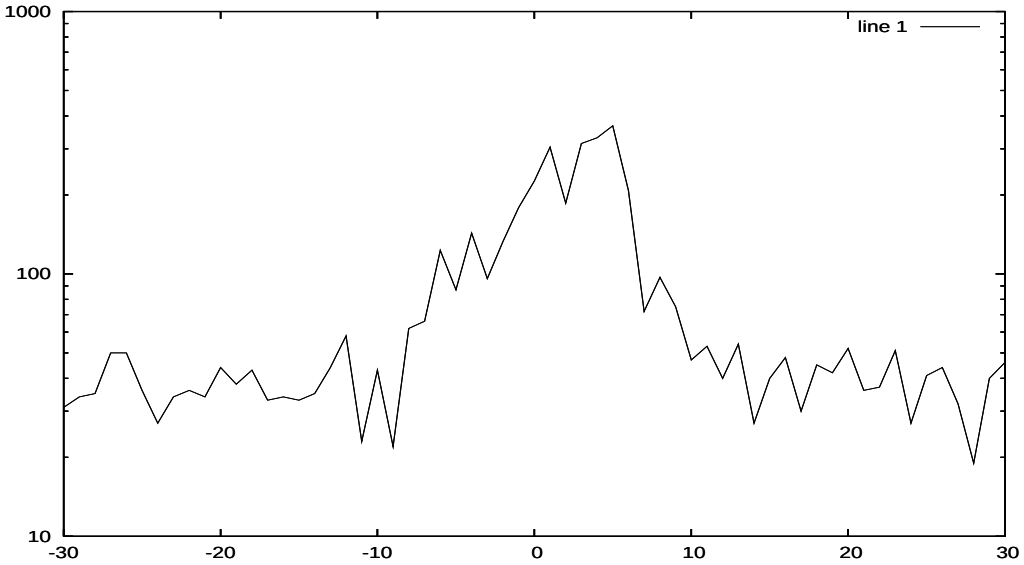


Figure 16.8: Distribution of sequence distances between donors and acceptors in arginine S-M bonds.

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.

16.4.2 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.

16.5 Exercises

Exercise 16.1 *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*

$$\phi(r) = \tfrac{1}{2} + \tfrac{3}{4}(r - \tfrac{1}{3}r^3) \tag{16.1}$$

and is in particular independent of ρ.

Exercise 16.2 *Repeat the analysis in Section 16.1 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 16.3 *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 16.1. 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 16.4 *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 17

Tree representations of data

The process of evolution is naturally expressed via a tree [75, 433]. For this reason, tree representations of data are widely used in biology. We have used a tree representation of relationships among different myoglobin molecules in various species in Figure 2.1. In that representation, the distances along the tree are intended to represent the evolutionary distance, as measured by sequence similarity, between different versions of myoglobin. This sequence distance was used to contrast differences in the numbers of dehydrons in the different versions of myoglobin.

Trees are not the only way to represent relationships in biology [220, 221, 252, 344, 426]. Simple clustering techniques [33, 344] are often used. On the other hand, one could represent distances by a general graph [220]. A compromise between a tree and a general graph is represented by 'reticulated' representations [252]. However, tree representations are very widely used in biology and thus deserve significant attention. In many cases, only the topology of the representation is significant [221], and we discuss some issues related to the uniqueness of the topology of tree representations.

We explain here some of the basic issues with tree representations from a mathematical point of view, mainly that they rarely represent the data exactly. There is a precise condition that must be satisfied for a set of distances to be representable by a tree, and generically this would not be satisfied. Moreover, different algorithms [33, 111, 116, 163, 176, 357] are commonly used, and they do not give the same trees in many cases. Although we are not able to explain the differences that may arise in general, we do give one hint of how this may be tolerated in practice. We also give various examples of interest as illustration.

Determining tree representations also suffers from challenges that are purely biological [41, 441, 443]. These mainly have to do with the difficulty of determining relationships between different biological entities for which the data is likely incomplete. This gets reflected in the mathematical problem of tree determination in that the definition of distance may be imprecise. Unfortunately, the compounding of two types of fuzziness does not tend to cancel the fuzziness, but perhaps the one makes the other irrelevant. We will derive at least one result that is robust with respect to uncertainties in the distances.

The objects represented by trees in biology can vary, from individual proteins to entire species. Our main application will be to compare different proteins. The notion of distance used to define

the trees will vary correspondingly, but in different directions. Even for a particular set of objects, such as proteins, it is possible to define potentially different notions of distance between individual proteins. For example, although the frequently used definition of distance derives from sequence similarly, we show that using particular features can also provide useful information. For completeness, we will begin with basic information about distance metrics and review the basic concepts of distance based on sequence alignment.

17.1 Distance metrics

The concept of distance is formalized mathematically in a **metric space**. Finite metric spaces are quite simple. They are characterized by having objects that can be labelled by integers $i = 1, \ldots, n$, and corresponding distances between the points given by a **distance matrix** $\mathcal{D}(i, j)$ where the individual values of the entries of the matrix are the distances in the space between objects i and j. For the time being, think of the objects as being different proteins and the distances some measure of similarity among them.

Several of the key properties of distance matrices are both easy to verify by inspection and easy to motivate. By definition, the diagonal of $\mathcal{D}$ is always zero, since the distance from one point to itself should be zero. Also, since they are distances, all of the entries are positive. Similarly, the matrix is assumed to be symmetric, reflecting the presumption that the cost of going from A to B is the same as going from B to A. In some cases, this might not be true, so a different kind of mathematics would need to be used.

The most important condition on a distance matrix is the **triangle inequality**. This condition encapsulates an important geometric condition and is not easily verified by inspection.

17.1.1 Triangle inequality

The triangle inequality for a distance matrix $\mathcal{D}$ requires that

$$\mathcal{D}(i, j) \leq \mathcal{D}(i, k) + \mathcal{D}(k, j) \tag{17.1}$$

for all $i, j, k = 1, \ldots, N$, where N is the total number of objects in the metric space. For example, if there are only three objects in the metric space (three proteins, say), then the general form of a distance matrix is

$$\begin{pmatrix} 0 & a & b \\ a & 0 & c \\ b & c & 0 \end{pmatrix} \tag{17.2}$$

where a, b, and c are positive parameters. Suppose that, without loss of generality, that c is the smallest:

$$c \leq \min\{a, b\}. \tag{17.3}$$

(If not, re-label the names of the elements of the metric space.) Then we leave as Exercise 17.2 that the triangle inequality implies that

$$|b - a| \leq c. \tag{17.4}$$

We leave as Exercise 17.3 to show that any coefficients a, b, and c satisfying the bounds (17.3) and (17.4) generate a distance matrix of the form (17.2) satisfying the triangle inequality. Thus the set of possible values for a, b, and c occupy a non-convex cone (with non-empty interior) in three dimensions defined by the bounds (17.3) and (17.4) (cf. Exercise 17.4).

17.1.2 Non-metric measurements

In many cases in biology, there are values $\widetilde{\mathcal{D}}(i,j)$ that represent relationships between objects i and j that are all positive but do not necessarily satisfy the triangle inequality (17.1). Thus they do not faithfully represent the 'cost' or 'distance' in going from i to j. For example, if $\widetilde{\mathcal{D}}(i,j) > \widetilde{\mathcal{D}}(i,k) + \widetilde{\mathcal{D}}(k,j)$, then the path

$$i \to k \to j \tag{17.5}$$

represents a cheaper way (with cost $\widetilde{\mathcal{D}}(i,k) + \widetilde{\mathcal{D}}(k,j)$) to get from i to j, rather than going directly. Thus $\widetilde{\mathcal{D}}$ no longer satisfies our general notion of a collection of direct distances. However, it is possible to define a metric related to $\widetilde{\mathcal{D}}$ that represents the relations faithfully, as follows.

Using the metaphor of distance, we can think of progressing from i to j via a finite sequence of steps like (17.5), i.e.,

$$i = k_0 \to k_1 \to \cdots \to k_r = j. \tag{17.6}$$

Then we define $\mathcal{D}(i,j)$ as the minimum cost over all possible paths of the type (17.6), viz.,

$$\mathcal{D}(i,j) = \min_{i=k_0,k_1,\ldots,k_r=j} \sum_{\ell=1}^{r} \widetilde{\mathcal{D}}(k_{\ell-1}, k_\ell). \tag{17.7}$$

Without loss of generality, we can assume that all paths satisfy $k_{\ell-1} \neq k_\ell$, since if they are equal there is no contribution to the sum in (17.7), and the corresponding paths with the repetitions of terms eliminated lead to the same value.

Lemma 17.1 *Suppose that the matrix $\widetilde{\mathcal{D}}$ is symmetric, has all off-diagonal entries positive and all diagonal entries zero. Then the matrix $\mathcal{D}$ defined by (17.7) is a metric (in particular, satisfies the triangle inequality) and*

$$\mathcal{D}(i,j) \leq \widetilde{\mathcal{D}}(i,j) \tag{17.8}$$

for all i and j.

Proof. Since one of the possible paths is $i = k_0, k_1 = j$, we have $\mathcal{D}(i,j) \leq \widetilde{\mathcal{D}}(i,j)$ for all i and j, which proves (17.8).

Let $\epsilon = \min\left\{\widetilde{\mathcal{D}}(i,j) \;\middle|\; i,j = 1,\ldots N\right\}$ and $\mu = \max\left\{\widetilde{\mathcal{D}}(i,j) \;\middle|\; i,j = 1,\ldots N\right\}$. Then

$$\sum_{\ell=1}^{r} \widetilde{\mathcal{D}}(k_{\ell-1}, k_\ell) \geq \epsilon r. \tag{17.9}$$

In view of (17.8), we can restrict to paths such that

$$\sum_{\ell=1}^{r} \widetilde{\mathcal{D}}(k_{\ell-1}, k_\ell) \leq \mu, \tag{17.10}$$

because paths with a larger sum will not change the value of the minimum in (17.7). So we can assume that $r \leq \mu/\epsilon$. Thus the number of paths can be assumed to be finite and the minimum in (17.7) is thus positive for all i and j. Moreover, we can assume that the minimum is attained by a particular path in (17.7):

$$\mathcal{D}(i,j) = \sum_{\ell=1}^{r} \widetilde{\mathcal{D}}(k_{\ell-1}, k_\ell), \tag{17.11}$$

for some path of the form (17.6).

The matrix $\mathcal{D}$ defined by (17.7) is symmetric because any path from i to j given via steps $i = k_0, k_1, \ldots, k_r = j$ gives a path from j to i given by $j = k_r, k_{r-1}, \ldots, k_1 = i$, and conversely. And the inequality (17.8) guarantees that $\mathcal{D}$ is zero on the diagonal.

Now let us confirm that the triangle inequality (17.1) holds for the matrix $\mathcal{D}$ defined by (17.7). Let i, j, and m be arbitrary. Let $i = k_0, k_1, \ldots, k_r = m$ be a path such that

$$\mathcal{D}(i,m) = \sum_{\ell=1}^{r} \widetilde{\mathcal{D}}(k_{\ell-1}, k_\ell). \tag{17.12}$$

Similarly, let $m = k'_0, k'_1, \ldots, k'_{r'} = j$ be a path such that

$$\mathcal{D}(m,j) = \sum_{\ell=1}^{r'} \widetilde{\mathcal{D}}(k'_{\ell-1}, k'_\ell). \tag{17.13}$$

Using the path $i = k_0, k_1, \ldots, k_r = m = k'_0, k'_1, \ldots, k'_{r'} = j$ as a possible path in (17.7) proves that

$$\mathcal{D}(i,j) \leq \sum_{\ell=1}^{r} \widetilde{\mathcal{D}}(k_{\ell-1}, k_\ell) + \sum_{\ell=1}^{r'} \widetilde{\mathcal{D}}(k'_{\ell-1}, k'_\ell) = \mathcal{D}(i,m) + \mathcal{D}(m,j). \tag{17.14}$$

QED

17.2 Sequence distance

We now consider notions of distance that are appropriate for entities that can be expressed as a sequence of letters, such as a protein (or DNA). The mathematical term for such a sequence is a **string**. A string is just an ordered list of letters coming from an alphabet Σ that is fixed at the beginning of the discussion. So for example, we can think of the protein alphabet {A,C,D, . . . ,T,Y,W} of twenty letters listed in Table 4.1. A string could correspond to a complete protein, or a fragment such as DRYYRE. The length of the strings are arbitrary but finite; the set of such strings is denoted by Σ^* [106, 176].

The simplest form of sequence comparison is based on comparing strings one letter at a time, and then summing up all of the individual comparisons. To start with, we need a metric on the set of letters in the alphabet Σ for our sets of sequences. Let D_Σ be a metric on the alphabet Σ. Then $D_\Sigma(\xi, \eta)$ measures the difference between the two letters ξ and η in Σ. We consider two ways to extend this to metrics d on the set of all finite strings Σ^*.

Many underlying metrics for amino acid residues are used in practice, and even standard metrics may undergo revision [392]. There are some obvious trends one would expect, such as D_Σ(Tyr,Gly) $>$ D_Σ(Tyr,Phe), or D_Σ(Ile,Leu) $<$ D_Σ(Arg,Asp). However, different models lead to different precise numbers for the distance between residues. We will not try to comment on this issue but rather assume that we are given a set of such numbers to work with.

17.2.1 Hamming distance

Given a metric D_Σ a metric on an alphabet Σ, we can define a metric on stings $\Sigma^N \subset \Sigma^*$ of length N by

$$d_H(\sigma, \tau) = \sum_{i=1}^{N} D_\Sigma(\sigma_i, \tau_i) \tag{17.15}$$

for all $\sigma, \tau \in \Sigma^N$, where σ_i denotes the i-th letter in the string σ, and similarly for τ. This can be extended to strings of different length in various ways, but we will keep things simple for the moment. It is elementary to show that d_H forms a metric on Σ^N for any N (see Exercise 17.8). The concept of the Hamming distance can be easily extended to other situations in which we sum up individual comparisons, as we discuss in Section 17.3.

17.2.2 Edit distance

The problems with Hamming distance are two fold. First of all, it is only defined on sequences of the same length. There are different ways to fix this, but the second problem is more serious. The insertion of a single letter in a sequence would shift every letter past it to the right by one and make the comparison useless in many cases. Consider for example the difference in spellings of the word 'color' (American) and 'colour' (British). Comparisons of English text would predict erroneous differences if we did not allow for the simple insertion of a letter that alters the spelling. Edit distance d_e is designed to fix this problem. We will describe how it works informally at first.

If two strings x and y differ only in the k-th position, then we set $d_e(x, y) = D_\Sigma(x_k, y_k)$. In general, when there are multiple replacements, string edit distance is based on just summing the effects. Thus it is the same as Hamming distance for simple replacements. However, string-edit distance also allows a different kind of change as well: insertion and deletion. For example, we can define $x_{\hat{k}}$ to mean the string x with the k-th entry removed. It might be that $x_{\hat{k}}$ agrees perfectly with the string y, and so we assign $d(x, y) = \delta$ where δ is the deletion penalty. Similarly, insertions of characters are allowed to determine edit distance. Clearly, if $y = x_{\hat{k}}$, then adding x_k to y at the k-th position yields x. Again, the effect of multiple insertions/deletions is additive, and this allows strings of different lengths to be compared, as we explain shortly.

The use of both replacements and instertion/deletions to determine edit distance means that an edit path from x to y is not unique. Edit distance is therefore defined by taking the minimum over all possible representations, as we define formally in (17.16). But this will not in general define a metric unless appropriate conditions on the gap penalty δ and alphabet metric D_Σ are satisfied. These conditions can be defined by extending the alphabet Σ and metric D_Σ to include a "gap" as a character, say "_" (let $\widetilde{\Sigma}$ denote the extended alphabet), and by assigning a distance $\delta = D_{\widetilde{\Sigma}}(x, \_)$ for each character x in the original alphabet. Although not strictly necessary, it is typical to have the 'gap' penalty be the same for all characters in the alphabet. We will make this assumption from now on. We now see how two strings of different length are compared: we just add sufficient _'s at the end of the shorter sequence to make it the same length as the longer one.

The edit distance d_e is derived from the extended alphabet distance $D_{\widetilde{\Sigma}}$ as follows. We introduce the notion of *alignment* $\mathcal{A}$ of sequences $(\tilde{\sigma}, \tilde{\tau}) = \mathcal{A}(\sigma, \tau)$ where $\tilde{\sigma}$ has the letters of σ in the same order but possibly with gaps _ inserted, and similarly for $\tilde{\tau}$. We suppose that $\tilde{\sigma}$ and $\tilde{\tau}$ have the same length even if σ and τ did not, which can always be achieved by adding gaps at one end or the other. Then we define

$$d_e(\sigma, \tau) = \min_{\mathcal{A}} \sum_i D_{\widetilde{\Sigma}}(\tilde{\sigma}_i, \tilde{\tau}_i) \qquad [(\tilde{\sigma}, \tilde{\tau}) = \mathcal{A}(\sigma, \tau)] \tag{17.16}$$

for any $\sigma, \tau \in \Sigma$. The minimum is over all alignments $\mathcal{A}$ and the sum extends over the length of the sequences. Fortunately, string-edit distance d_e, and even more complex metrics involving more complex gap penalties, can be computed efficiently by the dynamic programming algorithm [420]. Moreover, the following result shows that we have not left the realm of metric spaces by introducing gap penalties.

Theorem 9.4 of [420] says that d_e is a metric on strings of letters in Σ whenever $D_{\widetilde{\Sigma}}$ is a metric on the extended alphabet $\widetilde{\Sigma}$.

The simple string-edit distance d_e described here is useful in many contexts. However, more complex metrics would be required in other applications, including more complex ways of scoring sequences of gaps [420].

17.2.3 Two-letter alphabets

The simplest non-trivial example is an alphabet with two letters, say x and y, when there is only one distance $D_\Sigma(x, y)$ that is non-zero. The requirement that the triangle inequality hold for $D_{\widetilde{\Sigma}}$ reduces to three inequalities that can be expressed as

$$|D_{\widetilde{\Sigma}}(x, \_) - D_{\widetilde{\Sigma}}(y, \_)| \le D_\Sigma(x, y) \le D_{\widetilde{\Sigma}}(x, \_) + D_{\widetilde{\Sigma}}(y, \_). \tag{17.17}$$

Together with the condition that all distances be non-negative, we see that (17.17) characterizes completely the requirement for $D_{\widetilde{\Sigma}}$ to be a metric in the case of a two-letter alphabet Σ.

17.2.4 General alphabets

For a general alphabet Σ, if

$$\alpha \le D_\Sigma(x, y) \le 2\alpha \tag{17.18}$$

for all $x \neq y$ (including _) for some $\alpha > 0$, then D_Σ is a metric (that is, the triangle inequality holds). This is because

$$D_\Sigma(x, y) \leq 2\alpha \leq D_\Sigma(x, z) + D_\Sigma(z, y) \tag{17.19}$$

for any $z \in \Sigma$. One simple choice for a metric on letters is to choose $D_\Sigma(x, y) = 1$ for all $x \neq y$, and then to take $D_{\widetilde{\Sigma}}(x, \_) = 2$; the resulting $D_{\widetilde{\Sigma}}$ satisfies (17.18) for $\widetilde{\Sigma}$. However, condition (17.18) is far from optimal as the example (17.17) shows.

17.2.5 Distance versus score

Typically biologists prefer to work with a "score" that is large when two sequences are close as opposed to a distance which is small in such a case. The dynamic programming algorithm can equivalently be used to mininmize the distance or maximize a score. There is a formal correspondence that can be made between scores and distances, as follows [420]. It is clear that the minimization problem (17.16) is equivalent to the maximization problem

$$s_e(\sigma, \tau) = \max_{\mathcal{A}} \sum_i a - cD_{\widetilde{\Sigma}}(\tilde{\sigma}_i, \tilde{\tau}_i) \qquad [(\tilde{\sigma}, \tilde{\tau}) = \mathcal{A}(\sigma, \tau)] \tag{17.20}$$

for any constants a and c, where the maximum is over all alignments $\mathcal{A}$ and the sum extends over the length of the sequences. Thus any distance matrix is equivalent to a scoring matrix where we the scoring matrix is defined by

$$S_{\widetilde{\Sigma}}(x, y) = a - cD_{\widetilde{\Sigma}}(x, y), \tag{17.21}$$

or equivalently, the distance matrix is derived from the scoring matrix via

$$D_{\widetilde{\Sigma}}(x, y) = \big(a - S_{\widetilde{\Sigma}}(x, y)\big)/c. \tag{17.22}$$

A necessary condition for a scoring matrix to correspond to a metric according to (17.22) is that the scores $S_{\widetilde{\Sigma}}(x, x)$ for all $x \in \widetilde{\Sigma}$ be the same value (a). This is commonly the case for DNA alignment algorithms, but not for protein sequence scoring matrices. For example, the score S(Trp,Trp) is commonly much larger than the score S(Ser,Ser). Since Trp is much less common than Ser (cf. Table 7.2), this makes sense, but it is nevertheless difficult to interpret scores as related in a simple way to distances.

17.3 Feature distance

Another way to define metrics for biological entities is by comparing feature differences [267]. 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 homologs, and thus a distance metric can encapsulate such differences.

	Ack	Kit	Abl	Lck	Chk1	Pdk1
EGFR	2	2	3	3	4	4
Ack		1	4	4	5	5
Kit			4	4	5	5
Abl				1	4	4
Lck					4	4
Chk1						1

Table 17.1: Dehydron distances for seven homologous tyrosine kinase proteins. PDB files corrsponding to the abbreviations in the table are: Abl (1FPU), Ack1 (1U54), Chk1 (1IA8), C-kit (1T45), EGFR (1M17), Lck (3LCK), Pdk1 (1UVR).

17.3.1 H-bond distance

One feature that can be compared is the hydrogen bond location for homologous proteins. This can be done in a variety of ways, but one way is to first form an initial alignment and then consider the resulting alignment of hydrogen bonds. This first step is analogous to the transition step from Hamming distance (17.15) to edit distance (17.16), and it will depend on the alignment chosen. Thus it is possible to minimize over all such alignments. But for now, we consider the computation for just one fixed alignment.

Now that the two proteins have been aligned, we define indicator matrices to record the presence of hydrogen bonds. An indicator matrix for each protein is constructed that is indexed by all pairs of aligned residues. It consists only of zeros unless two residues are paired by a hydrogen bond, in which case the corresponding entry is one. More precisely, a matrix H_{ij} is constructed by choosing $H_{ij} = 1$ if residues i and j are paired by a hydrogen bond and $H_{ij} = 0$, otherwise. If the proteins are named n and m respectively, then we now have two hydrogen bond indicator matrices, $H(m)$ and $H(n)$.

Now, a Hamming-type distance [132] can be defined based on the number of disagreements between two such indicator matrices. More precisely, we define a distance $\mathcal{D}(m, n)$ by

$$\mathcal{D}(m, n) = \sum_{i<j} |H_{ij}(m) - H_{ij}(n)|, \tag{17.23}$$

where m and n are the names of the different proteins. Note that we have restricted to indices $i < j$ due to the symmetry of the hydrogen bond indicator matrices. If desired, the definition (17.23) could be refined by minimizing over all possible alignments of m and n.

17.3.2 Other metrics

Metrics based on comparison of hydrogen bond structures can also be refined by limiting to certain classes of hydrogen bonds. A very successful use of this idea is to restrict to the underwrapped hydrogen bonds (the dehydrons) [141]. Since this metric reflects defects in the packing of the

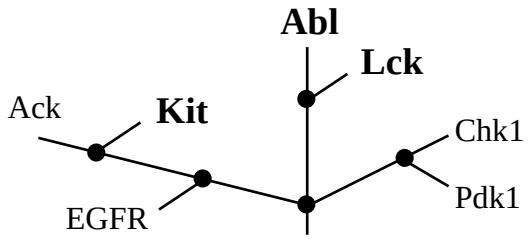


Figure 17.1: Packing similarity tree (PST) for the seven structurally aligned paralogs of Bcr-Abl, listed as 'Abl' in the diagram, based on the dehydron distances in Table 17.1. The paralogs in **bold** have the most similar packing in the region that aligns with the Gleevec-wrapped region in Bcr-Abl and are also primary targets of this inhibitor [114].

hydrogen bonds by hydrophobic groups, it is called the 'packing distance.' The packing distance for seven homolgous tyrosine kinase proteins [275] is depicted in Table 17.1.

Another example of restricting to special hydrogen bonds is depicted in Table 17.2. Here the distance is based on the intermolecular hydrogen bonds formed between the antibody and the antigen in homologous antibody-antigen structures.

A distance based on nonpolar regions on proteins (a.k.a., hydrophobic patches) can similarly be defined [73]. A pharmacological distance matrix can also be constructed based on affinity profiling against specified drugs [114, 132].

Many other features could potentially be employed to define useful metrics. For example, geometric [32] features such as surface curvature can be measured by

$$\mathcal{D}(m,n) = \sum_i |K_i(m) - K_i(n)| \tag{17.24}$$

where K_i is the curvature of the protein-ligand interface at the i-th position of aligned proteins m, n.

17.3.3 Relating distances to trees

Such distance matrices can be used to construct dendrograms (or trees) and are implicitly behind the dendrograms commonly seen [114, 150]. Such trees are often constructed by using standard algorithms for phylogenetic analysis [176]. In some cases, this relationship is easy to comprehend as is the case for Table 17.1 and Figure 17.1. The latter is a particularly simple tree in which the distance between the leaves and the parent node is exactly half the distance between the parent nodes themselves. If we take that distance to be one, then we obtain the distances given in Table 17.1. Correspondingly, we can relate the distances given in Table 17.1 to the number of parent nodes on the shortest path between leaves of the tree.

Since Table 17.1 encodes what we have called 'packing' distances, because it measures discrepancies in the arrangements of dehydrons, we refer to the tree in Figure 17.1 as a 'packing' similarity tree (PST). The resulting tree gives a visual indication of 'closeness' for the various proteins. 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.

The key concept of tree representation of similarity is that distance in the tree is intended to reflect distance as encoded in the distance matrix, as is the case in Table 17.1 and Figure 17.1. However, we will see that in general this does not easily hold. There are certain restrictions that apply to the distance matrix to get an exact representation [176]. There is often no tree that exactly represents the distance data, and the set of trees that closely approximate it is often not well defined. We will refer to this as a lack of uniqueness in the tree representation, althought the situation is really more complex than that. Thus the general use of trees to represent distance data is problemmatic.

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 [132]. In general, if a distance matrix is to be represented faithfully by a tree, it must satisfy a 'four-point condition' [176] 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 [132].

17.4 Tree representation of metrics

A tree naturally defines a metric space for the leaves of the tree, where the distance between leaves is the shortest path in the tree. We now invert this observation and ask which metrics can be represented by trees.

17.4.1 Three point metrics

A distance matrix $\mathcal{D}(i,j)$ for three points can always be represented by a tree. There are only three unique values in the distance matrix: $\mathcal{D}(1,2)$, $\mathcal{D}(2,3)$ and $\mathcal{D}(1,3)$. A tree that connects the points 1, 2, and 3 via a central node (call it 0) has only three distances to define it: the distance δ_i from 0 to node i for $i = 1, 2, 3$. The relatinship between the two representations of distance is $\mathcal{D}(i,j) = \delta_i + \delta_j$. It is easy to see that this 3×3 system of equations is invertible.

However, in general a distance matrix $\mathcal{D}(i,j)$ for $k > 3$ points cannot be exactly represented by the distances in a tree. We will see this explicitly in the case $k = 4$, but it is easy to see why there is a difficulty in this case. The tree of interest will take the form in Figure 17.4. There are only five internal distances available in this tree. But there are six different distinct values in a distance matrix $\mathcal{D}(i,j)$ for $k = 4$. Even if these six values are chosen to satisfy the triangle inequality, there is one too many of them to be able to be matched by five parameters.

Different algorithms are used in practice, and they have the property that given a distance matrix $\mathcal{D}(i,j)$ they will produce an answer. It is not known in general what the relattionship is between the products of different popular algorithms. However, for $k = 4$ we can at least give an interpretation of the extent of ambiguity.

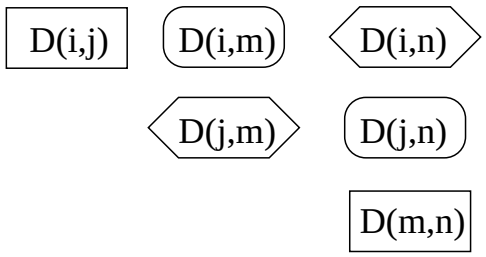


Figure 17.2: Entries in the distance matrix which are constrained by the four point condition.

17.4.2 Four point condition

In general, if a distance matrix is to be represented faithfully by a tree, it must satisfy the following **four-point condition** [36, 176, 362]

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \max\{\mathcal{D}(i,m) + \mathcal{D}(j,n), \mathcal{D}(j,m) + \mathcal{D}(i,n)\}, \tag{17.25}$$

for all i, j, m, and n. The four-point condition generalizes the triangle inequality (take $m = n$). For a distance matrix satisfying the triangle inequality, it suffices to take i, j, m, and n distinct. The relationship between the various values are depicted in Figure 17.2. In the case of only four points, the entire (upper-triangular part of the) distance matrix is depicted in Figure 17.2, and the entries making up the three terms in (17.25) are enclosed in like enclosures (rectangle, oval, hexagon). In the case of a general distance matrix, the same picture obtains after eliminating all rows except i, j and m, and all columns except j, m and n.

We will see that generically most distance matrices do not satisfy the four-point condition. This implies that typical biological trees will not rigorously represent a given biological distance matrix.

Definition 17.1 *A matrix that satisfies the four point condition is called* **additive**.

The following theorem may be found in [176].

Theorem 17.1 *A distance matrix can be represented by distances in a tree if and only if it satisfies the four point condition (i.e., it is additive).*

There is a simple interpretation of the four point condition, as follows. Given distinct values of i, j, m, and n, define parameters A, B, and C by

$$A = \mathcal{D}(i,j) + \mathcal{D}(m,n), \quad B = \mathcal{D}(i,m) + \mathcal{D}(j,n), \quad \text{and} \quad C = \mathcal{D}(j,m) + \mathcal{D}(i,n). \tag{17.26}$$

When we have three values A, B, and C, it is natural to expect that there are three distinct values given by

$$\min\{A,B,C\} \leq \text{mid}\{A,B,C\} \leq \max\{A,B,C\} \tag{17.27}$$

Here, the 'mid' function picks out the value between the min and the max. The four-point condition is the requirement that

$$\text{mid}\{A,B,C\} = \max\{A,B,C\}. \tag{17.28}$$

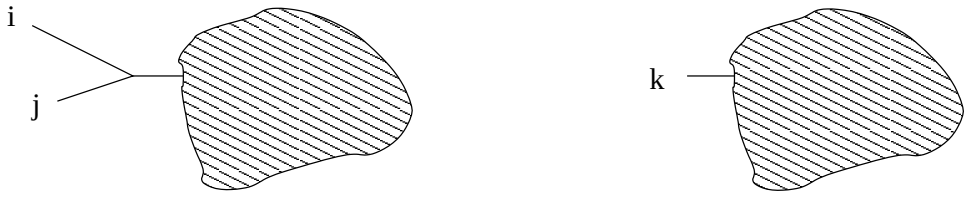


Figure 17.3: Original tree (left) and reduced tree (right) where nodes i and j have been removed and replaced by node k. The corresponding distances are defined via (17.30).

Otherwise said, of the three values A, B, and C, the two largest must be equal for the four-point condition to hold.

As an example, consider the intermolecular hydrogen bond distance data presented in Table 17.2. For a metric space with only four points, there is only one choice for the distinct values of i, j, m, and n, although relabelling is possible. For the data in Table 17.2, we find that

$$\{A, B, C\} = \{8, 12, 12\}, \tag{17.29}$$

so Table 17.2 is an additive matrix.

It is possible to motivate the four-point condition by considering a simple algorithm for reducing the size of the metric space. This algorithm is only suggestive, but it does make clear why the condition arises, and it also motivates one of the widely used algorithms for determining trees from data.

17.4.3 A reduction algorithm

Consider the following algorithm for constructing a tree from a distance matrix. Loop over all pairs i, j. Suppose there is a graph rerpresentation with nodes i and j appearing as leaves of a parent node as depicted on the left-hand side of Figure 17.3. Call the parent node k, where k is an index not being used in the current indexing scheme. Then define

$$\mathcal{D}(m,k) = \mathcal{D}(k,m) = \tfrac{1}{2}\left(\mathcal{D}(i,m) + \mathcal{D}(j,m) - \mathcal{D}(i,j)\right) \quad \forall m \neq i, j. \tag{17.30}$$

Create a new discrete space by eliminating i and j and adding k; in terms of the distance matrix, we eliminate the i and j rows and add the new information defined by (17.30). By the triangle inequality, this new matrix is non-negative. Moreover, if we find $\mathcal{D}(k,m) = 0$ we can take $k = m$ and avoid the addition to the discrete space, so we can assume that this new matrix is non-degenerate.

This new matrix satisfies the triangle inequality, that is $\mathcal{D}(k,m) \leq \mathcal{D}(k,n) + \mathcal{D}(n,m)$ for all n. This is equivalent to

$$\left(\mathcal{D}(i,m) + \mathcal{D}(j,m) - \mathcal{D}(i,j)\right) \leq \left(\mathcal{D}(i,n) + \mathcal{D}(j,n) - \mathcal{D}(i,j)\right) + 2\mathcal{D}(n,m), \tag{17.31}$$

or equivalently

$$\mathcal{D}(i,m) + \mathcal{D}(j,m) \leq \mathcal{D}(i,n) + \mathcal{D}(j,n) + 2\mathcal{D}(n,m). \tag{17.32}$$

But the triangle inequality for the original matrix implies this: we just add

$$\mathcal{D}(i,m) \leq \mathcal{D}(i,n) + \mathcal{D}(n,m) \tag{17.33}$$

to

$$\mathcal{D}(j,m) \leq \mathcal{D}(j,n) + \mathcal{D}(n,m) \tag{17.34}$$

to prove (17.32).

The reduction to a smaller metric space just described motivates the popular UPGMA algorithm. The algorithm proceeds by clustering nodes i and j for which $\mathcal{D}(i,j)$ is smallest. The heuristic is that smaller $\mathcal{D}(i,j)$ values should mean that i and j are closer in the tree (which is correct) and (which is not correct) the closest points must be children of the same parent node in the tree. We give an example to show how this fails in Section 17.5.2.

17.4.4 Obstruction to reduction

The difficulty arises in the assignment of the distances between the new point and the deleted points. If all were well, we would have

$$\mathcal{D}(i,k) = \mathcal{D}(i,m) - \mathcal{D}(m,k) = \tfrac{1}{2}\left(\mathcal{D}(i,m) - \mathcal{D}(j,m) + \mathcal{D}(i,j)\right), \tag{17.35}$$

for any m, and similarly for $\mathcal{D}(j,k)$:

$$\mathcal{D}(j,k) = \mathcal{D}(j,m) - \mathcal{D}(m,k) = \tfrac{1}{2}\left(\mathcal{D}(j,m) - \mathcal{D}(i,m) + \mathcal{D}(i,j)\right). \tag{17.36}$$

Since m is arbitrary in (17.35), it must hold as well for any other node n replacing m. Since the left-hand side of (17.35) remains unchanged, we must have

$$\mathcal{D}(i,m) - \mathcal{D}(j,m) + \mathcal{D}(i,j) = \mathcal{D}(i,n) - \mathcal{D}(j,n) + \mathcal{D}(i,j) \tag{17.37}$$

for any m and n, which is the same as saying

$$\mathcal{D}(i,m) + \mathcal{D}(j,n) = \mathcal{D}(i,n) + \mathcal{D}(j,m), \tag{17.38}$$

which we will see is equivalent to the 'mid=max' interpretation of the four-point condition.

To complete the derivation of the four-point condition from (17.38), we note that the common value in (17.38) may be written as (see Exercise 17.10)

$$\mathcal{D}(i,m) + \mathcal{D}(j,n) = \mathcal{D}(i,n) + \mathcal{D}(j,m) = \mathcal{D}(i,j) + \mathcal{D}(k,m) + \mathcal{D}(k,n). \tag{17.39}$$

By the triangle inequality,

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \mathcal{D}(i,j) + \mathcal{D}(m,k) + \mathcal{D}(k,n). \tag{17.40}$$

Combining (17.40) and (17.39) completes the derivation of the four-point condition. Moreover, it proves that $\mathcal{D}(i,j) + \mathcal{D}(m,n)$ has to be the 'min' value of these ABC values, for all m and n. This appears at first to be stronger than the four point condition, so we state it formally.

Lemma 17.2 *Suppose that the matrix $\mathcal{D}$ can be represented as a tree. Then there are indices i and j such that*

$$\mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \min\{\mathcal{D}(i,m) + \mathcal{D}(j,n), \mathcal{D}(i,n) + \mathcal{D}(j,m)\} \tag{17.41}$$

for all m and n.

The problem of identifying indices i and j such that the reduction algorithm can be applied is not simple, and there are different ways to identify them for an additive distance matrix. Moreover, the different methods lead to different approximation algorithms in the case of non-additive distances. However, one algorithm derives directly from Lemma 17.2 [162] (also see page 326 [362]). Define the expression

$$\begin{aligned} P_{ij} =& \#\big\{m,n \neq i,j \;\big|\; \\ & \mathcal{D}(i,j) + \mathcal{D}(m,n) \leq \min\{\mathcal{D}(i,m) + \mathcal{D}(j,n), \mathcal{D}(i,n) + \mathcal{D}(j,m)\big\}, \end{aligned} \tag{17.42}$$

where # means the cardinality of the set. The ADDTREE algorithm [362] consists of choosing i,j such that P_{ij} is maximized. In the definition of P_{ij} it is sufficient to assume that $m < n$. For an additive distance, Lemma 17.2 guarantees that there is some pair i,j such that the maximum possible value of P_{ij} is attained, so the maximization algorithm will always yield a pair satisfying (17.41). For non-additive distances, the ADDTREE algorithm also provides good approximations, as we will discuss shortly. An alternative way to express the ADDTREE algorithm is to define

$$R_{ij} = \#\left\{m,n \neq i,j \;\middle|\; \mathcal{D}(i,j) + \mathcal{D}(m,n) > \mathcal{D}(i,m) + \mathcal{D}(j,n)\right\}. \tag{17.43}$$

Then i,j are suitable neighbors if and only if $R_{ij} = 0$; the ADDTREE algorithm corresponds to minimizing R_{ij} for $i \neq j$ (cf. Exercise 17.11). Note that the computation of P can be simplified by assuming $m < n$, but the computation of R requires all m,n to be considered.

There is another condition that is known to determinine suitable indices i and j for additive distances. The condition is to minimize the expression

$$\begin{aligned} Q_{ij} =& (N-2)\mathcal{D}(i,j) - \sum_{k=1}^{N} \mathcal{D}(i,k) - \sum_{k=1}^{N} \mathcal{D}(j,k) \\ =& (N-4)\mathcal{D}(i,j) - \sum_{k\neq i,j;k=1}^{N} \mathcal{D}(i,k) + \mathcal{D}(j,k), \end{aligned} \tag{17.44}$$

where N is the total number of points in the metric space. Although discovered some time ago [357], the condition based on minimizing (17.44) has received a great deal of attention [163] and continues to be of interest [13, 12, 108, 111, 288]. The exact form (17.44) is due to Studier and Kepler [162]. The condition (17.44) is known to be unique [54] among certain functionals for determinining indices i and j.

The algorithm we have described for constructing a tree by recursively reducing the size of the metric space using the condition (17.44) is called **neighbor joining** (NJ). The neighbors being joined are i and j. The choice given by minimizing (17.44) is widely used, but not universal.

The UPGMA algorithm uses the far simpler heuristic of minimizing $\mathcal{D}(i,j)$. We explain why this approach leads to the wrong answer in Section 17.5.2 even for an additive metric.

The leaves i and j identified as above not only allow the reduction algorithm to construct a tree for an additive metric, they also are good choices in the case that a metric is not additive [12, 13, 54, 108, 111, 163, 288, 362]. Although we will see that there are many interesting biological metrics that are additive, in general this is not the case. Thus it is of interest to understand to what extent we can approximate a general distance matrix via an additive metric [33].

It is possible to assess theoretically the approximation quality for both neighbor joining, using the criterion of minimizing (17.44) [54, 108, 111, 163, 288], and the ADDTREE algorithm [362], using the criterion of maximizing (17.42). It is known that both of these algorithms are stable in maximum norm, as we now explain.

Define the ℓ^∞-norm for distance matrices:

$$\|\mathcal{D}\|_{\ell^\infty} = \max_{i<j} |\mathcal{D}(i,j)|. \tag{17.45}$$

Let f denote the mapping from a distance matrix $\mathcal{D}$ to an additive tree $\mathcal{T}$ given by either of the algorithms based on (17.42) or (17.44). For any such tree, let $\epsilon(\mathcal{T})$ denote the length of the shortest edge in the tree.

Theorem 17.2 *[362] For all distance matrices $\mathcal{D}'$ such that*

$$\|\mathcal{D} - \mathcal{D}'\|_{\ell^\infty} < \tfrac{1}{2}\epsilon(\mathcal{T}), \tag{17.46}$$

then the algorithms NJ and ADDTREE form the same tree that is,

$$f(\mathcal{D}') = f(\mathcal{D}). \tag{17.47}$$

This theorem says that the tree formation algorithms are continuous with respect to small perturbations in the data. It is known that Theorem 17.2 is sharp, in that there are $\mathcal{D}$ and $\mathcal{D}'$ such that $\|\mathcal{D} - \mathcal{D}'\|_{\ell^\infty} = \frac{1}{2}\epsilon(\mathcal{T})$ and $f(\mathcal{D}') \neq f(\mathcal{D})$ [362]. However, the set of distance matrices leading to the same additive tree according to a given algorithm is a more complex set [108].

Operationally, one may not know a priori whether a given distance matrix $\mathcal{D}'$ is close to being additive or not. But applying ADDTREE or NJ [271], we can obtain an additive tree $\mathcal{T} = f(\mathcal{D}')$, and we can form the corresponding additive matrix $\mathcal{D}$ that matches this tree. If $\|\mathcal{D}-\mathcal{D}'\|_{\ell^\infty} < \frac{1}{2}\epsilon(\mathcal{T})$ then we can have good confidence that this tree well represents the data $\mathcal{D}'$. There are also external measures to determine which trees should be picked to represent non-additive disatance matrices [96, 281].

In many cases, what is of most interest is not the lengths of the edges in the tree, but rather just the topology of the tree [33]. We now consider one simple case in which we can understand this approximation problem in some detail.

	1DQJ	1C08	1NDG
1NDM	3	7	6
1DQJ		6	5
1C08			5

Table 17.2: Distance matrix for intermolecular hydrogen bond interactions between antibody and antigen in the PDB complexes: a=1NDM, b=1DQJ, c= 1C08, d= 1NDG. Zeros on the diagonal and the redundant lower triangular part of the matrix have been omitted.

17.4.5 The ABC theorem

It is possible to show that the topology of the tree representation of four points is essentially unique under very mild conditions, as follows. Consider the three independent quantities that figure in the four point condition:

$$\begin{aligned} A &= \mathcal{D}(i,m) + \mathcal{D}(j,n) \\ B &= \mathcal{D}(i,n) + \mathcal{D}(j,m) \\ C &= \mathcal{D}(i,j) + \mathcal{D}(n,m) \end{aligned} \tag{17.48}$$

based on the three ways to partition the index set $\{i,j,m,n\}$ into distinct pairs. These quantities determine the topology of the tree representations, as follows. There are four distinct cases. Three of them involve two internal nodes and one internal edge, and are categorized by the following three distinct possibilities for additive matrices: $A = B > C$, $B = C > A$, and $C = A > B$. The fourth tree corresponds to $A = B = C$. Note that when $A = B = C$, the tree representing the distance matrix is a star. That is, there is one internal node k, and four edges joining the four indices to k.

We will show that even in the case that a matrix is not additive, a unique assignment of one of these topology classes is possible in most cases.

Suppose $\mathcal{D}$ is a general distance matrix that is not necessarily additive. Without loss of generality, by renaming the indices if necessary, we can assume that the terms are ordered:

$$A \geq B \geq C. \tag{17.49}$$

The four-point condition can now be stated simply: $A = B$. In this case, the distance matrix can be represented exactly by a tree. Now we consider the other case, that $A > B$. First, we define the ℓ^1-norm for distance matrices:

$$\|\mathcal{D}\|_{\ell^1} = \sum_{i<j} |\mathcal{D}(i,j)| \tag{17.50}$$

Note that we allow for negative entries, as we intend to apply the norm to differences of distance matrices.

Theorem 17.3 *Suppose that $A > B$. Then*

$$\inf\left\{ \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \;\middle|\; \mathcal{D}' \text{ satisfies the four-point condition} \right\} = A - B. \tag{17.51}$$

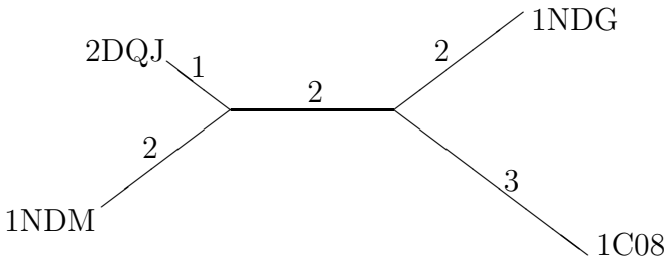


Figure 17.4: Tree representation of the distance matrix in Table 17.2.

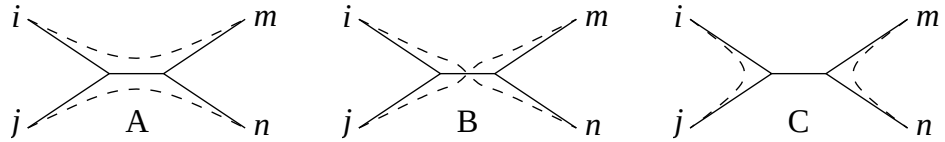


Figure 17.5: Relationship between the values A, B and C and the corresponing additive tree in the case that $C < \min\{A, B\}$. The dashed lines indicate the distances that correspond to A, B and C.

Moreover, if $B > C$, then all additive distance matrices $\mathcal{D}'$ which satisfy

$$\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B \tag{17.52}$$

have trees with the same topology, cf. Figure 17.5.

First of all, we note that there can be multiple 'best' distance matrices $\mathcal{D}'$ in (17.51). That is, we cannot define the 'best' tree in a simple way by just minimizing the ℓ^1 distance. However, we do assert in the second part of the theorem that the topology of all these trees is the same, provided that $B > C$.

When $B = C$, there is an ambiguity in representing $\mathcal{D}$ since there are additive matrices $\mathcal{D}'$ all equally close in ℓ^1 norm with different topology types. We leave as an exercise that there is a matrix $\mathcal{D}'$ with $A' = B' = C' = B = C$, as well as two others: one with $A' = B' = A$ and $C' = B = C$ and the other with with $A' = C' = A$ and $B' = B = C$.

Proof. To prove these assertions, we first show that

$$\inf\left\{\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \;\middle|\; \mathcal{D}' \text{ satisfies four-point condition}\right\} \leq A - B. \tag{17.53}$$

To so so, we simply need to exhibit a $\mathcal{D}'$ which satisfies the four-point condition and $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B$. We can do this if we keep $A' = A$ and increase B' to be equal to A. For example, we can set

$$\mathcal{D}'_{in} = \mathcal{D}_{in} + A - B, \tag{17.54}$$

leaving all other entries of $\mathcal{D}'$ the same as for $\mathcal{D}$. Thus by explicit construction, we have $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B$. Similarly, since we also have $A' = A = B'$, $\mathcal{D}'$ satisfies the four point condition.

Now we demonstrate the other inequality. We can clearly write

$$|A - A'| + |B - B'| + |C - C'| \leq \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} \tag{17.55}$$

for any distance matrices. Now suppose it were the case that for some any $\mathcal{D}'$ we have $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} < A - B$. Then by (17.55) we have

$$\begin{aligned} B' - A' + A - B =& A - A' + B' - B \\ \leq& \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} < A - B \end{aligned} \tag{17.56}$$

from which we conclude that $B' < A'$, so $\mathcal{D}'$ cannot satisfy the four point condition. This completes the proof of the equality (17.51).

Now suppose that $\mathcal{D}'$ is additive and satisfies (17.52). Then we want to show that

$$A \geq A' = B' \geq B. \tag{17.57}$$

Suppose that $A < B'$. Then $B' - B > A - B$ and applying (17.55) we find that $\|\mathcal{D} - \mathcal{D}'\|_{\ell^1} > A - B$. On the other hand, if $A' < B$ then $A - A' > A - B$, and again (17.55) yields a contradiction. Thus (17.57) has to hold.

Applying (17.57) in (17.55), we get

$$\begin{aligned} A - B =& A - A' + B' - B \\ =& |A - A'| + |B - B'| \\ \leq& |A - A'| + |B - B'| + |C - C'| \\ \leq& \|\mathcal{D} - \mathcal{D}'\|_{\ell^1} = A - B \end{aligned} \tag{17.58}$$

which means that equality holds throughout the experession (17.58), so we must have $C = C'$. In particular, we conclude that $A' = B' \geq C'$. If $B > C$, then we also have $B' > C'$.

Now it is easy to show that all additive matrices with $A' = B' > C'$ have the same topology. **QED**

17.5 Approximate algorithms

Since we see that there is a unique topology for trees representing generic four-point metric spaces, it is of interest to ask what various approximation algorithms compute in this simple case. We will see that neighbor joining always yields a tree with the correct topology. However, we will see that UPGMA does not, even in the case of an additive distance matrix.

17.5.1 What neighbor joining does

Neighbor joining chooses vertices based on the expression (17.44). So it makes sense to related those quantities to the determinants of topology. We use the notation of Section 17.4.5, and in particular that of Figure 17.5. Noting that $n = 4$, we can use the second form of (17.44) to see that

$$Q_{ij} = Q_{mn} = -(\mathcal{D}(i,m) + \mathcal{D}(i,n)) - (\mathcal{D}(j,m) + \mathcal{D}(j,n)) = -(A + B). \tag{17.59}$$

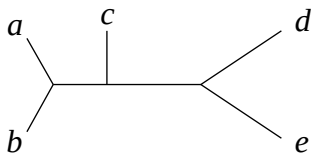


Figure 17.6: Tree for additive distance matrix $\mathcal{D}$ in (17.61) [289].

Similarly, it is easy to compute that

$$Q_{im} = Q_{jn} = -(B + C) \qquad \text{and} \qquad Q_{in} = Q_{jm} = -(A + C). \tag{17.60}$$

When $C \leq \min\{A, B\}$, then the indices that minimize the expression (17.44) are (i, j) and (m, n). Thus neighbor joining will choose the tree shown in Figure 17.5.

It is also clear that ADDTREE will also give the same tree, since it is based on ordering A, B and C. However, ADDTREE and neighbor joining will not give the same results [289] on the data in (17.61); $\mathcal{D}$ is additive, and $\mathcal{D}'$ is not:

$$\mathcal{D} = \begin{pmatrix} & b & c & d & e \\ a & 2 & 3 & 6 & 6 \\ b & & 3 & 6 & 6 \\ c & & & 5 & 5 \\ d & & & & 4 \end{pmatrix} \qquad \mathcal{D}' = \begin{pmatrix} & b & c & d & e \\ a & 2 & 3 & 6 & \mathbf{3} \\ b & & 3 & 6 & 6 \\ c & & & 5 & 5 \\ d & & & & 4 \end{pmatrix}. \tag{17.61}$$

The only difference between $\mathcal{D}$ and $\mathcal{D}'$ is that $\mathcal{D}(d, e) \neq \mathcal{D}'(d, e)$, as shown in bold in (17.61). The resulting tree for $\mathcal{D}$ is depicted in Figure 17.6.

The Q matrices (17.44) for $\mathcal{D}$ and $\mathcal{D}'$ in (17.61) are

$$Q(\mathcal{D}) = \begin{pmatrix} & b & c & d & e \\ a & -25 & -21 & -17 & -23 \\ b & & -24 & -20 & -17 \\ c & & & -22 & -19 \\ d & & & & \mathbf{-27} \end{pmatrix} \qquad Q(\mathcal{D}') = \begin{pmatrix} & b & c & d & e \\ a & -28 & -24 & -20 & -20 \\ b & & -24 & -20 & -20 \\ c & & & -22 & -22 \\ d & & & & \mathbf{-30} \end{pmatrix}, \tag{17.62}$$

indicating the same choice for coalescence of neighbors, d and e. The R matrices for $\mathcal{D}$ and $\mathcal{D}'$ in (17.43) are

$$R(\mathcal{D}) = \begin{pmatrix} & b & c & d & e \\ a & \mathbf{0} & 2 & 3 & 3 \\ b & & 2 & 3 & 3 \\ c & & & 2 & 2 \\ d & & & & \mathbf{0} \end{pmatrix} \qquad R(\mathcal{D}') = \begin{pmatrix} & b & c & d & e \\ a & 1 & 3 & 5 & 2 \\ b & & 1 & 3 & 6 \\ c & & & 2 & 4 \\ d & & & & \mathbf{0} \end{pmatrix}. \tag{17.63}$$

Note that there are two choices for $\mathcal{D}$ for coalescence of neighbors, (a, b) and (d, e).

	b	c	d
a	3	5	4
b		4	5
c			7

Table 17.3: Additive distance matrix for which UPGMA gives the wrong tree.

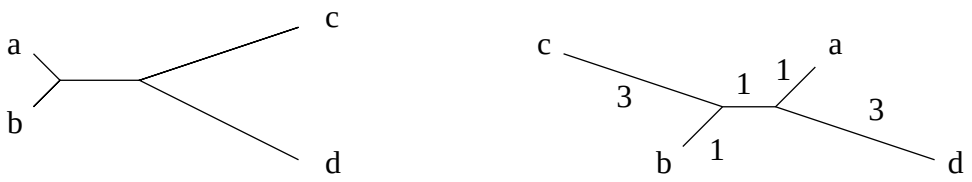


Figure 17.7: UPGMA tree (left) and additive tree for distance matrix in Table 17.3.

17.5.2 What UPGMA does

The UPGMA algorithm [176] applied to a distance matrix will coallesce the closest points, that is, the two for which the entry in the distance matrix is smallests, say $\mathcal{D}(i,j)$. One might hope that UPGMA would find the correct tree for an additive matrix.

The nearest neighbors in the tree for an additive matrix are the indices that combine to form the term C that is the smallest of the three terms involved in the four point condition: $C < B = A$. However, even though it may hold that $\mathcal{D}(i,j)$ is the smallest entry in the distance matrix, $C = \mathcal{D}(i,j) + \mathcal{D}(m,n)$ is not smaller than A or B. In this case, UPGMA finds the wrong tree. For example, consider the distance matrix in Table 17.3 with $A = 5 + 5 = 10$, $B = 4 + 4 = 8$, and $C = 3 + 7 = 10$. In Figure 17.7 we see both the tree that UPGMA will generate (left) for the data together with the tree (right) that precisely represents this additive matrix.

17.6 Exercises

Exercise 17.1 *Show that the distance matrices in Table 17.1, Table 17.2, and Table 17.3 satisfy the triangle inequality.*

Exercise 17.2 *Prove that any distance matrix (17.2) must satisfy $|b - a| \leq c$ provided that $c \leq \min\{a, b\}$. (Hint: apply the triangle inequality (17.1).)*

Exercise 17.3 *Suppose that a distance matrix (17.2) satisfies (17.3) and (17.4). Prove that it satisfies the triangle inequality (17.1). (Hint: reverse the derivation in Exercise 17.4.)*

Exercise 17.4 *Suppose that a distance matrix (17.2) satisfies $c \leq \min\{a, b\}$. Sketch the cone of values of a, b, and c that satisfy the triangle inequality (17.1). (Hint: apply Exercise 17.2.)*

Exercise 17.5 *Determine the set of the (six) allowable values for a distance matrix for a four-element metric space (cf. Exercise 17.4).*

Exercise 17.6 *Does that the matrix*

$$\begin{pmatrix} 0 & 4 & 2 \\ 4 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} \tag{17.64}$$

satisfy the triangle inequality (17.1)? (Hint: see Exercise 17.2.)

Exercise 17.7 *PAM and Blossum matrices attempt to encode evolutionary distance between sequence elements. Examine these to see if they satisfy the triangle inequality (17.1).*

Exercise 17.8 *Prove that (17.15) defines a metric on stings of length n, provided that D_Σ is a metric on the alphabet Σ.*

Exercise 17.9 *Compute the Q matrix for the data in Table 17.1.*

Exercise 17.10 *Prove (17.39).*

Exercise 17.11 *Show that the matrices $P(\mathcal{D})$ and $R(\mathcal{D})$ defined in (17.42) and (17.43) satisfy $P(i,j) + R(i,j) = c$ for all $i, j = 1, \ldots, N$, where c is a constant depending only on the size N of the distance matrix $\mathcal{D}$. Determine the value of the constant c as a function of N.*

Chapter 18

Quantum models

Computational models for approximating the quantum mechanics of chemical systems [79] represent one of the most computationally intense areas of computational science. There has been recent interest in computational methods for quantum chemistry that involve algorithms with improved computational complexity [370, 391]. Recent books [68, 69] present the subject from a more mathematical point of view. Here we review of the major ideas as a guide to understanding the basis for molecular force fields.

Our objective is to clarify the notions of polarizability and induced fields. We present a novel derivation of the asymptotic behavior of the London dispersion force which is based on the induced dipole. We confirm the well known R^{-6} dependence of the force due to an induced dipole, but our techniques do not rely on the standard eigenvalue (second-order) perturbation approach [325]. Instead, our approach performs a perturbation of the wavefunction and identifies a boundary value problem for the asymptotic form of the induced dipole, which incidentally provides a way to compute the coefficient of R^{-6} for any two systems.

18.1 Why quantum models?

One might ask why we include a discussion of quantum models in a discussion of protein interactions. 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 [396, 318, 198, 318, 179, 442]. But this may require water models which are currently considered quite expensive, and these models may require further examination at the quantum level.

18.2 The Schrödinger equation

The Schrödinger equation is taken as the fundamental model for the molecular systems considered here. It is a time-dependent partial differential equation for the *wave function* ψ for the system as a function of the positions in space of the nuclei and electrons in a molecule.

For one particle of mass m moving in a force field having potential V, it can be written

$$\iota\hbar\frac{\partial\psi}{\partial t} = \mathcal{H}\psi \tag{18.1}$$

where $\hbar$ is Planck's constant divided by 2π, ι is the imaginary unit and

$$\mathcal{H}\phi(\mathbf{r}) := -\frac{\hbar^2}{2m}\nabla^2\phi(\mathbf{r}) + V(\mathbf{r})\phi(\mathbf{r}). \tag{18.2}$$

For a system of k particles, each of mass m_i, the equation (18.1) remains valid with

$$\mathcal{H}_k\phi(r_1,\ldots,r_k) := -\sum_{i=1}^{k}\frac{\hbar^2}{2m_i}\nabla_i^2\phi(r_1,\ldots,r_k) + V_k(r_1,\ldots,r_k)\phi(r_1,\ldots,r_k) \tag{18.3}$$

and ∇_i denotes the gradient with respect to r_i.

The general form of the potential V_k is

$$V_k(r_1,\ldots,r_k) := -\sum_{i=1}^{k}\sum_{j=i+1}^{k}\frac{z_i z_j}{|r_i - r_j|} \tag{18.4}$$

where z_i denotes the charge of particle i.

The single particle Hamiltonian $\mathcal{H}_1$ will appear in subsequent approximate models for multi-particle systems.

18.2.1 Particle spin

In actuality, there are more independent variables than just position and time. The *spin* σ of the particles is a discrete variable which does not effect the form of the equation (18.1) but will just enter as a parameter [169, 321]. Thus the correct functional form of the wave function is $\psi(t, r_1, \ldots, r_k, \sigma_1, \ldots, \sigma_k)$. Pauli [321] postulated the antisymmetry of the wave function:

$$\psi(t, r_1, \ldots, r_k, \sigma_1, \ldots, \sigma_k) = (-1)^{|\mu|}\psi(t, r_{\mu(1)}, \ldots, r_{\mu(k)}, \sigma_{\mu(1)}, \ldots, \sigma_{\mu(k)}) \tag{18.5}$$

for any permutation μ of the integers $\{1,\ldots,k\}$, where $|\mu|$ is the signature of the permutation (zero for even permutations, one for odd permutations).

This mathematical expression implies the Pauli exclusion principle, that there is zero probability of having two particles with the same spin and the same position. Note however, two particles of different spin can occupy the same position with nonzero probability.

18.2.2 Interpretation of ψ

The solution $\psi(t, r_1, \ldots, r_k)$ of (18.1) is called the time-dependent wave function, and $|\psi(t, r_1, \ldots, r_k)|^2$ is the probability density of finding particles at positions $r_1, \ldots, r_k$ at time t.

The electric moment may thus be defined in terms of ψ. Let m_i denote the moment

$$m_i := \int r_i |\psi(t, r_1, \ldots, r_k)|^2 \, dr_1 \cdots dr_k, \tag{18.6}$$

which is a vector in $\mathbb{R}^3$. Then the electric moment m of the entire system is

$$m := \sum_{i=1}^{k} m_i z_i = \int \sum_{i=1}^{k} z_i r_i |\psi(t, r_1, \ldots, r_k)|^2 \, dr_1 \cdots dr_k, \tag{18.7}$$

where z_i is the charge of the i-th particle.

18.3 The eigenvalue problem

The eigenvalue problem for the Schrödinger equation provides sufficient information to model many chemical interactions. It is a time-independent partial-differential eigenvalue problem of the form

$$\mathcal{H}\Psi = \lambda\Psi \tag{18.8}$$

where $\mathcal{H}$ is as in Section 18.2.

The motivation for the eigenvalue problem is that $\psi = e^{i\lambda t}\Psi$ will be a typical time-dependent solution of (18.1). This assumption requires some thought, given the extreme complexity of the system. In particular, it represents only one mode of oscillation of the system. The related equation of Madelung [416] displays the potential for more complex behavior. However, we will assume that it is sufficient to consider only the fundamental mode of oscillation of the Schrödinger equation here.

In the case of a nucleus of charge Z and a single electron (the hydrogen atom if $Z = 1$) equation (18.8) can be written

$$-\frac{\hbar^2}{2m}\nabla^2\phi(r) - \frac{Z}{|r|}\phi = \lambda\phi \tag{18.9}$$

by taking the nucleus fixed at the origin. The solutions to this equation can be determined in closed form since the angular and radial variables can be separated, that is,

$$\phi(r) = \phi(|r|\theta) = u(|r|)v(\theta) \tag{18.10}$$

where v is a spherical harmonic and u decays exponentially at infinity.

In the case of two protons and two electrons (the helium atom), equation (18.8) can be written

$$-\frac{\hbar^2}{2m}\nabla_x^2\phi - \frac{\hbar^2}{2m}\nabla_y^2\phi - Z\frac{1}{|x|}\phi - Z\frac{1}{|y|}\phi + \frac{1}{|x-y|}\phi = \lambda\phi \tag{18.11}$$

by taking the nucleus again fixed at the origin. The solutions $\phi(x, y)$ to this equation apparently can not be separated into a product $\phi_1(x)\phi_2(y)$ of functions of the the x and y variables, respectively, due to the coupling term $\frac{1}{|x-y|}$.

It is interesting to note that the operator $\mathcal{H}_k$ defined in equation (18.3) has certain symmetry properties. We will say that a function f of the k variables $r_1, \ldots, r_k$ is *symmetric* if $f(r_1, \ldots, r_k) = f(x_{\mu(1)}, \ldots, r_{\mu(k)})$ for any permutation μ of the integers $\{1, \ldots, k\}$. Then it is easy to see that $\mathcal{H}_k f$ is symmetric as long as f is symmetric. Similarly, we will say that a function f of the k variables $r_1, \ldots, r_k$ is *antisymmetric* if $f(r_1, \ldots, r_k) = -f(x_{\mu(1)}, \ldots, r_{\mu(k)})$ for any permutation μ of the integers $\{1, \ldots, k\}$. Then $\mathcal{H}_k f$ is antisymmetric as long as f is antisymmetric.

In the case of $k = 2$ (as for the helium atom), any function can be decomposed into the sum of a symmetric and an antisymmetric part.

We recall that the Pauli exclusion principle is derived from the inclusion of spin variables σ and the assertion that only antisymmetric eigenfunctions are of physical interest. In the case of electrons, the only possible spin values are $\pm\frac{1}{2}$, however for k electrons there are then 2^k possible spin combinations.

Note that the inclusion of spin in the independent variables does not change the form of the eigenproblem. More precisely, first observe that solutions to the time dependent problem (18.1) which satisfy the condition of antisymmetry (18.5) can be expressed in terms of the eigenmodes of the problem (18.8) restricted to a space of antisymmetric functions satisfying

$$\phi(r_1, \ldots, r_k, \sigma_1, \ldots, \sigma_k) = (-1)^{|\mu|}\phi(r_{\mu(1)}, \ldots, r_{\mu(k)}, \sigma_{\mu(1)}, \ldots, \sigma_{\mu(k)}) \tag{18.12}$$

for any permutation μ of the integers $\{1, \ldots, k\}$, where $|\mu|$ is the signature of the permutation (zero for even permutations, one for odd permutations).

Secondly, one can show that the eigenvalue problem in a space of antisymmetric functions of space and spin (i.e. r and σ) is equivalent to the corresponding eigenvalue problem in a space of antisymmetric functions of space alone. If

$$\mathcal{H}\psi(r, \sigma) = \lambda_\sigma \psi(r, \sigma) \tag{18.13}$$

then how do we know that the function

$$\tilde{\psi}(r) = \sum_\sigma \psi(r, \sigma) \tag{18.14}$$

(where the summation is over all spin values) satisfies the eigenproblem

$$\mathcal{H}\psi(r) = \lambda\psi(r) \tag{18.15}$$

???

On the other hand, if $\psi(r)$ is any antisymmetric eigenfunction of (18.15) then defining $\psi(r, \sigma) = \psi(r)$ for all σ yields an eigenfunctions of (18.13).

Space of antisymmetric functions can be generated by linear combinations of determinants, as observed by Slater [379, 380] and later proved by Löwdin [262]. More precisely, suppose that a given space $\mathcal{F}$ of functions of k variables can be written as a k-fold tensor product of a space F of functions of one variable. Here, let us presume that "written as" means equality as topological vector spaces (e.g., Banach spaces). This means that $f(r_1, \ldots, r_k) \in \mathcal{F}$ can be written as a limit of linear combinations of products of k functions $f_i(r_i) \in F$ of a single variable. Then the antisymmetric functions in $\mathcal{F}$ can be expressed as limits of linear combinations of determinants of matrix expressions $f_i(r_j)$.

18.4 The Born-Oppenheimer approximation

The Born-Oppenheimer approximation makes the approximation that the movement of electrons does not strongly affect the position of the nuclei. In this case, we write the force field resulting from the interactions of k electrons and n nuclei in the form

$$\begin{aligned} V_k^{\mathrm{BO}}(x_1,\dots,x_k;\xi_1,\dots,\xi_n) &:= \sum_{i=1}^{k}\sum_{j=i+1}^{k}|x_i-x_j|^{-1} \\ &-\sum_{i=1}^{k}\sum_{\nu=1}^{n}Z_\nu|x_i-\xi_\nu|^{-1}-\sum_{\mu=1}^{n}\sum_{\nu=\mu+1}^{n}Z_\mu Z_\nu|\xi_\mu-\xi_\nu|^{-1} \end{aligned} \tag{18.16}$$

where Z_ν is the charge of the ν-th nucleus, $\{x_i \,:\, i=1,\dots,k\}$ are the locations of the electrons and $\{\xi_\nu \,:\, \nu=1\dots,n\}$ are the locations of the nuclei. We assume that the charge units have been arranged so that the charge of the electrons is -1.

Of course, with the appropriate renaming of variables, the Born-Oppenheimer potential V_k^{BO} is the same as the full potential V_{n+k} defined in equation (18.4). However, we choose to think of V_k^{BO} as a function of $3k$ variables with $3n$ parameters as opposed to a function of $3(k+n)$ variables. With this in mind, we write

$$V_k^{\mathrm{BO}}(x_1,\dots,x_k;\xi_1,\dots,\xi_n) = V_k(x_1,\dots,x_k) + V_k^{\mathrm{EF}}(x_1,\dots,x_k;\xi_1,\dots,\xi_n) \tag{18.17}$$

where $V_k(x_1,\dots,x_k)=\sum_{i=1}^{k}\sum_{j=i+1}^{k}|x_i-x_j|^{-1}$ represents the direct interaction of the k electrons, and V_k^{EF} represents the remaining two terms in (18.16) which we can think of producing an external field.

The full Hamiltonian can similarly be written in the form

$$\mathcal{H}_k = -\frac{\hbar^2}{2}\sum_{i=1}^{n}\frac{1}{M_i}\Delta_{\xi_i} - \frac{\hbar^2}{2m}\Delta_{\mathbf{x}} + V_k^{\mathrm{BO}}(\mathbf{x},\xi) \tag{18.18}$$

where m is the electron mass. Here, $\Delta_{\mathbf{x}}=\sum_{i=1}^{k}\Delta_{x_k}$. We will assume that

$$M := \min M_i >> m. \tag{18.19}$$

Note that H acts on $L^2(\mathbb{R}^{(3k+3n)})$. The Born-Oppenheimer approximation is essentially a perturbation scheme in the small parameter m/M, but as is clear from (18.18), this is a very singular perturbation.

The Born-Oppenheimer approximation consists of first studying the spectral problem for the "electronic Hamiltonian,"

$$\mathcal{H}_k^{\mathrm{BO}}(\xi) = -\frac{\hbar^2}{2m}\Delta_{\mathbf{x}} + V_k^{\mathrm{BO}}(\mathbf{x},\xi) \quad , \tag{18.20}$$

acting on $L^2(\mathbb{R}^{3k})$ (here $\mathbf{x}\in\mathbb{R}^{3k}$), and treating the nuclear positions as parameters. Thus, we will obtain a sequence of eigenvalues $\lambda_1(\xi)<\lambda_2(\xi)\le\lambda_3(\xi)\le\dots$, for each choice of the nuclear

positions $\underset{\sim}{\xi}$. Born and Oppenheimer then argued that the eigenvalues of the full problem are given (approximately) by the sequence of simplified "nuclear Hamiltonians"

$$\mathcal{H}_j^{\rm N} = -\frac{\hbar^2}{2M}\Delta_{\underset{\sim}{\xi}} + \lambda_j(\underset{\sim}{\xi}) \quad , \tag{18.21}$$

which act on $L^2(\mathbb{R}^{3n})$ (here $\underset{\sim}{\xi} \in \mathbb{R}^{3n}$).

Some insight into the relation between the spectrum of (18.20) and (18.21) comes from remarking that if $u_1(\cdot;\underset{\sim}{\xi})$ is the eigenfunction of (18.20), with eigenvalue $\lambda_1(\underset{\sim}{\xi})$, and if $\psi(\cdot)$ is an eigenfunction of the nuclear Hamiltonian H_1^N, then $\psi(\underset{\sim}{\xi})u_1(\mathbf{x},\underset{\sim}{\xi})$ is an eigenfunction of the full Hamiltonian (18.18). There is, however, no reason for all of the eigenfunctions of the full problem to factor in this way, and a better way of understanding the relationship between H^{BO} and H_j^N is through a reduction method first introduced by Feshbach in the context of nuclear physics [147]. Suppose that we expand an arbitrary function $\psi(\mathbf{x},\underset{\sim}{\xi}) = \sum_{n=0}^{\infty}\phi_n(\mathbf{x},\underset{\sim}{\xi})\psi_n(\underset{\sim}{\xi})$, where for each $\underset{\sim}{\xi}$, we choose a complete, orthonormal set of eigenfunctions $\{\phi_n(\cdot,\underset{\sim}{\xi})\}$ for $H^{BO}(\underset{\sim}{\xi})$. (One can choose these eigenfunctions to be at least C^2 with respect to $\underset{\sim}{\xi}$ [277].) If one then substitutes this expansion into the "full" eigenvalue problem:

$$H\psi(\mathbf{x},\underset{\sim}{\xi}) = \lambda\psi(\mathbf{x},\underset{\sim}{\xi}) \tag{18.22}$$

and then projects onto the eigenfunctions ϕ_j, one obtains the coupled system of eigenvalue problems

$$H_j^N\psi_j + \sum_k \frac{\hbar^2}{2M}\langle\phi_j, \Delta_{\underset{\sim}{\xi}}\phi_k\rangle_x\psi_k = \lambda\psi_j \quad j = 0,1,2,3\ldots \quad . \tag{18.23}$$

Here, $\langle\cdot,\cdot\rangle_x$ means that we take the inner product only with respect to the electron coordinates.

Feshbach then notes that if one defines projection operators P_0, which projects onto the ψ_0 direction, and $\hat{P}_0$, the orthogonal complement of P_0, then one can rewrite the $j=0$ equation in (18.23) as an uncoupled equation for ψ_0. If one applies P_0 and $\hat{P}_0$ to (18.22), one obtains:

$$P_0HP_0\psi + P_0H\hat{P}_0\psi = \lambda P_0\psi \tag{18.24}$$

$$\hat{P}_0\psi = (\hat{P}_0(J-\lambda)\hat{P}_0)^{-1}\hat{P}_0H\hat{P}\psi \quad , \tag{18.25}$$

where $(\hat{P}_0(J-\lambda)\hat{P}_0)^{-1}$ exists provided there is a gap between the lowest eigenvalue of H^{BO} and the remainder of its spectrum – *i.e.* provided $\delta = \inf_{\underset{\sim}{\xi}}\mathrm{dist}(\lambda_0(\underset{\sim}{\xi}), \mathrm{spec}(H^{BO}(\underset{\sim}{\xi})\backslash\lambda_0(\underset{\sim}{\xi})) > 0$. Inserting the first of these equations into the second gives

$$P_0HP_0 + P_0H(\hat{P}_0(J-\lambda)\hat{P}_0)^{-1}\hat{P}_0HP_0\psi = \lambda P_0\psi \tag{18.26}$$

Recalling that $H = -\frac{\hbar^2}{2M}\Delta_{\underset{\sim}{\xi}} + H^{BO}(\underset{\sim}{\xi})$, and that P_0 and $\hat{P}_0$ are spectral projections for H^{BO}, one finds that $\hat{P}_0HP_0 = -\frac{\hbar^2}{2M}\hat{P}_0\Delta_{\underset{\sim}{\xi}}P_0$, while $P_0HP_0 = -\frac{\hbar^2}{2M}P_0\Delta_{\underset{\sim}{\xi}}P_0 + \lambda_0(\underset{\sim}{\xi})P_0$. Thus, since $\psi_0 = P_0\psi$, (18.24) is equivalent to

$$\begin{aligned}\left(-\frac{\hbar^2}{2M}\Delta_{\underset{\sim}{\xi}} + \lambda_0(\underset{\sim}{\xi})\right)\psi_0 + &\frac{\hbar^2}{2M}\langle\phi_0, \Delta_{\underset{\sim}{\xi}}\phi_0\rangle_x\psi_0 \\ &+ \left(\frac{\hbar^2}{2M}\right)^2\hat{P}_0\Delta_{\underset{\sim}{\xi}}P_0\left(\hat{P}_0(J-\lambda)\hat{P}_0\right)^{-1}\hat{P}_0\Delta_{\underset{\sim}{\xi}}\psi_0 = \lambda\psi_0 \quad .\end{aligned} \tag{18.27}$$

Thus, up to the correction terms, $\frac{\hbar^2}{2M}\langle\phi_0, \Delta_{\underset{\sim}{\xi}}\phi_0\rangle_x$ and $(\frac{\hbar^2}{2M})^2\hat{P}_0\Delta_{\underset{\sim}{\xi}}P_0(\hat{P}_0(J-\lambda)\hat{P}_0)^{-1}\hat{P}_0\Delta_{\underset{\sim}{\xi}}$, the spectral problem for (18.22) is equivalent to that for H_0^N, at least when $\lambda \approx \lambda_0$.

Note further that the perturbations in (18.27) are no longer singular perturbations, since both of the perturbations are relatively bounded with respect to H_0^N. Indeed, while it is technically rather involved, the above approach can be made rigorous, [177], [178], [222], [277] and one finds that there exists an asymptotic expansion of the eigenvalues of (18.22) in powers of $(\frac{m}{M})^{1/2}$ with the leading order in this expansion given by the eigenvalues of H_0^N.

From now on, we will assume that the Born-Oppenheimer approximation is in force.

18.5 External fields

Now let us consider the effect of an external field on solutions of the eigenvalue problem for the Schrödinger equation. This takes the form

$$\mathcal{H}\Psi + \epsilon B\Psi = \lambda\Psi \tag{18.28}$$

where $\mathcal{H}$ is as in Section 18.2 and B represents the external field. We introduce a small parameter ϵ to facilitate discussion.

More precisely, we have $\mathcal{H} := \mathcal{H}_k + V_k^{\mathrm{BO}}$ and $B = \epsilon^{-1}V_k^{\mathrm{EF}}$ where

$$V_k^{\mathrm{EF}}(x_1, \ldots, x_k; \eta_1, \ldots, \eta_n) = -\sum_{i=1}^{k}\sum_{\nu=1}^{n} z_i Z_\nu |x_i - \eta_\nu|^{-1} - \sum_{\mu=1}^{n}\sum_{\nu=\mu+1}^{n} Z_\mu Z_\nu |\eta_\mu - \eta_\nu|^{-1} \tag{18.29}$$

where the charges Z_ν produce the external field. The second term is independent of x_i, so we can just treat it as a constant, C_Z. Suppose that the two systems are separated by a distance $R = \min\{|x_i - \eta_\nu|\}$, so that we can write $\eta_\nu = \eta'_\nu - R\mathbf{E}$ for some unit vector $\mathbf{E}$ and expand the first term in powers of R:

$$V_k^{\mathrm{EF}}(x_i, \eta_\nu) - C_Z = -\sum_{i=1}^{k}\sum_{\nu=1}^{n}\frac{z_i Z_\nu}{|x_i - \eta_\nu|} = -\frac{1}{R}\sum_{i=1}^{k}\sum_{j=1}^{k}\frac{z_i Z_\nu}{|E + R^{-1}(x_i - \eta_\nu)|}. \tag{18.30}$$

We then have the approximation

$$V_k^{\mathrm{EF}}(x_i, \eta_i) = C_Z + R^{-1}\left(A_0 + A_1 R^{-1} + A_2 R^{-2} + \cdots\right) \tag{18.31}$$

where

$$A_0 = -\sum_{i=1}^{k}\sum_{\nu=1}^{n} z_i Z_\nu \tag{18.32}$$

and

$$A_1 = \sum_{i=1}^{k}\sum_{\nu=1}^{n} z_i Z_\nu \mathbf{E}\cdot(x_i - \eta_\nu). \tag{18.33}$$

If we assume that $\sum_{j=1}^{k} z_j = 0$, then (18.32) simplifies to $A_0 = 0$. Similarly, A_1 becomes

$$A_1 = \sum_{i=1}^{k} \sum_{\nu=1}^{n} z_i Z_\nu \mathbf{E} \cdot x_i = \left(\sum_{\nu=1}^{n} Z_\nu \right) \mathbf{E} \cdot \left(\sum_{i=1}^{k} z_i x_i \right). \tag{18.34}$$

Thus we can write

$$V_k^{\mathrm{EF}}(r_i^1, r_i^2) \approx C_Z + \mathcal{E} \cdot M(x_i) \tag{18.35}$$

where $\mathcal{E}$ denotes the external electric field

$$\mathcal{E} = \frac{1}{R^2} \left(\sum_{\nu=1}^{n} Z_\nu \right) \mathbf{E}, \tag{18.36}$$

and

$$M(x_i) := \sum_{i=1}^{k} z_i x_i \tag{18.37}$$

is the electric moment of the system with charges z_i at positions x_i. With these approximations, our eigenvalue problem becomes

$$\mathcal{H}_k \Psi + \mathcal{E} \cdot M(x_i) \Psi = (\lambda - C_Z) \Psi \tag{18.38}$$

which is of the form (18.28) with $\epsilon B = \mathcal{E} \cdot M(x_i)$, and with the eigenvalue shifted by a constant.

An eigenvalue problem is a nonlinear system, so a perturbation is not simple to analyze, except in the case where the perturbation is small. Let us call the solution to (18.38) Ψ^ϵ. Hence Ψ^0 is the solution to the unperturbed eigenvalue problem (18.8). Let us suppose that we can write $\Psi^\epsilon \approx \Psi^0 + \epsilon \Psi'$ for some Ψ' at least for small ϵ.

In general [298], the dependence of λ^ϵ on ϵ can be quite singular. However, let us assume that $\lambda^\epsilon \approx \lambda^0 + \epsilon \lambda'$ for small ϵ. Then taking difference quotients and letting $\epsilon \to 0$, we find

$$(\mathcal{H} - \lambda^0 \mathcal{I}) \Psi' = -B\Psi^0 + \lambda' \Psi^0 \tag{18.39}$$

Since $\|\Psi^\epsilon\| = 1$ for all ϵ, we must have $\Psi' \in [\Psi^0]^\perp$, where $[\Psi^0]$ denotes the linear space spanned by Ψ^0 (Exercise 18.3). If λ^0 is a simple eigenvalue, then this implies that (18.39) is solvable provided $\lambda' = (\Psi^0, B\Psi^0)$, since this implies that $B\Psi^0 + \lambda' \Psi^0 \in [\Psi^0]^\perp$.

18.5.1 Polarization

Polarization of a chemical system is simply the result of the action of an external field. The **polarizability** α of a material can be defined as the infinitesimal response to a small external field. More precisely, α is the matrix that gives the change in the electric moment (18.7) as a response to the input external field.

The electric moment for a system where z_i is the charge of the i-th particle is defined in terms of Ψ as in (18.7) via

$$m := \sum_{i=1}^{k} z_i \int r_i |\Psi(r_1, \ldots, r_k)|^2 \, dr_1 \cdots dr_k. \tag{18.40}$$

If $\mathcal{E} = \epsilon \mathbf{e}^j$ is our external field, then the resulting infinitesimal change in i-th moment is given by

$$\begin{aligned}\Delta m_i^j :=& \lim_{\epsilon \to 0} \frac{1}{\epsilon} \int r_i \big(|\Psi^{\mathcal{E}}(r_1, \ldots, r_k)|^2 - |\Psi^0(r_1, \ldots, r_k)|^2 \big)\, dr_1 \cdots dr_k \\ =& 2 \int r_i \Psi^0(r_1, \ldots, r_k) \Psi_j'(r_1, \ldots, r_k)\, dr_1 \cdots dr_k,\end{aligned} \tag{18.41}$$

where $\Psi_j' \in [\Psi^0]^\perp$ solves the linear system

$$(\mathcal{H} - \lambda^0 \mathcal{I}) \Psi_j' = -e^j \cdot M(r) \Psi^0 + \lambda_j' \Psi^0, \tag{18.42}$$

with

$$\begin{aligned}\lambda_j' =& (\Psi^0, e^j \cdot M(r) \Psi^0) \\ =& \int \sum_{i=1}^k z_i r_i^j |\Psi^0(r_1, \ldots, r_k)|^2\, dr_1 \cdots dr_k.\end{aligned} \tag{18.43}$$

Thus α is the matrix

$$\alpha_{j,\ell} = 2 \int \sum_{i=1}^k z_i r_i^\ell \Psi^0(r_1, \ldots, r_k) \Psi_j'(r_1, \ldots, r_k)\, dr_1 \cdots dr_k. \tag{18.44}$$

Note that λ_j' is the j-th component of the electric moment of the unperturbed system defined in (18.40).

18.5.2 Induced fields

Induced dipoles represent an intriguing example of external fields. Each dipole provides the external field that induces the other dipole. This sounds like a circular argument, and so it is deserving of a more detailed examination, which we now present. When two parts of an atomic system are far apart from each other, it is reasonable to assume that the coupling should be quite weak. In particular, let us suppose that the two sub-systems are identical, for simplicity.

Let us return to the basic notation of (18.3), but now we refer to the variables of one system by r_i^1 and the other by r_i^2. The full system operator is thus

$$\begin{aligned}\mathcal{H}_k \phi(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2) :=& -\sum_{i=1}^k \frac{\hbar^2}{2m_i} \nabla^2_{r_i^1} \phi - \sum_{i=1}^k \frac{\hbar^2}{2m_i} \nabla^2_{r_i^2} \phi \\ &+ V_k(r_1^1, \ldots, r_k^1)\phi + V_k(r_1^2, \ldots, r_k^2)\phi + V_k^{\mathrm{EF}}(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2)\phi \\ =& \left(A_{r^1} + A_{r^2} + V_k^{\mathrm{EF}}(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2) \right) \phi\end{aligned} \tag{18.45}$$

where $\phi = \phi(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2)$ and we have separated the potential terms as in (18.17). Suppose that the two systems are separated by a distance R along an axis that is parallel to a unit vector E, so that by renaming variables by

$$r_j^2 \to -r_j^2 + R\mathbf{E} \tag{18.46}$$

we can expand $V_k^{\rm EF}$ as

$$\begin{aligned} V_k^{\rm EF}(r_1^1,\ldots,r_k^1,r_1^2,\ldots,r_k^2) &= -\sum_{i=1}^k\sum_{j=1}^k \frac{z_i z_j}{|r_i^1 - R\mathbf{E} + r_j^2|} \\ &= -\frac{1}{R}\sum_{i=1}^k\sum_{j=1}^k \frac{z_i z_j}{|E - R^{-1}(r_i^1 + r_j^2)|}. \end{aligned} \tag{18.47}$$

We can make the asymptotic expansion

$$V_k^{\rm EF}(r_1^1,\ldots,r_k^1,r_1^2,\ldots,r_k^2) = \frac{a_0}{R} + \frac{a_1}{R^2} + \frac{a_2}{R^3} + \cdots \tag{18.48}$$

where we can easily verify that

$$a_0 = -\sum_{i=1}^k\sum_{j=1}^k z_i z_j = -\sum_{i=1}^k z_i\left(\sum_{j=1}^k z_j\right). \tag{18.49}$$

If the net charge on each molecule is zero, then

$$\sum_{j=1}^k z_j = 0, \tag{18.50}$$

so (18.48) simplifies since $a_0 = 0$. We will show subsequently that

$$a_1 = \sum_{i=1}^k\sum_{j=1}^k z_i z_j \mathbf{E}\cdot(r_i^1 + r_j^2). \tag{18.51}$$

Expanding we find that

$$\begin{aligned} a_1 =& \sum_{i=1}^k\sum_{j=1}^k z_i z_j \mathbf{E}\cdot r_i^1 + \sum_{i=1}^k\sum_{j=1}^k z_i z_j \mathbf{E}\cdot r_j^2 \\ =& \left(\sum_{j=1}^k z_j\right)\sum_{i=1}^k z_i \mathbf{E}\cdot r_i^1 + \left(\sum_{i=1}^k z_i\right)\sum_{j=1}^k z_j \mathbf{E}\cdot r_j^2 = 0, \end{aligned} \tag{18.52}$$

provided (18.50) holds.

To clarify the computations, let us write $\epsilon = 1/R$ and then

$$V_k^{\rm EF}(r_1^1,\ldots,r_k^1,r_1^2,\ldots,r_k^2) = -\epsilon S(\epsilon), \tag{18.53}$$

where the expression S is given by

$$S(\epsilon) = \sum_{i=1}^k\sum_{j=1}^k \frac{z_i z_j}{|\mathbf{E} - \epsilon(r_i^1 + r_j^2)|}. \tag{18.54}$$

We have $S(0) = a_0$, and computing the derivative of S with respect to ϵ yields a_1. Let S_{ij} denote an individual term in the double summation (18.54):

$$S_{ij}(\epsilon) = \frac{z_i z_j}{|\mathbf{E} - \epsilon r|} = z_i z_j \left(|\mathbf{E}|^2 - 2\epsilon \mathbf{E}\cdot r + \epsilon^2 |r|^2\right)^{-1/2}, \tag{18.55}$$

where $r = r_i^1 + r_j^2$. Therefore

$$S'_{ij}(\epsilon) = z_i z_j \left(\mathbf{E}\cdot r - \epsilon |r|^2\right)\left(|\mathbf{E}|^2 - 2\epsilon \mathbf{E}\cdot r + \epsilon^2 |r|^2\right)^{-3/2}, \tag{18.56}$$

so that setting $\epsilon = 0$ confirms the computation of $a_1 = S'(0)$, since $|\mathbf{E}| = 1$. Differentiating (18.56), we find

$$\begin{aligned} S''_{ij}(\epsilon) =& 3 z_i z_j \left(\mathbf{E}\cdot r - \epsilon |r|^2\right)^2 \left(|\mathbf{E}|^2 - 2\epsilon \mathbf{E}\cdot r + \epsilon^2 |r|^2\right)^{-5/2} \\ & - z_i z_j |r|^2 \left(|\mathbf{E}|^2 - 2\epsilon \mathbf{E}\cdot r + \epsilon^2 |r|^2\right)^{-3/2}, \end{aligned} \tag{18.57}$$

and thus $S''_{ij}(0) = 3 z_i z_j \left(\mathbf{E}\cdot r\right)^2 - z_i z_j |r|^2$. Therefore

$$\begin{aligned} S''(0) &= \sum_{i=1}^k \sum_{j=1}^k z_i z_j \left(3\left(\mathbf{E}\cdot (r_i^1 + r_j^2)\right)^2 - |r_i^1 + r_j^2|^2\right) \\ &= \sum_{i=1}^k \sum_{j=1}^k z_i z_j \left(3\left((\mathbf{E}\cdot r_i^1)^2 + 2(\mathbf{E}\cdot r_i^1)(\mathbf{E}\cdot r_j^2) + (\mathbf{E}\cdot r_j^2)^2\right) - \left(|r_i^1|^2 + 2 r_i^1 \cdot r_j^2 + |r_j^2|^2\right)\right) \\ &= \sum_{i=1}^k \sum_{j=1}^k z_i z_j \left(6(\mathbf{E}\cdot r_i^1)(\mathbf{E}\cdot r_j^2) - 2\left(r_i^1 \cdot r_j^2\right)\right), \end{aligned} \tag{18.58}$$

provided (18.50) holds. Define the electric moment

$$D(r_1^\ell, \ldots, r_k^\ell) = \sum_{i=1}^k z_i r_i^\ell, \ \ell = 1, 2. \tag{18.59}$$

Note that $a_2 = \frac{1}{2} S''(0)$. Then if (18.50) holds, (18.58) implies

$$\begin{aligned} a_2 =& \sum_{i=1}^k \sum_{j=1}^k z_i z_j \left(3\left(\mathbf{E}\cdot r_i^1\right)\left(\mathbf{E}\cdot r_j^2\right) - \left(r_i^1 \cdot r_j^2\right)\right) \\ =& 3\left(\left(\mathbf{E}\cdot D(\mathbf{r}^1)\right)\left(\mathbf{E}\cdot D(\mathbf{r}^2)\right) - D(\mathbf{r}^1)\cdot D(\mathbf{r}^2)\right). \end{aligned} \tag{18.60}$$

where we set $\mathbf{r}^\ell = (r_1^\ell, \ldots, r_k^\ell)$. From (18.48), we conclude that

$$V_k^{\rm EF}(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2) \approx \frac{a_2}{R^3} = \frac{1}{R^3}\left(3(\mathbf{E}\cdot D(\mathbf{r}^1))(\mathbf{E}\cdot D(\mathbf{r}^2)) - D(\mathbf{r}^1)\cdot D(\mathbf{r}^2)\right). \tag{18.61}$$

Now the eigenvalue problem can be written

$$Af_\epsilon + \epsilon B f_\epsilon = \lambda_\epsilon f_\epsilon, \tag{18.62}$$

where $A = A_{\mathbf{r}^1} + A_{\mathbf{r}^2}$, $\epsilon = R^{-3}$ and B is the multiplication operator defined by

$$B(\mathbf{r}^1, \mathbf{r}^2) = B(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2) = 3(\mathbf{E}\cdot D(\mathbf{r}^1))(\mathbf{E}\cdot D(\mathbf{r}^2)) - D(\mathbf{r}^1)\cdot D(\mathbf{r}^2). \tag{18.63}$$

Suppose that $f_\epsilon = f_0 + \epsilon f' + \mathcal{O}(\epsilon^2)$ and that $f_0(\mathbf{r}^1, \mathbf{r}^2) = \hat{f}_0(\mathbf{r}^1)\hat{f}_0(\mathbf{r}^2)$ with the property that $\hat{f}_0$ satisfies

$$\int D(\mathbf{r})|\hat{f}_0(r_1, \ldots, r_k)|^2\, dr_1 \cdots dr_k = 0. \tag{18.64}$$

Here, the integrand is a vector, so we are saying the integral of all three components is zero. If we take the dot-product with $\mathbf{E}$, we also find

$$\int \mathbf{E}\cdot D(\mathbf{r})|\hat{f}_0(r_1, \ldots, r_k)|^2\, dr_1 \cdots dr_k = 0. \tag{18.65}$$

Therefore (18.64) implies that

$$\begin{aligned}
&\int B(\mathbf{r}^1, \mathbf{r}^2)|f_0(\mathbf{r}^1, \mathbf{r}^2)|^2\, d\mathbf{r}^1 d\mathbf{r}^2 = \\
&\int \left(3(\mathbf{E}\cdot D(\mathbf{r}^1))(\mathbf{E}\cdot D(\mathbf{r}^2)) - D(\mathbf{r}^1)\cdot D(\mathbf{r}^2)\right)|\hat{f}_0(\mathbf{r}^1)\hat{f}_0(\mathbf{r}^2)|^2\, d\mathbf{r}^1 d\mathbf{r}^2 = \\
&3\int (\mathbf{E}\cdot D(\mathbf{r}^1))|\hat{f}_0(\mathbf{r}^1)|^2\, d\mathbf{r}^1 \int (\mathbf{E}\cdot D(\mathbf{r}^2))|\hat{f}_0(\mathbf{r}^2)|^2\, d\mathbf{r}^2 \\
&- \int D(\mathbf{r}^1)|\hat{f}_0(\mathbf{r}^1)|^2\, d\mathbf{r}^1 \cdot \int D(\mathbf{r}^2)|\hat{f}_0(\mathbf{r}^2)|^2\, d\mathbf{r}^2 = 0.
\end{aligned} \tag{18.66}$$

Note that f' satisfies $f' \perp f_0$ and

$$\begin{aligned}
-\sum_{i=1}^k \frac{\hbar^2}{2m_i}\nabla^2_{r_i^1} f'(\mathbf{r}^1, \mathbf{r}^2) - \sum_{i=1}^k \frac{\hbar^2}{2m_i}\nabla^2_{r_i^2} f'(\mathbf{r}^1, \mathbf{r}^2) + \left(V_k(\mathbf{r}^1) + V_k(\mathbf{r}^2) - \lambda_0\right) f'(\mathbf{r}^1, \mathbf{r}^2) \\
= - B(\mathbf{r}^1, \mathbf{r}^2)\hat{f}_0(\mathbf{r}^1)\hat{f}_0(\mathbf{r}^2)
\end{aligned} \tag{18.67}$$

since f_0 is orthogonal to Bf_0, by (18.66). Even though f_0 factors into a product, we cannot solve (18.67) by separation of variables as a product. The coupling of the right-hand side term $D(\mathbf{r}^1)\cdot D(\mathbf{r}^2)$ prevents this. Thus the dipole $\widetilde{D}_\epsilon$ induced is

$$\begin{aligned}
\widetilde{D}_\epsilon &= \int D(\mathbf{r}^1)|f_\epsilon(\mathbf{r}^1, \mathbf{r}^2)|^2\, d\mathbf{r}^1 d\mathbf{r}^2 \\
&\approx \int D(\mathbf{r}^1)|\left(f_0 + \epsilon f'\right)(\mathbf{r}^1, \mathbf{r}^2)|^2\, d\mathbf{r}^1 d\mathbf{r}^2 + \mathcal{O}(\epsilon^2) \\
&= 2\epsilon \int D(\mathbf{r}^1)\left(f_0 f'\right)(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2 + \mathcal{O}(\epsilon^2) \\
&= \frac{2}{R^3}\int D(\mathbf{r}^1)\left(f_0 f'\right)(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2 + \mathcal{O}(R^{-6}).
\end{aligned} \tag{18.68}$$

Let us give a name to the dipolar coefficient of R^{-3} in (18.68):

$$\mathcal{D} = \int D(\mathbf{r}^1)\,(f_0 f')\,(\mathbf{r}^1, \mathbf{r}^2)\,d\mathbf{r}^1 d\mathbf{r}^2. \tag{18.69}$$

Then we can state our main result as follows: the induced dipole $\widetilde{D}_\epsilon$ is

$$\widetilde{D}_\epsilon = \int D(\mathbf{r}^1)|f_\epsilon(\mathbf{r}^1, \mathbf{r}^2)|^2\,d\mathbf{r}^1 d\mathbf{r}^2 = \frac{2}{R^3}\mathcal{D} + \mathcal{O}(R^{-6}), \tag{18.70}$$

where $\mathcal{D}$ is defined in (18.69).

It is noteworthy that the term $\frac{2}{R^3}$ is significant for $R = 4$, a distance at which van der Waals forces are often a dominant effect. The error term $\epsilon^2 = 2^{-18}$ is quite small for $R = 4$, so we can trust the accuracy of the prediction (18.70) in this range.

In the approach reported in [325], the perturbation analysis is based on the beavhior of λ^ϵ instead of the wave function f_ϵ as we have done here. The deviation $\lambda^\epsilon - \lambda^0$ is a direct measure of the energy of the induced dipole. As we have determined, $\lambda' = 0$, so it would appear that

$$\lambda^\epsilon - \lambda^0 = \mathcal{O}(\epsilon^2) = \mathcal{O}(R^{-6}) \tag{18.71}$$

but we have not identified an experession for the coefficient λ'' required to quantify this asymptotic estimate. The degeneracy caused by having $\lambda' = 0$ necessitates a so-called second-order perturbation method in [325], and only an approximate version of that is utilized.

By emphasizing the calculation of the wave function, we have derived an expression that is not degenerate in ϵ and indeed gives the expected behavior of the induced dipole strength as a function of R. Of course, evaluating numerically the coefficient of R^{-3} presented in (18.68) requires some work, but conceptually it is clear how to do this by solving the Schrödinger equation together with a closely related auxiliary equation.

18.5.3 Hartree approximation

The Hartree approximation can sometimes provide interesting information. It uses the approximation

$$\phi(r_1^1, \ldots, r_k^1, r_1^2, \ldots, r_k^2) \approx f^1(r_1^1, \ldots, r_k^1) f^2(r_1^2, \ldots, r_k^2). \tag{18.72}$$

We will see that this leads to an interesting interpretation of the induced dipole, with similar asymptotic behvior.

Let us write $\mathcal{H}_k = \mathcal{H}_k^1 + \mathcal{H}_k^2 + V_k^{\mathrm{EF}}$. Then

$$\begin{aligned}
\mathcal{H}_k\phi \approx & f^2\mathcal{H}_k^1 f^1 + f^1\mathcal{H}_k^2 f^2 + V_k^{\mathrm{EF}} f^1 f^2 \\
= & f^2(\mathbf{r}^2)\left(\mathcal{H}_k^1 f^1 + \tfrac{1}{2}V_k^{\mathrm{EF}}(\cdot, \mathbf{r}^2) f^1\right) + f^1(\mathbf{r}^1)\left(\mathcal{H}_k^2 f^2 + \tfrac{1}{2}V_k^{\mathrm{EF}}(\mathbf{r}^1, \cdot) f^2\right) \\
\approx & f^2(\mathbf{r}^2)\left(\mathcal{H}_k^1 f^1 + \tfrac{1}{2}R^{-3}B(\cdot, \mathbf{r}^2) f^1\right) + f^1(\mathbf{r}^1)\left(\mathcal{H}_k^2 f^2 + \tfrac{1}{2}R^{-3}B(\mathbf{r}^1, \cdot) f^2\right)
\end{aligned} \tag{18.73}$$

where we used the approximation (18.61). We further approximate by replacing the dipoles in the definition of B by ones given by their expected positions. That is, let $\hat{r}_i^1$ denote the expected position of the i-th atom,

$$\hat{r}_i^\ell := \int r_i^\ell |f^\ell(r_1^\ell, \ldots, r_k^\ell)|^2 \, dr_1^\ell \cdots dr_k^\ell, \tag{18.74}$$

and define

$$\widehat{D}(r^\ell) = \widehat{D}^\ell = \sum_{i=1}^k z_i \hat{r}_i^\ell \tag{18.75}$$

for $\ell = 1, 2$. Thus

$$\widehat{D}^\ell := \sum_{i=1}^k z_i \int r_i^\ell |f^\ell(r_1^\ell, \ldots, r_k^\ell)|^2 \, dr_1^\ell \cdots dr_k^\ell. \tag{18.76}$$

We can thus write the expression $B(\mathbf{r}^1, \mathbf{r}^2)$ as either

$$B(\mathbf{r}^1, \mathbf{r}^2) \approx 3(\mathbf{E} \cdot \widehat{D}^1)(\mathbf{E} \cdot D(\mathbf{r}^2)) - \widehat{D}^1 \cdot D(\mathbf{r}^2) \tag{18.77}$$

or

$$B(\mathbf{r}^1, \mathbf{r}^2) \approx 3(\mathbf{E} \cdot \widehat{D}^2)(\mathbf{E} \cdot D(\mathbf{r}^1)) - \widehat{D}^2 \cdot D(\mathbf{r}^1), \tag{18.78}$$

depending on the context. Then (18.73) becomes

$$\begin{aligned} \mathcal{H}_k \phi \approx f^2(\mathbf{r}^2) \Big(\mathcal{H}_k^1 f^1 + \tfrac{1}{2} R^{-3} \left(3(\mathbf{E} \cdot \widehat{D}^2)(\mathbf{E} \cdot D(\mathbf{r}^1)) - \widehat{D}^2 \cdot D(r^1) f^1 \right) \Big) \\ + f^1(\mathbf{r}^1) \Big(\mathcal{H}_k^2 f^2 + \tfrac{1}{2} R^{-3} \left(3(\mathbf{E} \cdot \widehat{D}^1)(\mathbf{E} \cdot D(\mathbf{r}^2)) - \widehat{D}^1 \cdot D(r^2) f^2 \right) \Big). \end{aligned} \tag{18.79}$$

Let us make the ansatz that $f^1 = f^2$. This reduces the eigenvalue problem to one for a single distribution $f = f^1 = f^2$. Then the eigenvalue problem $\mathcal{H}_k \phi = \lambda \phi$ would collapse into a single eigenvalue problem

$$\mathcal{H}_k^\rho f + \tfrac{1}{2} R^{-3} \left(3(\mathbf{E} \cdot \widehat{D})(\mathbf{E} \cdot D(\rho)) - \widehat{D} \cdot D(\rho) \right) f = \lambda f, \tag{18.80}$$

where ρ denotes the spatial variable name, and the induced dipole is

$$\widehat{D} := \sum_{i=1}^k z_i \int \rho_i |f(\rho_1, \ldots, \rho_k)|^2 \, d\rho_1 \cdots d\rho_k. \tag{18.81}$$

Then (18.80) appears to be a single eigenvalue problem that is parameterized by the coefficient $R^{-3}\widehat{D}$, but the problem is that $\widehat{D}$ also depends on f as well. Now we proceed to remove this recursive dependence on f.

Suppose that we know by symmetry that $\widehat{D} = d\mathbf{E}$ where $\mathbf{E}$ is the unit vector used in the change of variables (18.46). Then we have

$$d = \mathbf{E} \cdot \widehat{D} = \sum_{i=1}^k z_i \int \mathbf{E} \cdot \rho_i |f(\rho_1, \ldots, \rho_k)|^2 \, d\rho_1 \cdots d\rho_k, \tag{18.82}$$

and the term multiplying $\frac{1}{2}R^{-3}$ in (18.80) can be simplified as

$$3(\mathbf{E}\cdot\widehat{D})(\mathbf{E}\cdot D(\rho)) - \widehat{D}\cdot D(\rho) = 3d(\mathbf{E}\cdot D(\rho)) - d\mathbf{E}\cdot D(\rho) = 2d(\mathbf{E}\cdot D(\rho)). \tag{18.83}$$

Therefore the eigenvalue problem (18.80) can be written

$$\mathcal{H}_k^\rho f_\epsilon + \epsilon B f_\epsilon = \lambda_\epsilon f_\epsilon, \tag{18.84}$$

where we make the definitions

$$\epsilon = dR^{-3} \tag{18.85}$$

and

$$B = \mathbf{E}\cdot D(\rho). \tag{18.86}$$

Written in this way, we have eliminated the recursive dependence of f_ϵ on f_ϵ, but at the expense of losing R as an independent variable. But we can recover R as needed from (18.85). Thus we have $R = (d/\epsilon)^{1/3}$ so that (18.82) implies

$$R = R_\epsilon = \left(\frac{1}{\epsilon}\sum_{i=1}^k z_i \int E\cdot\rho_i |f_\epsilon(\rho_1,\dots,\rho_k)|^2\, d\rho_1\cdots d\rho_k\right)^{1/3}. \tag{18.87}$$

Suppose that $f_\epsilon = f_0 + \epsilon f' + \mathcal{O}(\epsilon^2)$ and that

$$\sum_{i=1}^k z_i \int E\cdot\rho_i |f_0(\rho_1,\dots,\rho_k)|^2\, d\rho_1\cdots d\rho_k = 0. \tag{18.88}$$

Then

$$R_\epsilon \approx \left(2\left(\sum_{i=1}^k z_i \int E\cdot\rho_i f_0(\rho_1,\dots,\rho_k) f'(\rho_1,\dots,\rho_k)\, d\rho_1\cdots d\rho_k\right) + \mathcal{O}(\epsilon)\right)^{1/3}. \tag{18.89}$$

Note that f' satisfies $f' \perp f_0$ and

$$(\mathcal{H}_k^\rho - \lambda_0 I) f' = -Bf_0 = -\mathbf{E}\cdot D(\rho) f_0 \tag{18.90}$$

since f_0 is orthogonal to Bf_0. The system (18.90) can be viewed as a Hartree approximation to (18.67).

Recalling (18.87), we find

$$R_\epsilon \to R_0 := \left(2\left(\sum_{i=1}^k z_i \int \mathbf{E}\cdot\rho_i f_0(\rho_1,\dots,\rho_k) f'(\rho_1,\dots,\rho_k)\, d\rho_1\cdots d\rho_k\right)\right)^{1/3}. \tag{18.91}$$

Note the similarlity between the expression for R_0 and the one for $\mathcal{D}$ defined in (18.69); the expression (18.91) can be viewed as a Hartree approximation to (18.69). More precisely, define

$$\mathcal{D}^H = \int D(\rho) f_0(\rho_1,\dots,\rho_k) f'(\rho_1,\dots,\rho_k)\, d\rho_1\cdots d\rho_k. \tag{18.92}$$

Then we have

$$\begin{aligned} |\widehat{D}| =& d = \epsilon R_\epsilon^3 \\ \approx& \epsilon R_0^3 + o(\epsilon) \\ =& 2\epsilon E \cdot \mathcal{D}^H + o(\epsilon). \end{aligned} \tag{18.93}$$

Suppose that

$$R_\epsilon \approx R_0 - \epsilon R' + o(\epsilon), \tag{18.94}$$

so that $\epsilon = (R_0 - R)/R' + o(\epsilon)$. Then

$$d(R) \approx \frac{2(R_0 - R)}{R'} E \cdot \mathcal{D}^H + o(R_0 - R). \tag{18.95}$$

We get agreement with (18.70) provided that

$$\frac{(R_0 - R)}{R'} E \cdot \mathcal{D}^H \approx \frac{2}{R^3} E \cdot \mathcal{D}. \tag{18.96}$$

These results tell us that the Hartree model yields a bifurcation interpretation of the induced dipole (London dispersion) interaction. For $R > R_0$, the two atoms do not interact. In principle, for $R \leq R_0$, they might also not interact, but this state is presumably not stable. At $R = R_0$, a bifurcation occurs, and the induced dipole grows linearly in the parameter $R_0 - R$ according to the expression (18.95).

18.6 Comparisons and conclusions

We now summarize the main conclusions of our results on the approximation of the induced dipole, a.k.a., London dispersion. We have been able to derive the asymptotic form (R^{-6}) of the interaction of induced dipoles, by determining a formula for the asymptotic form of the electron distribution polarization due to the induction. This provides a length scale at which the induced dipole becomes significant in size.

The main result was the asympototic formula for the polarization due to an induced dipole, encapsulated in (18.69). We can interpret the asymptotic formula (18.70) as defining a length scale R_1 at which the induced dipole achieves an order-one effect. It should be remembered that this is dimensionally correct, in that the units of polarization are volume (Section 14.7.2). More precisely, let us define

$$R_1 = (E \cdot \mathcal{D})^{1/3}. \tag{18.97}$$

We now compare the Hartree approximation with the exact derivation in which electron correlations were included.

Comparing (18.97) with (18.91), we can say that R_0 is the Hartree approximation of R_1. The latter is the length scale at which the induced dipole appears (according to the asymptotic theory) to develop an order-one influence. Similarly, in the Hartree theory, R_0 is the scale at which the

induced dipole influence begins. In the Hartree theory, the change in size of the induced dipole is linear in R, so it becomes of unit size almost instantaneously.

The Hartree approximation fails to predict the asymptotic form of the decay of the induced dipole. On the other hand, the exact theory only tells us reliable estimates of the induced dipole only for $R >> R_1$. So in a sense, both theories fail to tell us reliable information for R smaller than the distances (R_0 or R_1) predicted by the asymptotic form of the induced dipole moment. But we can say that the Hartree approximation correctly predicts the length scale at which the induced dipole becomes significant, provided that $R_0 \approx R_1$.

It would be of interest to investigate the form of the electron distribution at distances comparable to, and smaller than, R_1 (and R_0). This would include the effects of repulsive forces (Section 18.7) as well.

18.7 Repulsive molecular forces

In addition to the force of attraction due to London dispersion, it is also of interest to investigate the asymptotic form of the force of repulsion as R decreases. This is often modelled via a Lennard-Jones potential as behaving like R^{-12}, although there is no particular physical basis for this. It is often quoted that an exponential form is more realistic, e.g., espression (2.19) in [81] and the discussion in Chapter XII of [325]. However, these are not exponentials in R^{-1} but rather of the form $c_0 e^{-c_1 R}$ (for constants c_0 and c_1) and hence tend to a finite limit as $R \to 0$. Thus the expressions may only capture the form of the repulsion for large R and not the behavior of the repulsive force as $R \to 0$. Thus it would be of interest to determine exactly what this is.

We do not attempt a general treatment of the repulsive force here, but instead we review calculations of the potential energey of interaction between two hydrogen atoms as given in equation (43-11) on page 343 of [325]. The dominant term in the energy as the distance r_{AB} between the two hydrogen nuclei goes to zero is proportional to $1/r_{AB}$ (and not a higher power). The repulsive effects of the electron-electron interactions is mollified by the integral against the kernel $r_{12}^{-1} = |\mathbf{r}^1 - \mathbf{r}^2|^{-1}$ which acts as a smoothing operator and remain bounded as $r_{AB} \to 0$. This statement requires a proof, which we initiate subsequently with the formal definition of the smoothing operator in (18.98).

The asymptotic behavior of the interaction energy as $r_{AB} \to \infty$ is more complex to describe. It decays like exponentials in $-r_{AB}$ (with various constants) times polynomials in r_{AB} [325]. Thus there is not a simple exponential decay. However, it is possible to establish bounds for decay rates. Again, certain integrals are the key to doing this.

The most complex computation that arises in the analysis of the hydrogen molecule (H_2) in [325] is an integral of the form

$$\int \int f(\mathbf{r}^1) g(\mathbf{r}^2) |\mathbf{r}^1 - \mathbf{r}^2|^{-1} \, d\mathbf{r}^1 d\mathbf{r}^2 \tag{18.98}$$

that can be interpreted (and computed) as follows. The function

$$u(\mathbf{r}^2) := \int f(\mathbf{r}^1) |\mathbf{r}^1 - \mathbf{r}^2|^{-1} \, d\mathbf{r}^1 \tag{18.99}$$

is a solution to the differential equation $-\Delta u = f$ together with Dirichlet boundary conditions ($u(\mathbf{r}) \to 0$ as $|\mathbf{r}| \to \infty$) provided that f decays at infinity sufficiently rapidly (which holds, say, for the wave function for the hydrogen atom). Thus

$$\int\int f(\mathbf{r}^1)g(\mathbf{r}^2)|\mathbf{r}^1 - \mathbf{r}^2|^{-1}\, d\mathbf{r}^1 d\mathbf{r}^2 = \int u(\mathbf{r}^2)g(\mathbf{r}^2)d\mathbf{r}^2. \tag{18.100}$$

In the particular case that $g = f$, this simplifies to

$$\begin{aligned}\int\int f(\mathbf{r}^1)f(\mathbf{r}^2)|\mathbf{r}^1 - \mathbf{r}^2|^{-1}\, d\mathbf{r}^1 d\mathbf{r}^2 &= \int u(\mathbf{r}^2)f(\mathbf{r}^2)d\mathbf{r}^2 \\ &= \int u(\mathbf{r}^2)(-\Delta u(\mathbf{r}^2))d\mathbf{r}^2 \\ &= \int |\nabla u(\mathbf{r}^2)|^2 d\mathbf{r}^2.\end{aligned} \tag{18.101}$$

Since $-\nabla\cdot(\nabla u) = f$, we can think of ∇u as an anti-derivative of f. Hence, the expression in (18.101) may be viewed as a type of Sobolev H^{-1} norm of f. That is,

$$\left(\int\int f(\mathbf{r}^1)f(\mathbf{r}^2)|\mathbf{r}^1 - \mathbf{r}^2|^{-1}\, d\mathbf{r}^1 d\mathbf{r}^2\right)^{1/2} \le \sup_{v\in H_0^1} \frac{\int v(\mathbf{r})f(\mathbf{r})d\mathbf{r}}{\left(\int |\nabla v(\mathbf{r})|^2 d\mathbf{r}\right)^{1/2}}. \tag{18.102}$$

This explains why the electron-electron repulsion is less strong than the proton-proton repulsion, as $|\mathbf{r}_A - \mathbf{r}_B| \to 0$.

To prove (18.102), choose $v = u$ and use (18.101) twice to get

$$\begin{aligned}\left(\int\int f(\mathbf{r}^1)f(\mathbf{r}^2)|\mathbf{r}^1 - \mathbf{r}^2|^{-1}\, d\mathbf{r}^1 d\mathbf{r}^2\right)^{1/2} &= \frac{\int u(\mathbf{r})f(\mathbf{r})d\mathbf{r}}{\left(\int |\nabla u(\mathbf{r})|^2 d\mathbf{r}\right)^{1/2}} \\ &\le \sup_{v\in H_0^1} \frac{\int v(\mathbf{r})f(\mathbf{r})d\mathbf{r}}{\left(\int |\nabla v(\mathbf{r})|^2 d\mathbf{r}\right)^{1/2}}.\end{aligned} \tag{18.103}$$

By Sobolev's inequality, $H_0^1 \subset L^6$; that is $\|v\|_{L^6} \le c\|\nabla v\|_{L^2}$ (in three dimensions). Therefore Hölder's inequality yields

$$\begin{aligned}\left(\int\int f(\mathbf{r}^1)f(\mathbf{r}^2)|\mathbf{r}^1 - \mathbf{r}^2|^{-1}\, d\mathbf{r}^1 d\mathbf{r}^2\right)^{1/2} &\le \sup_{v\in H_0^1} \frac{\int v(\mathbf{r})f(\mathbf{r})d\mathbf{r}}{\left(\int |\nabla v(\mathbf{r})|^2 d\mathbf{r}\right)^{1/2}} \\ &\le \sup_{v\in H_0^1} \frac{\|v\|_{L^6}\|f\|_{L^{6/5}}}{\left(\int |\nabla v(\mathbf{r})|^2 d\mathbf{r}\right)^{1/2}} \\ &\le c\|f\|_{L^{6/5}}.\end{aligned} \tag{18.104}$$

The expression (43-10) on page 343 of [325] is of the form (18.101), with

$$f(\mathbf{r}) = e^{-|\mathbf{r}-\mathbf{r}_A|-|\mathbf{r}-\mathbf{r}_B|}, \tag{18.105}$$

whereas expression (43-8) on page 342 of [325] is of the more general form (18.100) with

$$f(\mathbf{r}) = e^{-2|\mathbf{r}-\mathbf{r}_A|},\ g(\mathbf{r}) = e^{-2|\mathbf{r}-\mathbf{r}_B|}. \tag{18.106}$$

Applying (18.104) to f as in (18.105), we easily see that

$$\begin{aligned}\int\int f(\mathbf{r}^1)f(\mathbf{r}^2)|\mathbf{r}^1-\mathbf{r}^2|^{-1}\,d\mathbf{r}^1 d\mathbf{r}^2 \leq& c^2\|f\|^2_{L^{6/5}}\\ =&c^2\left(\int e^{-(6/5)(|\mathbf{r}-\mathbf{r}_A|+|\mathbf{r}-\mathbf{r}_B|)}\,d\mathbf{r}\right)^{5/3}\\ \leq& C_m e^{-m|\mathbf{r}_A-\mathbf{r}_B|}.\end{aligned} \tag{18.107}$$

for any $m < 2$, where C_m is some constant. The latter estimate follows because for all $\mathbf{r}$

$$e^{-(|\mathbf{r}-\mathbf{r}_A|+|\mathbf{r}-\mathbf{r}_B|)} \leq e^{-|\mathbf{r}_A-\mathbf{r}_B|} \tag{18.108}$$

and because $e^{-\epsilon(|\mathbf{r}-\mathbf{r}_A|+|\mathbf{r}-\mathbf{r}_B|)}$ is integrable for any $\epsilon > 0$. Examining the remaining terms in equation (43-11) on page 343 of [325], we see that

$$W_S - 2W_H \leq C_m e^{-m|\mathbf{r}_A-\mathbf{r}_B|} \text{ as } |\mathbf{r}_A-\mathbf{r}_B| \to \infty \tag{18.109}$$

for any $m < 2$, where C_m is some constant. Here, W_S is energy of the symmetric hydrogen pair, W_H is the energy of a single hydrogen, so that $W_S - 2W_H$ represents the energy increase due to the bond.

Note that the calculations in Chapter XII of [325] that we have reviewed in this section do not reflect the effects of the induced dipole. Thus they only provide a rough guide to potential asymptotic effects. Being based on a Galerkin approximation (with only two basis functions), they only provide an upper-bound on the energy. That is why they miss the significant contribution from the induced dipole. This also means that we cannot prove that there is an $|\mathbf{r}_A - \mathbf{r}_B|^{-1}$ dependence on the energy as $|\mathbf{r}_A - \mathbf{r}_B| \to 0$, only that it is no worse than this. However, they do give an indication that there are contributions to a repulsive force which grow exponentially from zero as $|\mathbf{r}_A - \mathbf{r}_B|$ decreases from infinity, and these can presumably balance the attractive force of the induced dipole, defining the iteratomic distance for a stable configuration. We leave as an exercise (Exercise 18.6) to include basis functions based on the limiting form of the induced dipole to see what new information this approach can provide.

We also note that the calculations in Chapter XII of [325] indicate that there could be a stronger repulsive force, proportional to $|\mathbf{r}_A - \mathbf{r}_B|^{-2}$, for two hydrogen atoms whose electrons have incompatible (meaning: the same) spins.

18.8 The hydrogen atom

The equations for the hydrogen atom, assuming the Born-Oppenheimer approximation, are

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{x}) + V(\mathbf{x})\psi(\mathbf{x}) = \lambda\psi(\mathbf{x}) \tag{18.110}$$

where ∇ denotes the gradient with respect to $\mathbf{x}$ and

$$V(\mathbf{x}) = -\frac{1}{|\mathbf{x}|}. \tag{18.111}$$

There is a simple scaling with respect to the spatial variable. Define $\phi(\mathbf{x}) = \psi(\rho^{-1}\mathbf{x})$, so that $\psi(\mathbf{x}) = \phi(\rho\mathbf{x})$. Then $\nabla^2\psi(\mathbf{x}) = \rho^2\nabla^2\phi(\rho\mathbf{x})$. Also, $V(\rho\mathbf{x}) = \rho^{-1}V(\mathbf{x})$. Thus

$$-\frac{\hbar^2}{2m}\rho^2\nabla^2\phi(\rho\mathbf{x}) + \rho V(\rho\mathbf{x})\phi(\rho\mathbf{x}) = \lambda\phi(\rho\mathbf{x}). \tag{18.112}$$

Dividing by ρ, we find

$$-\frac{\hbar^2}{2m}\rho\nabla^2\phi(\rho\mathbf{x}) + V(\rho\mathbf{x})\phi(\rho\mathbf{x}) = (\lambda/\rho)\phi(\rho\mathbf{x}). \tag{18.113}$$

This means that we can scale the spatial variable such that the constant $\hbar = \sqrt{m}$, and the only change is the scale of the eigenvalues λ. From now on, we will make this assumption.

The solution ψ is spherically symmetric in $\mathbf{x}$, so we use spherical coordinates. In these coordinates ($x = r\cos\theta\,\sin\phi$, $y = r\sin\theta\,\sin\phi$, $z = r\cos\phi$),

$$\Delta v = \frac{\partial^2}{\partial r^2}v + \frac{2}{r}\frac{\partial}{\partial r}v + \frac{1}{r_2}\left(\frac{1}{\sin^2\phi}\frac{\partial^2}{\partial\theta^2}v + \frac{\cos\phi}{\sin\phi}\frac{\partial}{\partial\phi}v + \frac{\partial^2}{\partial\phi^2}v\right). \tag{18.114}$$

Consider the function

$$v(x_1, x_2, x_3) = e^{-c\sqrt{x_1^2+x_2^2+x_3^2}} = e^{-cr}. \tag{18.115}$$

Using the formula (18.114), we find

$$\begin{aligned}\Delta v(x_1, x_2, x_3) =& \left(c^2 - \frac{2c}{r}\right)e^{-cr} \\ =& c^2 v + 2cV(\mathbf{x})v.\end{aligned} \tag{18.116}$$

Thus if we take $c = 1$, we find

$$-\tfrac{1}{2}\Delta v + Vv = -\tfrac{1}{2}v. \tag{18.117}$$

Thus the pair $(v = e^{-r}, \lambda = -\frac{1}{2})$ is an eigenpair for (18.110). This eigenpair is called the **ground state** of the hydrogen atom.

18.9 Ground state modifications

Now consider $w = uv$ where $v = e^{-cr}$ and u is arbitrary. First note that

$$\Delta(uv) = v\Delta u + 2\nabla u\cdot\nabla v + u\Delta v, \tag{18.118}$$

and

$$\nabla r\cdot\nabla u = \frac{\partial u}{\partial r}. \tag{18.119}$$

Furthermore,

$$\nabla v = -ce^{-cr}\nabla r = -cv\nabla r. \tag{18.120}$$

Therefore by (18.116)

$$\begin{aligned}\Delta(uv) =& v\Delta u - 2cv\frac{\partial u}{\partial r} + u\left(c^2 v + 2cV(\mathbf{x})v\right)\\ =& v\left(\Delta u - 2c\frac{\partial u}{\partial r} + u\left(c^2 + 2cV(\mathbf{x})\right)\right).\end{aligned} \tag{18.121}$$

Therefore

$$-\tfrac{1}{2}\Delta(uv) + V(\mathbf{x})uv = v\left(-\tfrac{1}{2}\Delta u + c\frac{\partial u}{\partial r} - u\left(\tfrac{1}{2}c^2 + (c-1)V(\mathbf{x})\right)\right). \tag{18.122}$$

If we choose $c = 1$ and $\lambda = -\frac{1}{2}$, this becomes

$$-\tfrac{1}{2}\Delta(uv) + V(\mathbf{x})uv = v\left(-\tfrac{1}{2}\Delta u + \frac{\partial u}{\partial r} + \lambda u\right). \tag{18.123}$$

Therefore

$$-\tfrac{1}{2}\Delta(uv) + V(\mathbf{x})uv - \lambda uv = v\left(-\tfrac{1}{2}\Delta u + \frac{\partial u}{\partial r}\right). \tag{18.124}$$

18.10 Two hydrogen interaction

Now suppose that $w = w(\mathbf{x}^1, \mathbf{x}^2) = u(\mathbf{x}^1, \mathbf{x}^2)v(\mathbf{x}^1)v(\mathbf{x}^2)$, with $v(\mathbf{x}) = e^{-r}$. By (18.123), we find

$$\begin{aligned}-\tfrac{1}{2}\Delta_1 w - \tfrac{1}{2}\Delta_2 w + V(\mathbf{x}^1)w + V(\mathbf{x}^2)w =& -\tfrac{1}{2}v(\mathbf{x}^2)\Delta_1 uv - \tfrac{1}{2}v(\mathbf{x}^1)\Delta_2 uv + V(\mathbf{x}^1)w + V(\mathbf{x}^2)w\\ =& v(\mathbf{x}^2)v(\mathbf{x}^1)\left(-\tfrac{1}{2}\Delta_1 u + \frac{\partial u}{\partial r_1} + \lambda u\right)\\ &+ v(\mathbf{x}^1)v(\mathbf{x}^2)\left(-\tfrac{1}{2}\Delta_2 u + \frac{\partial u}{\partial r_2} + \lambda u\right)\\ =& v(\mathbf{x}^1)v(\mathbf{x}^2)\left(-\tfrac{1}{2}\Delta_1 u - \tfrac{1}{2}\Delta_2 u + \frac{\partial u}{\partial r_1} + \frac{\partial u}{\partial r_2} + 2\lambda u\right).\end{aligned} \tag{18.125}$$

Therefore

$$-\tfrac{1}{2}\Delta_1 w - \tfrac{1}{2}\Delta_2 w + V(\mathbf{x}^1)w + V(\mathbf{x}^2)w - 2\lambda w = e^{-r_1 - r_2}\left(-\tfrac{1}{2}\Delta_1 u - \tfrac{1}{2}\Delta_2 u + \frac{\partial u}{\partial r_1} + \frac{\partial u}{\partial r_2}\right). \tag{18.126}$$

Thus

$$-\tfrac{1}{2}\Delta_1 w - \tfrac{1}{2}\Delta_2 w + V(\mathbf{x}^1)w + V(\mathbf{x}^2)w - 2\lambda w = e^{-r_1 - r_2}f(\mathbf{x}^1, \mathbf{x}^2) \tag{18.127}$$

if and only if

$$-\tfrac{1}{2}\Delta_1 u - \tfrac{1}{2}\Delta_2 u + \frac{\partial u}{\partial r_1} + \frac{\partial u}{\partial r_2} = f(\mathbf{x}^1, \mathbf{x}^2). \tag{18.128}$$

For example, consider $f = ax_1^1 x_1^2 + x_2^1 x_2^2 + x_3^1 x_3^2$. If we define $\hat{x}_i^j = -x_i^j$ for a fixed value of i and j, and all other entries of $\hat{\mathbf{x}}^k$ agree with $\mathbf{x}^k$, $k = 1, 2$, then $f(\hat{\mathbf{x}}^1, \hat{\mathbf{x}}^2) = -f(\mathbf{x}^1, \mathbf{x}^2)$. Thus it suffices to solve (18.128) on the domain

$$\Omega = \left\{ \mathbf{x} \in \mathbb{R}^6 \;\middle|\; x_i \geq 0 \right\}, \tag{18.129}$$

and the solution on the rest of $\mathbb{R}^6$ can be defined by reflection. Moreover, $u = 0$ on the set

$$\Gamma = \left\{ \mathbf{x} \in \mathbb{R}^6 \;\middle|\; (x_1, x_2, x_3) = 0 \quad \text{or} \quad (x_4, x_5, x_6) = 0 \right\}. \tag{18.130}$$

The vanishing of u comes from the fact that $\lim_{\mathbf{x}\to 0} |V(\mathbf{x})| = \infty$.

Note that we can write

$$f(\mathbf{x}^1, \mathbf{x}^2) = r_1 r_2 F(\omega) \tag{18.131}$$

where $\omega = (\phi^1, \theta^1, \phi^2, \theta^2)$. Define

$$\phi(\mathbf{x}^1, \mathbf{x}^2) = (r_1^2 r_2 + r_1 r_2^2) F(\omega). \tag{18.132}$$

Then by (18.114)

$$\begin{aligned}
&-\tfrac{1}{2}\Delta_1 \phi - \tfrac{1}{2}\Delta_2 \phi + \frac{\partial \phi}{\partial r_1} + \frac{\partial \phi}{\partial r_2} = \\
&-F(\omega)(r_2 + r_1 + 2r_2 + r_1^{-1} r_2^2 + r_2^{-1} r_1^2 + 2r_1 - 2r_1 r_2 - r_2^2 - 2r_2 r_1 - r_1^2) \\
&\quad + \frac{1}{r_1^2}\mu^1 + \frac{1}{r_2^2}\mu^2 \\
&= - F(\omega)(3r_2 + 3r_1 + r_1^{-1} r_2^2 + r_2^{-1} r_1^2 - 4r_1 r_2 - r_2^2 - r_1^2) \\
&\quad + \frac{1}{r_1^2}\mu^1 + \frac{1}{r_2^2}\mu^2
\end{aligned} \tag{18.133}$$

18.10.1 Further eigenvalues

Similarly, consider $w = x_i v$ where $v = e^{-cr}$. we find

$$\begin{aligned}
\Delta w =& 2E^i \cdot \nabla v + x_i \Delta v \\
=& x_i e^{-cr}\left(-\frac{2c}{r} + c^2 - \frac{2c}{r}\right) = x_i e^{-cr}\left(-\frac{4c}{r} + c^2\right),
\end{aligned} \tag{18.134}$$

where $E^i = \nabla x_i$ is the i-th standard unit vector. Therefore

$$-\tfrac{1}{2}\Delta w + V w = \left(\frac{2c-1}{r} - \tfrac{1}{2}c^2\right) w. \tag{18.135}$$

By choosing $c = \frac{1}{2}$, we see that the pair $(w = x_i e^{-r/2}, \lambda = -1/8)$ is an eigenpair for (18.110) for $i = 1, 2, 3$.

In general, if we take $w(\mathbf{x}) = P(\mathbf{x})v(r)$ where $v(r) = e^{-cr}$, we find

$$\begin{aligned}\Delta w =& v\Delta P + 2\nabla P \cdot \nabla v + P\Delta v \\ =& e^{-cr}\left(\Delta P - \frac{2c}{r}\mathbf{x}\cdot\nabla P + \left(c^2 - \frac{2c}{r}\right)P\right) \\ =& c^2 w + e^{-cr}\left(\Delta P - \frac{2c}{r}(\mathbf{x}\cdot\nabla P + P)\right)\end{aligned} \tag{18.136}$$

Thus if P is a harmonic polynomial, we find

$$-\tfrac{1}{2}\Delta w + Vw = -\tfrac{1}{2}c^2 w + \frac{e^{-cr}}{r}\left(c(\mathbf{x}\cdot\nabla P + P) - P\right) \tag{18.137}$$

Thus if

$$\mathbf{x}\cdot\nabla P(\mathbf{x}) = \left(\frac{1}{c} - 1\right)P(\mathbf{x}), \tag{18.138}$$

then $(w = P(\mathbf{x})e^{-cr}, \lambda = -\frac{1}{2}c^2)$ is an eigenpair for (18.110). Suppose that P is homogeneous of degree k, so that

$$P(r\omega) = r^k P(\omega) \tag{18.139}$$

for all $r \geq 0$ and ω on the unit sphere. Then

$$\begin{aligned}\mathbf{x}\cdot\nabla P(r\omega) =& r\lim_{h\to 0}\frac{P((r+h)\omega) - P(r\omega)}{h} \\ =& rP(\omega)\lim_{h\to 0}\frac{(r+h)^k - r^k}{h} \\ =& kr^k P(\omega) = kP(r\omega).\end{aligned} \tag{18.140}$$

Thus (18.138) holds with $c = 1/(k+1)$. The harmonic polynomials that are homogeneous of degree two are linear combinations of

$$x_1x_2,\ x_1x_3,\ x_3x_2,\ x_1^2 - x_2^2,\ x_3^2 - x_2^2. \tag{18.141}$$

In fact, the only polynomials that are homogeneous of degree two and are *not* harmonic are scalar multiples of $r_2 = x_1^2 + x_2^2 + x_3^2$.

18.11 The helium atom

The equations for the helium atom, assuming the Born-Oppenheimer approximation, are of the form

$$-\frac{\hbar^2}{2m}(\nabla^2_{\mathbf{r}^1} + \nabla^2_{\mathbf{r}^2})\psi(\mathbf{r}^1,\mathbf{r}^2) + V(\mathbf{r})\psi(\mathbf{r}^1,\mathbf{r}^2) = \lambda\psi(\mathbf{r}^1,\mathbf{r}^2) \tag{18.142}$$

where $\nabla_{\mathbf{r}^i}$ denotes the gradient with respect to $\mathbf{r}^i$ and

$$V(\mathbf{r}^1,\mathbf{r}^2) = -\frac{2}{|\mathbf{r}^1|} - \frac{2}{|\mathbf{r}^2|} + \frac{1}{|\mathbf{r}^1 - \mathbf{r}^2|}. \tag{18.143}$$

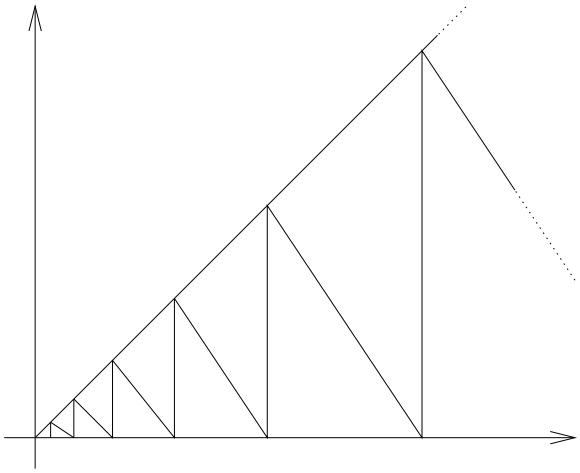


Figure 18.1: Triangulation of the domain $\widetilde{D} = \left\{(r_1, r_2) \in \mathbb{R}^2_+ \;\middle|\; r_2 \le r_1\right\}$ via a Cartesian product grid.

The solution ψ is radially symmetric in each variable $\mathbf{r}^i$, but the final term in (18.143) provides a challenge in terms of integration. Consider an expression of the form

$$\int_{\mathbb{R}^6} \frac{\psi(\mathbf{r}^1, \mathbf{r}^2) w(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2}{|\mathbf{r}^1 - \mathbf{r}^2|} = \int_{\mathbb{R}^2} \psi(r_1, r_2) w(r_1, r_2) v(r_1, r_2) r_1^2 r_2^2\, dr_1 dr_2, \tag{18.144}$$

where we have assumed that the test function w is radially symmetric and

$$v(r_1, r_2) = \int_{S^2 \times S^2} \frac{d\omega d\omega'}{|r_1\omega - r_2\omega'|} = \frac{1}{r_1} \int_{S^2 \times S^2} \frac{d\omega d\omega'}{|\omega - \rho\omega'|}, \quad \rho = r_2/r_1. \tag{18.145}$$

Let us define

$$\phi(\rho) = \int_{S^2 \times S^2} \frac{d\omega d\omega'}{|\omega - \rho\omega'|}. \tag{18.146}$$

By symmetry,

$$\begin{aligned}
\phi(\rho) =& 4\pi \int_{S^2} \frac{d\omega'}{|(0,0,1) - \rho\omega'|} \\
=& 4\pi \int_0^1 \int_0^{2\pi} \frac{(r/\sqrt{1-r^2}) dr d\theta}{|(0,0,1) - \rho(r\cos\theta, r\sin\theta, \sqrt{1-r^2})|} \\
=& 8\pi^2 \int_0^1 \frac{(r/\sqrt{1-r^2}) dr}{(\rho^2 - 2\rho\sqrt{1-r^2} + 1)^{1/2}} \\
=& 8\pi^2 \int_0^1 \frac{ds}{(\rho^2 - 2\rho s + 1)^{1/2}} = \frac{4\pi^2}{\rho} \int_{(\rho-1)^2}^{\rho^2+1} \frac{dt}{\sqrt{t}} \\
=& \frac{8\pi^2}{\rho} \left(\sqrt{\rho^2 + 1} - |\rho - 1| \right).
\end{aligned} \tag{18.147}$$

Note that $\phi(1/\rho) = \rho\phi(\rho)$ for any ρ; also $\phi(0) = 1$, $\phi(1) = \sqrt{2}$ and ϕ is strictly increasing on $[0, 1]$. Therefore

$$v(r_1, r_2) = \frac{\phi(r_2/r_1)}{r_1} = \frac{8\pi^2}{r_1 r_2}\left(\sqrt{r_2^2 + r_1^2} - |r_2 - r_1|\right). \tag{18.148}$$

The vaiational formulation of the Laplace equation in polar coordinates is standard:

$$\begin{aligned}\int_{\mathbb{R}^3} \nabla_{\mathbf{r}}^2 \psi(\mathbf{r}) w(\mathbf{r})\, d\mathbf{r} =& 4\pi \int_0^\infty \left(\frac{\partial^2}{\partial {r_1}^2} + \frac{2}{r_1}\frac{\partial}{\partial r_1}\right)\psi(r_1) w(r_1) r_1^2\, dr_1 \\ =& 4\pi \int_0^\infty \left(\psi''(r_1) + \frac{2}{r_1}\psi'(r_1)\right) w(r_1) r_1^2\, dr_1 \\ =& 4\pi \int_0^\infty -\psi'(r_1)\left(\frac{\partial}{\partial r_1} w(r_1) r_1^2\right) + 2r_1\psi'(r_1) w(r_1)\, dr_1 \\ =& -4\pi \int_0^\infty \psi'(r_1) w'(r_1) r_1^2\, dr_1\end{aligned} \tag{18.149}$$

Therefore

$$-\int_{\mathbb{R}^6} \left((\nabla_{\mathbf{r}^1}^2 + \nabla_{\mathbf{r}^2}^2)\psi(\mathbf{r}^1, \mathbf{r}^2)\right) w(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2 = (4\pi)^2 a(\psi, w), \tag{18.150}$$

where the energy form $a(\cdot, \cdot)$ is defined by

$$a(\psi, w) = \int_0^\infty \int_0^\infty \nabla_{(r_1, r_2)}\psi(r_1, r_2) \cdot \nabla_{(r_1, r_2)} w(r_1, r_2) r_1^2 r_2^2\, dr_1 dr_2. \tag{18.151}$$

The potential term takes the form

$$\begin{aligned} b(\psi, w) =& \int_{\mathbb{R}^6} V(\mathbf{r}^1, \mathbf{r}^2)\psi(\mathbf{r}^1, \mathbf{r}^2) w(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2 \\ =& \int_0^\infty \int_0^\infty \left(v(r_1, r_2) - 32\pi^2/r_1 - 32\pi^2/r_2\right)\psi(r_1, r_2) w(r_1, r_2) r_1^2 r_2^2\, dr_1 dr_2 \\ =& 8\pi^2 \int_0^\infty \int_0^\infty \left(\sqrt{r_2^2 + r_1^2} - |r_2 - r_1| - 4r_2 - 4r_1\right)\psi(r_1, r_2) w(r_1, r_2) r_1 r_2\, dr_1 dr_2. \end{aligned} \tag{18.152}$$

The L^2 inner-product also takes the form

$$\begin{aligned}(\psi, w) =& \int_{\mathbb{R}^6} \psi(\mathbf{r}^1, \mathbf{r}^2) w(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2 \\ =& (4\pi)^2 \int_0^\infty \int_0^\infty \psi(r_1, r_2) w(r_1, r_2) r_1^2 r_2^2\, dr_1 dr_2.\end{aligned} \tag{18.153}$$

Thus we can interpret the problem as having an energy form with a weight

$$\alpha(r_1, r_2) = r_1^2 r_2^2 \tag{18.154}$$

and a potential form with a weight

$$\beta(r_1, r_2) = r_1 r_2 \left(\sqrt{r_2^2 + r_1^2} - |r_2 - r_1| - 4r_2 - 4r_1 \right). \tag{18.155}$$

That is, if we define the r_1, r_2 domain to be D, then the forms are

$$a(\psi, w) = \int_D \alpha \nabla\psi \cdot \nabla w \, dX \quad \text{and} \quad b(\psi, w) = \int_D \beta \psi w \, dX. \tag{18.156}$$

If we divide by $8\pi^2$, the Schrödinger equation is then

$$\frac{\hbar^2}{m} a(\psi, w) + b(\psi, w) = \frac{\lambda}{8\pi^2}(\psi, w) = 2\lambda \int_D \alpha \psi w \, dX. \tag{18.157}$$

The test and trial spaces for w and ψ consist of functions that are anti-symmetric with respect to r_1, r_2. Thus we can restrict to the domain $\widetilde{D}$ consisting of point where $r_1 \geq r_2$ and use the boundary condition $w = \psi = 0$ on $r_1 = r_2$. In this domain, we can write β as

$$\beta(r_1, r_2) = r_1 r_2 \left(\sqrt{r_2^2 + r_1^2} - 3r_2 - 5r_1 \right). \tag{18.158}$$

There are no boundary conditions to be imposed on the lower boundary of $\widetilde{D}$ (where $r_2 = 0$), but there is a compatibility condition arising from the interplay between the $1/r_2$ potential and the $1/r_2$ term in the polar decomposition of the Laplace operator, namely,

$$\frac{\partial \psi}{\partial r_2}(r_1, 0) + \psi(r_1, 0) = 0. \tag{18.159}$$

However, such a condition cannot be imposed variationally due to the vanishing of the weight function in the inner-products. Since $\psi > 0$, this implies that $\psi(r_1, 0)$ must be decreasing. Since $\frac{\partial \psi}{\partial r_2}(r_1, 0) < 0$, ψ is not smooth at $(r_1, 0)$. However, in spherical polar coordinates, this singularity is not apparent, since it occurs at the boundary.

Using a Cartesian product grid as shown in Figure 18.1 raises some interesting questions regarding the quadrature required to compute the various integrals. If we use the trapezoidal rule, then the two integrals without derivatives would vanish, since the weights vanish on the r_1-axis and the test and trial functions vanish on the diagonal $r_1 = r_2$. Thus a higher-order quadrature is required. One simple implementation would be to interpolate the weights α and β in a finite element space. If we choose this space to be quartics, then the representation is exact except for the $\sqrt{r_1^2 + r_2^2}$ term.

18.12 The hydrogen molecule

The equations for the hydrogen molecule, assuming the Born-Oppenheimer approximation, are similar to those for the helium atom (Section 18.11):

$$-\frac{\hbar^2}{2m}(\nabla^2_{\mathbf{r}^1} + \nabla^2_{\mathbf{r}^2})\psi(\mathbf{r}^1, \mathbf{r}^2) + V(\mathbf{r})\psi(\mathbf{r}^1, \mathbf{r}^2) = \lambda\psi(\mathbf{r}^1, \mathbf{r}^2) \tag{18.160}$$

Figure 18.2: Coordinates for the hydrogen molecule.

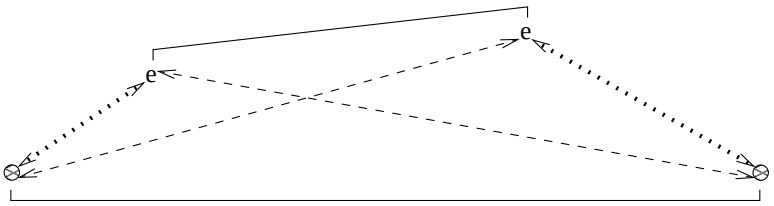


Figure 18.3: Triangulation of the domain $\widetilde{D} = \left\{(r_1, r_2) \in \mathbb{R}^2_+ \;\middle|\; r_2 \leq r_1\right\}$ via a Cartesian product grid.

where $\nabla_{\mathbf{r}^i}$ denotes the gradient with respect to $\mathbf{r}^i$ and

$$V(\mathbf{r}^1, \mathbf{r}^2) = -\frac{1}{|\mathbf{r}^1|} - \frac{1}{|\mathbf{r}^2|} + V_\epsilon(\mathbf{r}^1, \mathbf{r}^2), \tag{18.161}$$

where $\epsilon = 1/R$ and R is the separation distance between the two protons. Here, we have chosen to write the variables for each electron centered on each proton, as indicated in Figure 18.2.

The solution ψ is cylindrically symmetric in each variable $\mathbf{r}^i$, so we use coordinates

$$\mathbf{r}^i = (x_i, r_i \cos\theta_i, r_i \sin\theta_i) \tag{18.162}$$

again as indicated in Figure 18.2. In these coordinates, and for cylindrically symmetric functions v,

$$\Delta v = \frac{\partial^2}{\partial r^2} v + \frac{1}{r}\frac{\partial}{\partial r} v + \frac{\partial^2}{\partial x^2} v. \tag{18.163}$$

Consider the function

$$v(x, y, z) = x e^{-c\sqrt{x^2+y^2+z^2}}. \tag{18.164}$$

Note that $\nabla\sqrt{x^2+y^2+z^2} = \left(\sqrt{x^2+y^2+z^2}\right)^{-1}(x, y, z)$. Then

$$\begin{aligned} \Delta v(x, y, z) &= (1, 0, 0) \cdot \nabla e^{-c\sqrt{x^2+y^2+z^2}} + x\Delta e^{-c\sqrt{x^2+y^2+z^2}} \\ &= \left(\frac{-cx}{\sqrt{x^2+y^2+z^2}}\right) e^{-c\sqrt{x^2+y^2+z^2}} \end{aligned} \tag{18.165}$$

For the potential, we can write

$$\begin{aligned}
-V_\epsilon(\mathbf{r}^1,\mathbf{r}^2) &= \frac{1}{\left((x_1-R)^2+r_1^2\right)^{1/2}} + \frac{1}{\left((x_2-R)^2+r_2^2\right)^{1/2}} - \frac{1}{R} \\
&- \frac{1}{\left((x_1+x_2-R)^2+(r_1\cos\theta_1 - r_2\cos\theta_2)^2+(r_1\sin\theta_1-r_2\sin\theta_2)^2\right)^{1/2}} \\
&= \frac{1}{R}\Big(\left((\epsilon x_1-1)^2+(\epsilon r_1)^2\right)^{-1/2} + \left((\epsilon x_2-1)^2+(\epsilon r_2)^2\right)^{-1/2} - 1 \\
&- \left((\epsilon(x_1+x_2)-1)^2+(\epsilon r_1\cos\theta_1-\epsilon r_2\cos\theta_2)^2+(\epsilon r_1\sin\theta_1-\epsilon r_2\sin\theta_2)^2\right)^{-1/2}\Big) \\
&= \frac{1}{R}\Big(\left((\epsilon x_1-1)^2+(\epsilon r_1)^2\right)^{-1/2} + \left((\epsilon x_2-1)^2+(\epsilon r_2)^2\right)^{-1/2} - 1 \\
&- \left((\epsilon(x_1+x_2)-1)^2+\epsilon^2(y_1-y_2)^2+\epsilon^2(z_1-z_2)^2\right)^{-1/2}\Big)
\end{aligned} \tag{18.166}$$

Therefore, if we write $\rho_i = \sqrt{x_i^2+r_i^2}$, then since $(1+t)^{-1/2} \approx 1 - \frac{1}{2}t + \frac{3}{2}t^2 + \mathcal{O}(t^3)$

$$\begin{aligned}
-RV_\epsilon(\mathbf{r}^1,\mathbf{r}^2) &= \left(1-2\epsilon x_1+\epsilon^2\rho_1^2\right)^{-1/2} + \left(1-2\epsilon x_2+\epsilon^2\rho_2^2\right)^{-1/2} - 1 \\
&-\Big(1-2\epsilon(x_1+x_2)+\epsilon^2(x_1+x_2)^2 \\
&+\epsilon^2\left(r_1^2+r_2^2-2r_1r_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2)\right)\Big)^{-1/2} \\
&\approx -\tfrac{1}{2}\epsilon^2\left(\rho_1^2+\rho_2^2-(x_1+x_2)^2-r_1^2-r_2^2+2r_1r_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2)\right) \\
&+6\epsilon^2\left(x_1^2+x_2^2-(x_1+x_2)^2\right)+\mathcal{O}(\epsilon^3) \\
&= -\epsilon^2\big(11x_1x_2+r_1r_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2)\big)+\mathcal{O}(\epsilon^3) \\
&= -\epsilon^2\big(11x_1x_2+y_1y_2+z_1z_2\big)+\mathcal{O}(\epsilon^3)
\end{aligned} \tag{18.167}$$

Thus

$$R^3V_\epsilon(\mathbf{r}^1,\mathbf{r}^2) = 11x_1x_2 + r_1r_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2)+\mathcal{O}(\epsilon) \tag{18.168}$$

as $\epsilon = R^{-1} \to 0$. Thus the equation (**??**) reduces asymptotically to

$$\left(-\frac{\hbar^2}{2m}\left(\nabla^2_{\mathbf{r}^1}+\nabla^2_{\mathbf{r}^2}\right) - \frac{1}{|r_1|} - \frac{1}{|r_2|} - \lambda_0\right)\psi'(\mathbf{r}^1,\mathbf{r}^2) = (11x_1x_2+y_1y_2+z_1z_2)\psi_0(\mathbf{r}^1,\mathbf{r}^2). \tag{18.169}$$

To evaluate integrals involving V_ϵ, we need to compute integrals of the potential v defined by

$$\begin{aligned}
v(x_1,r_1,x_2,r_2,\theta_1,\theta_2) &= \left((x_1+x_2-R)^2+(r_1\cos\theta_1-r_2\cos\theta_2)^2+(r_1\sin\theta_1-r_2\sin\theta_2)^2\right)^{-1/2} \\
&= \left((x_1+x_2-R)^2+r_1^2+r_2^2-2r_1r_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2)\right)^{-1/2} \\
&= \frac{1}{\sqrt{2r_1r_2}}\left(\frac{(x_1+x_2-R)^2}{2r_1r_2}+\frac{r_1^2+r_2^2}{2r_1r_2}-(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2)\right)^{-1/2}.
\end{aligned} \tag{18.170}$$

In particular, we require the integrals

$$\gamma(t) = \int_0^{2\pi} \int_0^{2\pi} \frac{d\theta_1 d\theta_2}{\sqrt{t - (\cos\theta_1 \cos\theta_2 + \sin\theta_1 \sin\theta_2)}} \tag{18.171}$$

for $t \geq 1$. The asymptotics of γ as $t \to 1$ and $t \to \infty$ are the same as the function

$$2\frac{\sqrt{1 + 400(t-1)}}{\sqrt{t(t-1)}}, \tag{18.172}$$

although they are different for t close to 2.

The energy inner-product is of the form

$$a(\psi, w) = \int_0^\infty \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty (\nabla_{(x_1,r_1,x_2,r_2)}\psi) \cdot (\nabla_{(x_1,r_1,x_2,r_2)} w)\; r_1 r_2 \, dx_1 dx_2 dr_1 dr_2. \tag{18.173}$$

The L^2 inner-product also takes the form

$$\begin{aligned}(\psi, w) =& \int_{\mathbb{R}^6} \psi(\mathbf{r}^1, \mathbf{r}^2) w(\mathbf{r}^1, \mathbf{r}^2)\, d\mathbf{r}^1 d\mathbf{r}^2 \\ =& (2\pi)^2 \int_0^\infty \int_0^\infty \int_{-\infty}^\infty \int_{-\infty}^\infty \psi(x_1, r_1, x_2, r_2) w(x_1, x_2, r_1, r_2) r_1 r_2 \, dx_1 dx_2 dr_1 dr_2.\end{aligned} \tag{18.174}$$

The potential form can again be written

$$b(\psi, w) = \int_D \beta \psi w \, dx_1 dx_2 dr_1 dr_2, \tag{18.175}$$

where the domain D now is $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ and the weight is defined by

$$\begin{aligned}\beta(x_1, r_1, x_2, r_2) =& -\frac{r_1 r_2}{(x_1^2 + r_1^2)^{1/2}} - \frac{r_1 r_2}{(x_2^2 + r_2^2)^{1/2}} \\ &+ \frac{r_1 r_2}{((x_1 - R)^2 + r_1^2)^{1/2}} + \frac{r_1 r_2}{((x_2 - R)^2 + r_2^2)^{1/2}} - \frac{r_1 r_2}{R} \\ &- \frac{\sqrt{r_1 r_2}}{\sqrt{2}} \gamma\left(\left((x_1 + x_2 - R)^2 + r_1^2 + r_2^2\right)/2 r_1 r_2\right).\end{aligned} \tag{18.176}$$

If we divide by $4\pi^2$, the Schrödinger equation is then

$$\frac{\hbar^2}{m} a(\psi, w) + b(\psi, w) = \lambda \int_D \psi w \; r_1 r_2 \, dx_1 dx_2 dr_1 dr_2. \tag{18.177}$$

The test and trial spaces for w and ψ consist of functions that are anti-symmetric with respect to their positions. However, since the coordinates are different for the different electrons, this is

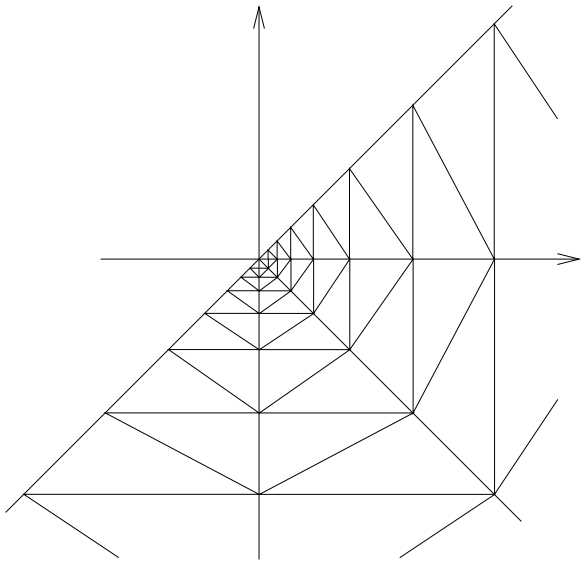


Figure 18.4: Triangulation of the domain $\widetilde{D}_x = \left\{(x_1, \hat{x}_2) \in \mathbb{R}^2 \;\middle|\; \hat{x}_2 \leq x_1\right\}$ via a Cartesian product grid.

somewhat hard to interpret. It becomes easier if we introduce the new variable $\hat{x}_2 = R - x_2$, so that the old variable $x_2 = R - \hat{x}_2$. In these coordinates, the antisymmetry is easy to express:

$$\psi(x_1, r_1, \hat{x}_2, r_2) = -\psi(\hat{x}_2, r_2, x_1, r_1). \tag{18.178}$$

The (x_1, x_2) coordinates of the corresponding domain form the domain $\widetilde{D}_x$ which is depicted in Figure 18.4 together with a Cartesian product mesh. For each point in this domain, there is a domain of (r_1, r_2) coordinates in a domain that looks like Figure 18.1.

The energy and L^2 inner-products are unchanged by the change of coordinates to $\hat{x}_2$, and the new potential term takes the form

$$\begin{aligned}\hat{\beta}(x_1, r_1, \hat{x}_2, r_2) = &- \frac{r_1 r_2}{\left(x_1^2 + r_1^2\right)^{1/2}} - \frac{r_1 r_2}{\left((\hat{x}_2 - R)^2 + r_2^2\right)^{1/2}} \\ &+ \frac{r_1 r_2}{\left((x_1 - R)^2 + r_1^2\right)^{1/2}} + \frac{r_1 r_2}{\left(\hat{x}_2^2 + r_2^2\right)^{1/2}} - \frac{r_1 r_2}{R} \\ &- \frac{\sqrt{r_1 r_2}}{\sqrt{2}} \gamma \left(\left((x_1 - \hat{x}_2)^2 + r_1^2 + r_2^2\right)/2 r_1 r_2\right).\end{aligned} \tag{18.179}$$

One quantity of interest is the induced dipole in the hydrogen molecule. In the orginal coordinates, this can be written as

$$\mu = \int_{\mathbb{R}^6} \mathbf{r}^1 |\psi(\mathbf{r}_1, \mathbf{r}_2)|^2 \, d\mathbf{r}^1 d\mathbf{r}^2, \tag{18.180}$$

where μ is a vector. Note that we are utilizing the fact that the proton is at the origin and the charges are of unit size. Writing (18.180) in terms of cylindrical coordinates, i.e., $\mathbf{r}^1 = (x_1, r_1 \cos\theta_1, r_1 \sin\theta_1)$,

we see that only the x component of μ is nonzero, since ψ is independent of θ^i. Thus the remaining component of the induced dipole is

$$\mu_x = 4\pi^2 \int_D |\psi(x_1, r_1, \hat{x}_2, r_2)|^2 \, x_1 r_1 r_2 \, dx_1 d\hat{x}_2 dr_1 dr_2. \tag{18.181}$$

This can be reduced to the antiysmmetry domain $\widetilde{D}$ as

$$\mu_x = 16\pi^2 \int_{\widetilde{D}} |\psi(x_1, r_1, \hat{x}_2, r_2)|^2 \, x_1 r_1 r_2 \, dx_1 d\hat{x}_2 dr_1 dr_2. \tag{18.182}$$

One thing of interest is to compute how μ_x depends on R as $R \to \infty$. Another is to study the behavior of μ_x for small R, e.g., to determine if there is a value of R where $\mu_x = 0$.

18.13 The Madelung equation

The Schrödinger equation can be recast by a simple change of variables. Write $\psi = e^{R+iS}$.

But there is a constraint [416] that must be satisfied by $\mathbf{v} = \nabla S$, namely,

$$\oint_L \mathbf{v} \cdot d\ell = 2\pi j \tag{18.183}$$

for some j, for any closed loop L.

18.14 Exercises

Exercise 18.1 *Verify the derivation of (18.39).*

Exercise 18.2 *Verify the computations in (18.61).*

Exercise 18.3 *Suppose that $X(s)$ is a smooth function from $[0,1]$ into an inner-product space with the property that $\|X(s)\| = 1$ for all s. Let Y be the derivative of X at $s = 0$. Show that Y and $X(0)$ are orthogonal. (Hint: just expand*

$$1 = \|X(s)\|^2 = \|X(0) + sY + \mathcal{O}(s^2)\| = \|X(0)\|^2 + 2s(X(0), Y) + \mathcal{O}(s^2).)$$

Exercise 18.4 *Derive an expression for the term R' in (18.94) in terms of f', f_0, and any other required quantities. Under what conditions can you assert that $R' \neq 0$?*

Exercise 18.5 *Carry out the computations indicated in (18.69) and (18.92) in the case of the hydrogen atom (in which case f_0 is known in closed form [325]). This requires solving (18.67) and (18.90) to determine f'.*

Exercise 18.6 *Improve the computations in section 43a. in [325] by including as basis function f' computed in Exercise 18.5. Note that f_0 and f' are orthogonal.*

Exercise 18.7 *Carry out a full numerical simulation of the exact problem for the interaction of the hydrogen molecule (cf. Exercise 18.5 and Exercise 18.6). Consider using a Galerkin method in which the basis functions used in Exercise 18.6 are augmented in some way and an adaptive scheme is used for mesh refinement. This may be essential to control the memory requirements for this six-dimensional problem. Also explore dimensional reductions that may be possible due to symmetries of the problem.*

Chapter 19

Continuum equations for electrostatics

The basic equations of electrostatics for a collection of charges of strength q_i at positions $\mathbf{r}_i$ can be derived from the simple expression

$$\nabla\cdot(\epsilon_0\mathbf{e}) = \sum_i q_i\delta(\mathbf{r}-\mathbf{r}_i) \tag{19.1}$$

where ϵ_0 is the permittivity of the vacuum. Here $\mathbf{e}$ is the induced electric field.

We will split the 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 charge groups with net charge zero, and ρ denotes the remainder of the charge density.

19.1 Understanding dielectrics

Dipole molecules can moderate charge by aligning with an external field. This effect is small on a small scale, but the effect of many dielectric molecules can coordinate on a large scale to achieve a substantial effect. In Figure 19.1, we present a simple example of a single charge moderated by a string of dipoles along a single line. There is a detectable depression in the average potential due to the dipole alignments.

Water is really a double dipole, with an effective dipole. In Tip5P, there are partial charges $q_1 = -q_2 = 0.241$, positioned at distances $l_1 = 0.9572$ and $l_2 = 0.70$ from the oxygen center.

19.2 Dielectric materials

The dielectric properties of materials are important in many contexts [45, 378, 279]. A **dielectric** medium [53] is characterized by the fact that the charges are organized in local groups with net charge zero. Specifically we assume that (at least part of) the $\mathbf{r}_i$ and q_i 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:

$$\sum_k q_{jk} = 0\,. \tag{19.2}$$

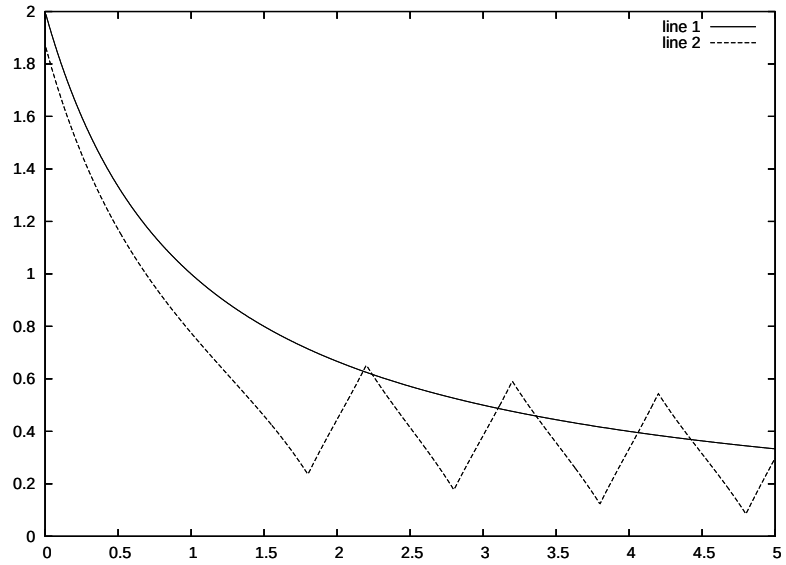


Figure 19.1: Single charge (+2) at zero and dipoles at 2,3,...,10 of width 0.4 and strength 1.

Then the expression for the charge density can be simplified as

$$\gamma(\mathbf{r}) = \sum_j \sum_k q_{jk} \delta(\mathbf{r} - \mathbf{r}_j - \mathbf{r}_{jk}) \tag{19.3}$$

The expression δ in (19.1) 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 [369]. In [186], 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 [186]. A closer approximation might be obtained by letting q be fractional, with positions $\mathbf{r}_{jk}$ chosen to improve the representation [270].

Let us suppose that the charge groups are homogeneous in the sense that

$$\mathbf{r}_{jk} = \mathcal{R}(\theta_j)\rho_k \tag{19.4}$$

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 [270]. Then

$$\begin{aligned} \sum_k q_{jk} \delta(\mathbf{r} - \mathbf{r}_j - \mathbf{r}_{jk}) =& \sum_k q_k \delta(\mathbf{r} - \mathbf{r}_j - \mathcal{R}(\theta_j)\rho_k) \\ =& \mathcal{F}(\theta_j, \mathbf{r} - \mathbf{r}_j) \end{aligned} \tag{19.5}$$

where $\mathcal{F}$ is defined by

$$\mathcal{F}(\theta, \mathbf{r}) = \sum_k q_k \delta(\mathbf{r} - \mathcal{R}(\theta)\rho_k). \tag{19.6}$$

Now suppose that δ is rotationally invariant. Then

$$\begin{aligned} \mathcal{F}(\theta, \mathcal{R}(\theta)\mathbf{r}) =& \sum_k q_k \delta(\mathcal{R}(\theta)\mathbf{r} - \mathcal{R}(\theta)\rho_k) \\ =& \sum_k q_k \delta(\mathbf{r} - \rho_k) \\ =& \nabla\cdot\mathcal{W}(\mathbf{r}), \end{aligned} \tag{19.7}$$

with $\mathcal{W}(\mathbf{r}) = \nabla\psi(\mathbf{r})$ where ψ solves a Poisson equation of the form

$$\Delta\psi = \sum_k q_k \delta(\mathbf{r} - \rho_k). \tag{19.8}$$

If δ is the Dirac δ-function, then $\mathcal{W}$ is a generalized multipole expression

$$\mathcal{W}(\mathbf{r}) = -\sum_k q_k \frac{\mathbf{r} - \rho_k}{|\mathbf{r} - \rho_k|^3}. \tag{19.9}$$

Then

$$\begin{aligned} \sum_k q_{jk} \delta(\mathbf{r} - \mathbf{r}_j - \mathbf{r}_{jk}) =& \mathcal{F}(\theta_j, \mathbf{r} - \mathbf{r}_j) \\ =& \nabla\cdot\mathcal{W}(\mathcal{R}(\theta_j)^t(\mathbf{r} - \mathbf{r}_j)). \end{aligned} \tag{19.10}$$

Therefore we have an exact representation of the dielectric field γ defined in (19.3), viz.,

$$\gamma(\mathbf{r}) = \sum_j \nabla\cdot\mathcal{W}(\mathcal{R}(\theta_j)^t(\mathbf{r} - \mathbf{r}_j)). \tag{19.11}$$

19.3 Polarization field

Now, let us suppose that there is an additional charge density ρ that is not related to the dielectric, and thus the equations (19.1) would be written

$$\nabla\cdot(\epsilon_0\mathbf{e}) = \rho + \sum_j \nabla\cdot\mathcal{W}(\mathcal{R}(\theta_j)^t(\mathbf{r} - \mathbf{r}_j)). \tag{19.12}$$

Let us define the **polariztion p** by

$$\mathbf{p}(\mathbf{r}) = -\sum_j \mathcal{W}(\mathcal{R}(\theta_j)^t(\mathbf{r} - \mathbf{r}_j)), \tag{19.13}$$

and set $\mathbf{d} = \epsilon_0 \mathbf{e} + \mathbf{p}$. Then

$$\nabla \cdot \mathbf{d} = \epsilon_0 \nabla \cdot \mathbf{e} + \nabla \cdot \mathbf{p} = \gamma + \nabla \cdot \mathbf{p} = \rho \tag{19.14}$$

The electric field $\mathbf{e}$ thus satisfies $\epsilon_0 \mathbf{e} = \mathbf{d} - \mathbf{p}$.

The directions θ_j that determine the polarization tend on average to cause $\mathbf{p}$ to line up with the induced field. In fact, in a thermalized system, there will be fluctuations in the angles θ_j, and we can only talk about the mean angles. Debye [89] suggested that we can write $\widetilde{\mathbf{p}} = (\epsilon - \epsilon_0)\widetilde{\mathbf{e}}$ where ϵ denotes an effective permittivity. Here, $\widetilde{\mathbf{p}}$ (resp., $\widetilde{\mathbf{e}}$) denotes a temporal average over a timescale that is long with respect to the basic thermal motions. The latter occur on the order of fractions of picoseconds, so we could imaging a time average of the order of picoseconds.

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. That is, we write

$$\mathbf{p} = (\epsilon - \epsilon_0)\mathbf{e} + \zeta \mathbf{e}^{\perp}, \tag{19.15}$$

where ϵ is defined by

$$\epsilon = \epsilon_0 + \frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}}, \tag{19.16}$$

with the appropriate optimism that $\mathbf{p} = 0$ when $\mathbf{e} = 0$. 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

$$\widetilde{\epsilon} = \epsilon_0 + \left\langle \frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}} \right\rangle, \tag{19.17}$$

should be well behaved. For simplicity, we will drop the tildes and think from now on that everything represents temporal, or spatial, averages.

The expression (19.17) 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

$$\frac{\mathbf{p} \cdot \mathbf{e}}{\mathbf{e} \cdot \mathbf{e}}, \tag{19.18}$$

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 $\widetilde{\epsilon} < 0$.

19.4 Frequency dependence

The relationship proposed by Debye between $\mathbf{p}$ and $\mathbf{e}$ depends on frequency. In particular, we have

$$\epsilon(\nu) = \epsilon_0 + \frac{\epsilon_1 - \epsilon_0}{1 + \nu^2 \tau^2} \tag{19.19}$$

where τ is a characteristic time associated with the dielectric material.

This relationship has been verified extensively by experimental data, as indicated in Figure 19.2. But it is instructive to review the theoretical derivation that Debye gave to justify the behavior.

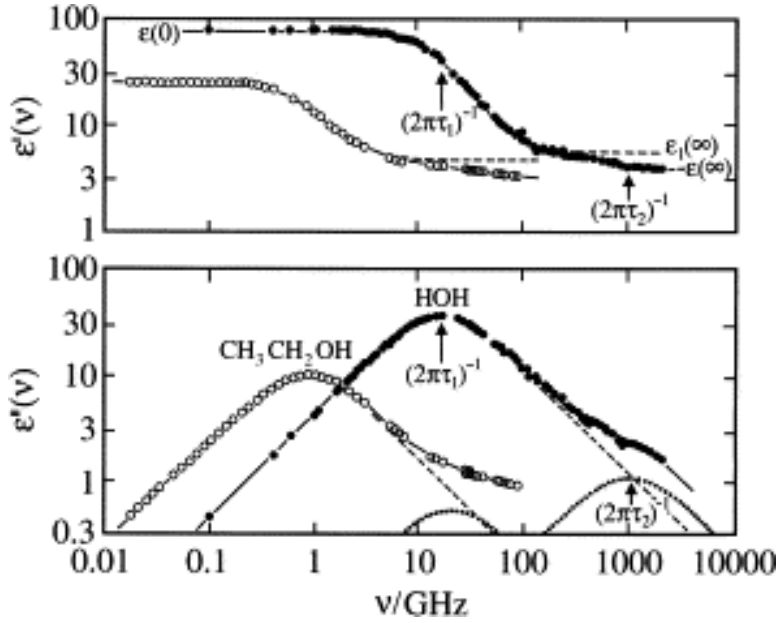


Figure 19.2: This figure is reproduced from [212] and we intend to get the permission of the authors and journal.

19.5 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 correspond to even higher frequencies, although at a smaller amplitude.

In addition, the dielectric coefficient varies by a factor of nearly one hundred from inside the protein to the dielectric bulk 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. This further engenders high frequency components in the electric field. Thus there are high-frequency components in the electric field in both 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 use the ansatz that the dielectric properties depend on spatial wave number ξ proportional to a factor κ where

$$\kappa(\xi) = \epsilon_0 + \frac{\epsilon_1 - \epsilon_0}{1 + |\lambda\xi|^2} \tag{19.20}$$

where λ is the length scale of the transition from one value of the dielectric to the other. In general, λ could be a matrix, allowing for anisotropy. But for the time being we will think of it as a scalar.

Empirical evidence suggests that the length scale λ should be approximately 1.7 Ångstroms [371] when modelling water in a bulk environment.

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 $\nu = 1/\lambda$, and write

$$\kappa(\mathbf{r}, \xi) = \epsilon_0 + \nu(\mathbf{r}) \frac{\epsilon_1 - \epsilon_0}{\nu(\mathbf{r}) + |\xi|^2} \tag{19.21}$$

where $\nu(\mathbf{r})$ is a spatial frequency scale. However, for simplicity we assume that the model (19.20) is sufficient for the moment.

19.5.1 Poisson-Debye equation: bulk case

We can expand $\mathbf{e}$ and $\mathbf{p}$ in a Fourier series and use the Debye-like relationship (19.20) to relate the resulting coefficients in the series. That is, we have (using the inverse Fourier transform)

$$\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi) \widehat{\mathbf{e}}(\xi)\, d\xi = \mathbf{p}(\mathbf{r})\,, \tag{19.22}$$

where here and subsequently we use the notation $\widehat{u}$ to denote the Fourier transform of a function u:

$$\widehat{u}(\xi) := \int_{\mathbb{R}^3} e^{i\xi\cdot\mathbf{r}} u(\mathbf{r})\, d\mathbf{r}\,. \tag{19.23}$$

Therefore the basic equation is

$$\nabla\cdot \left(\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi) \widehat{\mathbf{e}}(\xi)\, d\xi \right) = 4\pi\rho(\mathbf{r})\,. \tag{19.24}$$

We can write $\mathbf{e} = \nabla\phi$ using Maxwell's equations. Therefore $\widehat{\mathbf{e}}(\xi) = i\xi\widehat{\phi}(\xi)$. Therefore, (19.23) becomes

$$\nabla\cdot \left(\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi) i\xi \widehat{\phi}(\xi)\, d\xi \right) = 4\pi\rho(\mathbf{r})\,. \tag{19.25}$$

Taking the Fourier transform (19.23) of (19.25) provides the simple relation

$$\widehat{\phi}(\xi) = \frac{4\pi\widehat{\rho}(\xi)}{|\xi|^2 \kappa(\xi)} \tag{19.26}$$

which can be used to compute ϕ (and thus $\mathbf{e}$) from ρ.

The expression (19.24) can be simplified in certain limits. We have

$$\frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} e^{-i\mathbf{r}\cdot\xi} \kappa(\xi) \widehat{\mathbf{e}}(\xi)\, d\xi \approx \epsilon_j \mathbf{e}(\mathbf{r}) \tag{19.27}$$

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 (19.24) as a partial differential equation, except approximately in special cases.

When a non-dielectric material is immersed in a dielectric (e.g., a protein in water), it might be plausible to approximate (19.27) 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 macromolecular systems in solvent [268]. However, it is not clear what to do when very small non-dielectric objects, such as nanotubes [] 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 [371], 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 lose freedom 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 (19.24) involves a Fourier integral operator [104]. Due to the special form of $\kappa(\nu)$, it is possible to write (19.24) as a fourth-order elliptic partial differential operator for the potential ϕ:

$$\nabla\cdot\left(\left(\epsilon_1 - \epsilon_0\lambda^2\Delta\right)\nabla\phi\right) = \left(1 - \lambda^2\Delta\right)\rho \tag{19.28}$$

provided that λ is constant. 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.

19.5.2 Response to a point charge

The first question to ask with the model (19.24) is what the electric field (or potential) looks like for a single charge $\rho = 4\pi\delta$. More precisely, (19.24) with $\rho = 4\pi\delta$ defines a family of potentials $\phi_{\epsilon_r,\lambda}$ for any given ϵ_r and λ, where $\epsilon_r = \epsilon_1/\epsilon_0$. Thus a simple change of variables implies that

$$\phi(\mathbf{r}) = \frac{\epsilon_0}{\lambda}\phi_{\epsilon_r,1}(\mathbf{r}/\lambda) \tag{19.29}$$

where $\phi_{\epsilon_r,1}$ is defined using the kernel $\kappa = 1 + \frac{\epsilon_r - 1}{1+\xi^2}$. In the limit $\lambda \to \infty$ we find $\phi_\infty = c/\epsilon_0 r$.

The computation of $\phi_{\epsilon_r,1}$ requires us to consider the Fourier transform of a radially symmetric function u (which is itself radially symmetric). The following formula holds:

$$|\xi|\widehat{u}(\xi) := 4\pi\int_0^\infty ru(r)\sin(|\xi|r)\,dr \tag{19.30}$$

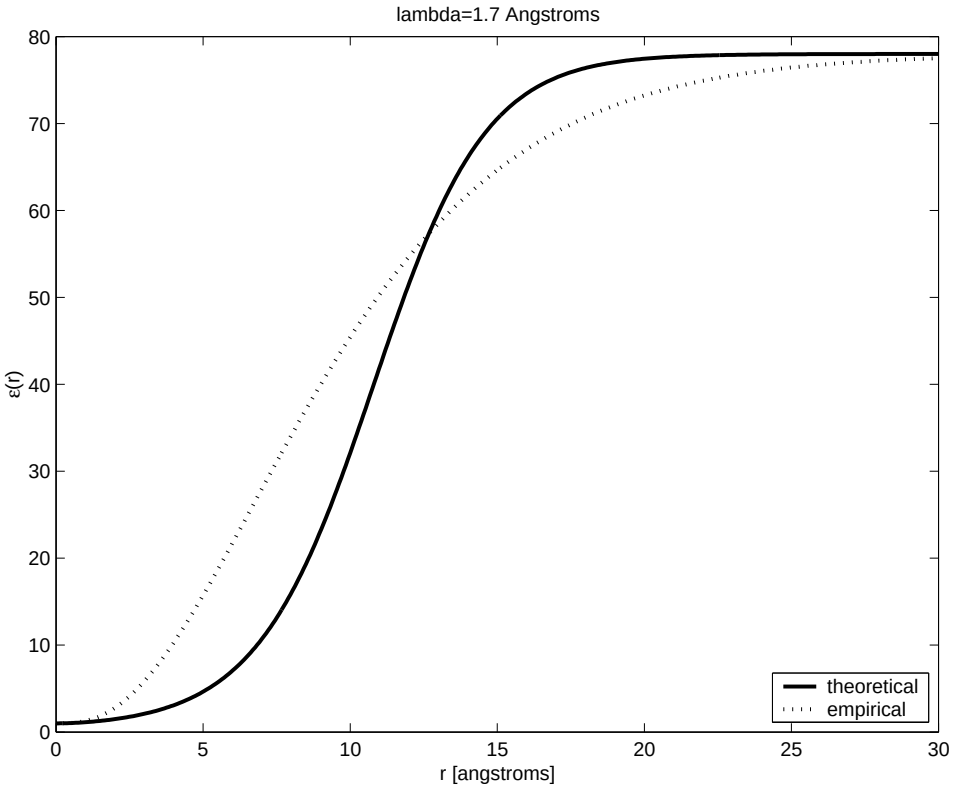


Figure 19.3: Comparison of empirical permittivity formula in [440] (dotted line) with the model (19.24) (solid line) where κ is defined in (19.20) with $\lambda = 1.7$. Plotted are the ratios of the effective permittivity experienced by a dipole to the vacuum permittivity as a function of the separation distance in Å, where the bulk permittivity is that of water.

where $u(r)$ means $u(\mathbf{r})$ with $r = |\mathbf{r}|$.

Thus the expression $\epsilon_0\phi_\infty/\phi_\lambda = c/r\phi_\lambda$ is depicted in Figure 19.3. This indicates that ϕ_λ behaves like ϕ_∞, the vacuum potential for a point charge, near the point charge. At a distance of 20 Å from the charge, the potential has been reduced by a factor of nearly eighty due to the effect of the dielectric. However, note that the dielectric does not take full effect until a distance of over 10λ is reached.

There is a physical interpretation for the suppression of the dielectric effect near the point charge. In the immediate vicinity of the charge, water molecules are strongly structured. In particular, the immediate layer of hydrogen bonds between the point charge and the first layer of water molecules would be expected to be nearly constant. The immobilization of such a layer of water immediately surrounding the charge leads to significant water structuring in subsequent layers. Thus only when one reaches a significant distance (e.g., 10λ) are the water molecules without noticeable effect from the point charge.

It is useful to observe that the permittivity of free space ϵ_0 can be written in convenient units as $\epsilon_0 = 0.24 q_e^2 (\text{kcal/mol})^{-1} \mu\text{m}^{-1}$ where $-q_e$ is the charge of the electron.

19.5.3 Non-local relationship between p and e

A model has been used by different people [244, 230, 119, 74] 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

$$\mathbf{p}(\mathbf{r}) = \int K(\mathbf{r}, \mathbf{r}')\mathbf{e}(\mathbf{r}')\, d\mathbf{r}' \tag{19.31}$$

where the averaging kernel K satisfies

$$K(\mathbf{r}, \mathbf{r}') = K(\mathbf{r} - \mathbf{r}') = \int e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')}\kappa(\mathbf{k})\, d\mathbf{k} \tag{19.32}$$

with the expression κ representing the Debye-like frequency dependence (19.20). Taking Fourier transforms, we see that the "non-local" model (19.20) is the same as (19.22). However, it is not possible to represent the mollifier K as an ordinary function. Clearly, $K = \epsilon_0\delta + (\epsilon_1 - \epsilon_0)\widetilde{K}$ where

$$\widetilde{K}(\mathbf{r}) = \int e^{i\mathbf{k}\cdot\mathbf{r}}\tilde{\kappa}(\mathbf{k})\, d\mathbf{k} \tag{19.33}$$

and $\tilde{\kappa}$ is defined by

$$\tilde{\kappa}(\mathbf{k}) = \frac{1}{1 + |\mathbf{k}|^2\lambda^2}. \tag{19.34}$$

We easily identify $\tilde{\kappa}$ as the Fourier transform of the fundamental solution of the Laplace operator $1 - \lambda^2\Delta$, so that

$$\widetilde{K}(\mathbf{r}) = \frac{\lambda e^{-|\mathbf{r}|/\lambda}}{4\pi|\mathbf{r}|}. \tag{19.35}$$

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 [369]) as

$$K(\mathbf{r}) = \epsilon_0\delta + \frac{\lambda(\epsilon_1 - \epsilon_0)e^{-|\mathbf{r}|/\lambda}}{4\pi|\mathbf{r}|}. \tag{19.36}$$

More precisely, we have

$$\mathbf{p}(\mathbf{r}) = \epsilon_0\mathbf{e}(\mathbf{r}) + \frac{\lambda(\epsilon_1 - \epsilon_0)}{4\pi}\int \frac{e^{-|\mathbf{r}-\mathbf{r}'|/\lambda}}{|\mathbf{r} - \mathbf{r}'|}\mathbf{e}(\mathbf{r}')\, d\mathbf{r}'. \tag{19.37}$$

19.6 The Poisson-Debye equation: general case

A general relationship between $\mathbf{p}$ and $\mathbf{e}$ of the form $\mathbf{p} = \mathcal{P}\mathbf{e}$ leads to an equation for the electric field of the following form:

$$\nabla\cdot\mathcal{P}\nabla\phi = \rho. \tag{19.38}$$

We can write a general relationship between the electric field and the polarization vector using an operator $\mathcal{P}$ (defined component-wise) given by

$$\mathcal{P}v(\mathbf{r}) = \int_{\mathbb{R}^3}\int_{\mathbb{R}^3} e^{i\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}')}\kappa(\mathbf{k},\mathbf{r},\mathbf{r}')v(\mathbf{r}')\,d\mathbf{k}\,d\mathbf{r}' \tag{19.39}$$

where we define the symbol κ by the Debye-like relationship

$$\kappa(\mathbf{k},\mathbf{r},\mathbf{r}') = \epsilon_0 + \frac{\epsilon_1-\epsilon_0}{1+|\lambda(\mathbf{r},\mathbf{r}')\mathbf{k}|^2}, \tag{19.40}$$

where we have assumed that the length scale is allowed to vary as a function of the spatial coordinate.

Our justification for such an approach is based on the fact that models in which λ is constant but of different size depending on the context have been successful in modelling dielectric behavior near boundaries [371]. At boundaries, λ increases to infinity, so it is perhaps easier to express κ in terms of $\nu(\mathbf{r},\mathbf{r}') = 1/\lambda$:

$$\kappa(\mathbf{k},\mathbf{r},\mathbf{r}') = \epsilon_0 + \frac{(\epsilon_1-\epsilon_0)\nu(\mathbf{r},\mathbf{r}')^2}{\nu(\mathbf{r},\mathbf{r}')^2+|\mathbf{k}|^2}. \tag{19.41}$$

In this description, κ tends to a well defined limit as $\nu \to 0$. However, note that the limit depends on whether $|\mathbf{k}| = 0$ or not. In particular, $\kappa(0,\mathbf{r},\mathbf{r}') = \epsilon_1$ for all ν. However, for $|\mathbf{k}| > 0$, $\kappa(0,\mathbf{r},\mathbf{r}')$ tends to ϵ_0 as ν tends to zero.

The dependence on the length scale λ on the distance from a hydrophobic surface was discussed in [371]. In terms of $\nu(\mathbf{r},\mathbf{r}') = 1/\lambda$ it is given by

$$\nu(\mathbf{r},\mathbf{r}') = \frac{1}{\lambda_b}\prod_j \left(1 - e^{-(|\mathbf{r}-\mathbf{r}_j|+|\mathbf{r}'-\mathbf{r}_j|)/\lambda_b}\right) \tag{19.42}$$

where λ_b denotes the bulk coordination length and hydrophobic entities are located at the points $\mathbf{r}_j$.

This can be interpreted in the following ways. First of all, it says that $\nu(\mathbf{r},\mathbf{r}') \approx 1/\lambda_b$ whenever either $\mathbf{r}$ or $\mathbf{r}'$ is far from the hydrophobes. In particular, if there are no hydrophobes at all (the bulk case), we have $\nu(\mathbf{r},\mathbf{r}') = 1/\lambda_b$.

When both $\mathbf{r}$ and $\mathbf{r}'$ are near a hydrophobe, say $\mathbf{r}_j$, then $e^{-(|\mathbf{r}-\mathbf{r}_j|+|\mathbf{r}'-\mathbf{r}_j|)/\lambda_b} \approx 1$. This means that $\nu(\mathbf{r},\mathbf{r}') \approx 0$ for such such $\mathbf{r},\mathbf{r}'$. In such a case, we find $\kappa(\mathbf{k},\mathbf{r},\mathbf{r}') \approx \epsilon_0$.

This approach does not explicitly involve a definition of the interior or boundary of the protein. This could be done using the function ν, for example by defining

$$\Omega = \left\{\mathbf{r} \;\middle|\; \nu(\mathbf{r},\mathbf{r}) < \tau\right\} \tag{19.43}$$

for some fixed tolerance $\tau > 0$, e.g., $\tau = 0.1$. But we do not explicitly need an expression for such a domain, and in particular there is no need to provide a mesh that is senstitive to the boundary $\partial\Omega$. This simplified representation of the effect of the protein in modulating the dielectric may compenstate for the fact that the resulting model is more complex.

19.7 Solving the Poisson-Debye equation

The Poisson-Debye equation (19.38) involves the Fourier Integral Operator $\mathcal{P}$ defined by (19.39). Clearly, the symbol $\kappa(\mathbf{k}, \mathbf{r}, \mathbf{r}') \geq \epsilon_0$ for all $\mathbf{r}, \mathbf{r}', \mathbf{k}$ (recall that $\epsilon_1 > \epsilon_0$). Similarly, the symbol of $\mathcal{P}$ is bounded above by a fixed constant for all $\mathbf{r}, \mathbf{r}', \mathbf{k}$. Thus a standard variational theory gives existence, uniqueness and stability for the Poisson-Debye equation (19.38).

19.7.1 Numerical methods for FIO's

Curvlets have been applied to FIO's [70] but efficiency seems to require smooth coefficients.

In our case,

$$\kappa(\mathbf{k}, \mathbf{r}, \mathbf{r}') = \epsilon_0 + \frac{\epsilon_1 - \epsilon_0}{1 + (|\mathbf{k}|/\nu(\mathbf{r}, \mathbf{r}'))^2} \tag{19.44}$$

is homogeneous in $|\mathbf{k}|/\nu$ and thus singular when $\nu \to 0$.

Therefore, direct application does not provide a sparse representation via curvelets.

19.8 Modeling DNA

The environment of DNA has been modeled as a dielectric cylinder with a given surface charge density immersed in an electrolite [314].

Chapter 20

Statistical mechanics

Statistical mechanics provides a way to include thermal fluctations in energetic models in a way that satisfies the laws of thermodynamics (cf. Section 2.7) [112, 117, 166, 216, 280, 315, 322, 328, 336, 342, 355, 356, 368, 401, 402, 406]. More precisely, suppose that we have a model U of the potential energy of a system that is characterized by the positions of n entities $\mathbf{r}_1, \ldots, \mathbf{r}_n$. The total energy of the system is a combination of the potential energy $U(\mathbf{r}_1, \ldots, \mathbf{r}_n)$ and the kinetic energy. For a molecular system with masses m_i, the kinetic energy of the i-th molecule is $m_i|v_i(t)|^2$, where

$$v_i(t) = \frac{\partial \mathbf{r}_i}{\partial t} \tag{20.1}$$

denotes the velocity of the i-th mass.

The temperature of a molecular system can be defined as

$$T(t) := \frac{1}{Nk_B} \sum_{i=1}^{N} m_i |v_i(t)|^2, \tag{20.2}$$

where k_B is Boltzmann's constant. Thus the temperature is just a suitable average of the kinetic energy. If temperature is in degrees Kelvin, velocities are measured in Ångstroms per picosecond, and masses in atomic mass units, then $k_B \approx 0.831$, that is, of order unity.

Thermodynamics views a system as a whole in which the individual states, e.g., the positions of the particles $\mathbf{r}_i(t)$, lose meaning. The Helmholtz (free) energy A is a thermodynamic potential, which measures the useful work obtainable from a closed thermodynamic system at a constant temperature and volume. It is often expressed as

$$A = E - Sk_BT, \tag{20.3}$$

where T is temperature, E is the 'internal energy' of the system (often called 'chemical energy'), and S is the entropy. The entropy S provides a measure of the degrees of freedom of the system. The energy E can be expressed as the expected potential energy

$$E = \frac{1}{Z} \int_{\Omega} U(\mathbf{r}_1, \ldots, \mathbf{r}_n) e^{-\beta U(\mathbf{r}_1, \ldots, \mathbf{r}_n)} \, d\mathbf{r}_1 \cdots d\mathbf{r}_n, \tag{20.4}$$

where $\beta = 1/k_B T$ and Z is a dimensionless number called the **partition function** for the system:

$$Z = \int_\Omega e^{-\beta U(\mathbf{r}_1,\ldots,\mathbf{r}_n)}\, d\mathbf{r}_1 \cdots d\mathbf{r}_n. \tag{20.5}$$

Here Ω denotes the state space, that is, the set of allowable positions $\mathbf{r}_i$. Thus Z is simply a normalization factor so that the quantity

$$P(\mathbf{r}_1, \ldots, \mathbf{r}_n) = \frac{1}{Z} e^{-\beta U(\mathbf{r}_1,\ldots,\mathbf{r}_n)} \tag{20.6}$$

forms a probability distribution over the state space; it is only a *function* of β, although it is also a function of the thermodynamic system, which might change due to exernal influences.

With this notation, $E = \int_\Omega U\, dP$. The probability distribution P in (20.5) is called the **Boltzmann distribution**. It says that lower energy states (smaller values of U) are more likely. The interpretation of E is that in a thermalized system all states are possible, with probability P; the lowest energy U has more influence but is not the only contributor. As $T \to 0$, P would tend to a discrete distribution with point masses at the states with lowest energy configurations $U_{\min}$, and in this limit $A(0) = \lim_{T\to 0} A(T) = U_{\min}$.

We can rationalize the exponential behavior of the Boltzmann distribution by considering the combination of two systems. If we take two systems with potential energy U^1 and U^2 respectively and combine them, we would expect the probability of states of the combined system to be the product of probabilities of the individual states. The only function f satisfying $f(U^1) + f(U^2) = f(U^1)f(U^2)$ is the exponential function $f(x) = e^{cx}$.

By considering the limit of zero temperature, the following fundamental identity can be derived:

$$A = -\frac{1}{\beta} \log Z = -k_B T \log \left(\int_\Omega e^{-\beta U(\mathbf{r}_1,\ldots,\mathbf{r}_n)}\, d\mathbf{r}_1 \cdots d\mathbf{r}_n \right). \tag{20.7}$$

Formula (20.7) reveals a fundamental connection between thermodynamics and statistical mechanics. We can use (20.7) to express the entropy as

$$S = \frac{E - A}{k_B T} = \left(\frac{1}{k_B T} \int_\Omega U\, dP \right) - \log Z, \tag{20.8}$$

in view of (20.3) and (20.4).

20.1 Example

None of these formulæ are computationally tractable in general, but they can be used to give some insight. Suppose we have a simple system with a finite number of states $R^1, \ldots, R^N$, such that the energies of the individual states are close:

$$U(R^j) = \overline{U} + \epsilon_j, \tag{20.9}$$

where the ϵ_j's are small, and $\overline{U}$ is the mean:

$$\sum_{j=1}^{n} \epsilon_j = 0. \tag{20.10}$$

Here the states R^j denote particular positions $\mathbf{r}_1^j, \ldots, \mathbf{r}_n^j$. With this notation,

$$Z = \sum_{j=1}^{N} e^{-\beta U(R^j)} = \sum_{j=1}^{N} e^{-\beta(\overline{U}+\epsilon_j)}. \tag{20.11}$$

We can use Taylor's theorem to approximate

$$e^{-\beta(\overline{U}+\epsilon)} \approx e^{-\beta(\overline{U})} \left(1 - \beta\epsilon + \mathcal{O}\left((\beta\epsilon)^2\right)\right). \tag{20.12}$$

Thus (20.11) yields

$$Z \approx \sum_{j=1}^{N} e^{-\beta\overline{U}}(1-\beta\epsilon_j) = e^{-\beta\overline{U}}(N - \beta\sum_{j=1}^{N}\epsilon_j) = Ne^{-\beta\overline{U}}, \tag{20.13}$$

by (20.10). Therefore

$$-\beta A = \log Z \approx -\beta\overline{U} + \log N. \tag{20.14}$$

Thus

$$A \approx \overline{U} - k_B T \log N. \tag{20.15}$$

We can relate this to the internal energy (20.4) using (20.12) and (20.13):

$$\begin{aligned} E =& \frac{1}{Z}\sum_{j=1}^{N} U(R^j) e^{-\beta U(R^j)} \approx \frac{e^{\beta\overline{U}}}{N}\sum_{j=1}^{n}(\overline{U}+\epsilon_j)e^{-\beta\overline{U}}(1-\beta\epsilon_j) \\ =& \frac{1}{N}\sum_{j=1}^{N}(\overline{U}+\epsilon_j)(1-\beta\epsilon_j) = \frac{1}{N}\sum_{j=1}^{N}(\overline{U}+\epsilon_j-\beta\epsilon_j\overline{U}-\beta\epsilon_j^2) \\ =& \overline{U} - \frac{\beta}{N}\sum_{j=1}^{N}\epsilon_j^2 \approx \overline{U}, \end{aligned} \tag{20.16}$$

where we used (20.10) at the penultimate step. Combining (20.16) with (20.14), we find

$$A \approx E - k_B T \log N, \tag{20.17}$$

so that

$$S \approx \log N. \tag{20.18}$$

Molecular systems can have a very large N, but the logarithm renders the energy expression manageable. The log (all of these are natural logarithms) of Avogadro's number is less than 55.

Chapter 21

Water structure

Each water molecule can potentially form hydrogen bonds with four other molecules, and typicaly does so when frozen, when water ice takes the form of a perfect lattice with all possible hydrogen bonds satisfied. However, the exact bonding structure of liquid water is still being studied and debated [1, 423, 382].

21.1 Understanding dielectrics

Dipole molecules can moderate charge. Induced dipole interactions lead to van der Waals forces.

21.2 Tetrahedral structure of water

Water is really a quadrupole, as indicated in Figure 21.3. The two positive charge regions provided by the hydrogens are matched by two negatively chaged regions oriented in a plane perpendicular to the plane of the water molecule. These negative lobes have been referred to variously as rabbit ears [238] and squirrel ears [276]. The deviation from a spherical electron density is actually quite small, yet distinct enough to localize hydrogen bonding.

Recent research has suggested that water is typically involved in only about half of its possible hydrogen bonds [423, 382].

21.3 Structural water in proteins

There is growing evidence that water plays some sort of structural role in proteins.

The frequent appearance of like-charged sidechain pairs has been noted previously.[269, 411] The stability of such repellent pairs comes from their interaction with stable clusters of water molecules.[269] Such favorable close approach of charged groups in water, due to polarization of the intervening water, has been demonstrated by ab initio quantum mechanical calculations and by observation of proteins in the PDB.[411] The role of backbone solvation and electrostatics in generating preferred peptide backbone conformations has recently been studied.[18] The effect of

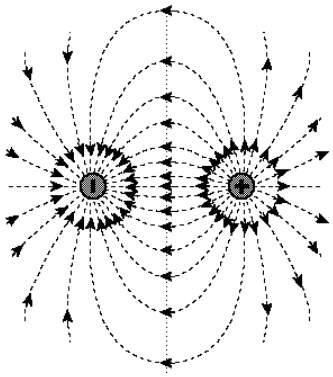


Figure 21.1: Single charge (+2) at zero and dipoles at 2,3,...,10 of width 0.4 and strength 1. Sharp peaks on the graph are located at the positive charge of the dipole, and the valleys are at the negative charge.

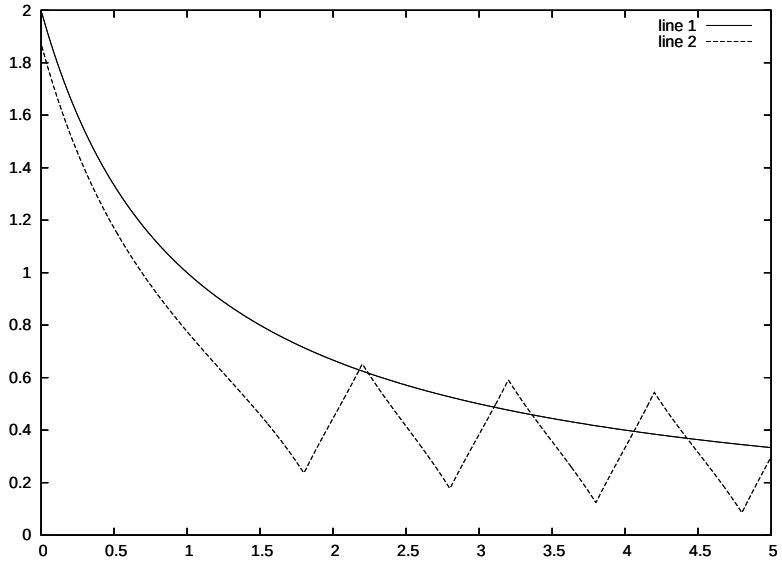


Figure 21.2:

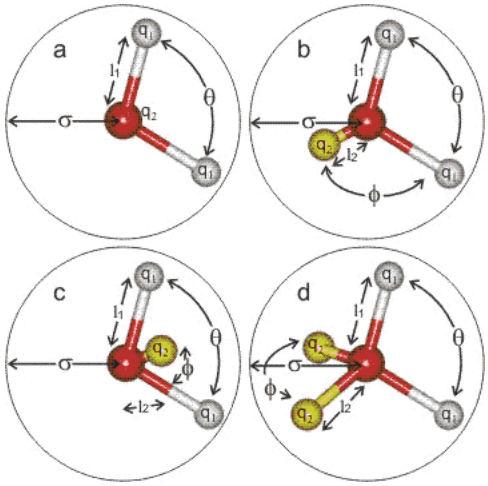


Figure 21.3: In Tip5P, see (d), $q_1 = -q_2 = 0.241$, $l_1 = 0.9572$, $l_2 = 0.70$, $\theta = 104.52$, $\phi = 109.47$.

neighboring residues on backbone conformation has been attributed to peptide backbone solvation and shielding from water of peptide groups by adjacent sidechains.[19]

21.4 Polarizable water models

Standard water models, such as Tip5P [253], do not allow for the polarizability of water ($\alpha \approx 1.47\text{Å}^3$). More complex models incorporate this explicitly [62, 435]. Apparently, the water dipole can be significantly enhanced by the polarization inherent in a condensed phase [171]. For a polarizability of $\alpha = 1\text{Å}^3$, a charge density of 0.2084 q_e per Å^2 would result in a change of one Debye. The dipole moment of an isolated water molecule is about 1.9 Debye and increases to 2.5 Debye in the condensed phase [171]. This would be consistent with a typical charge density of 0.1 q_e per Å^2.

21.5 A two-D water model

21.6 Two waters

Recent research has suggested that water can adopt (at least) two different phases [99, 206, 250].

Graph theory has been used to model interactions in water [236].

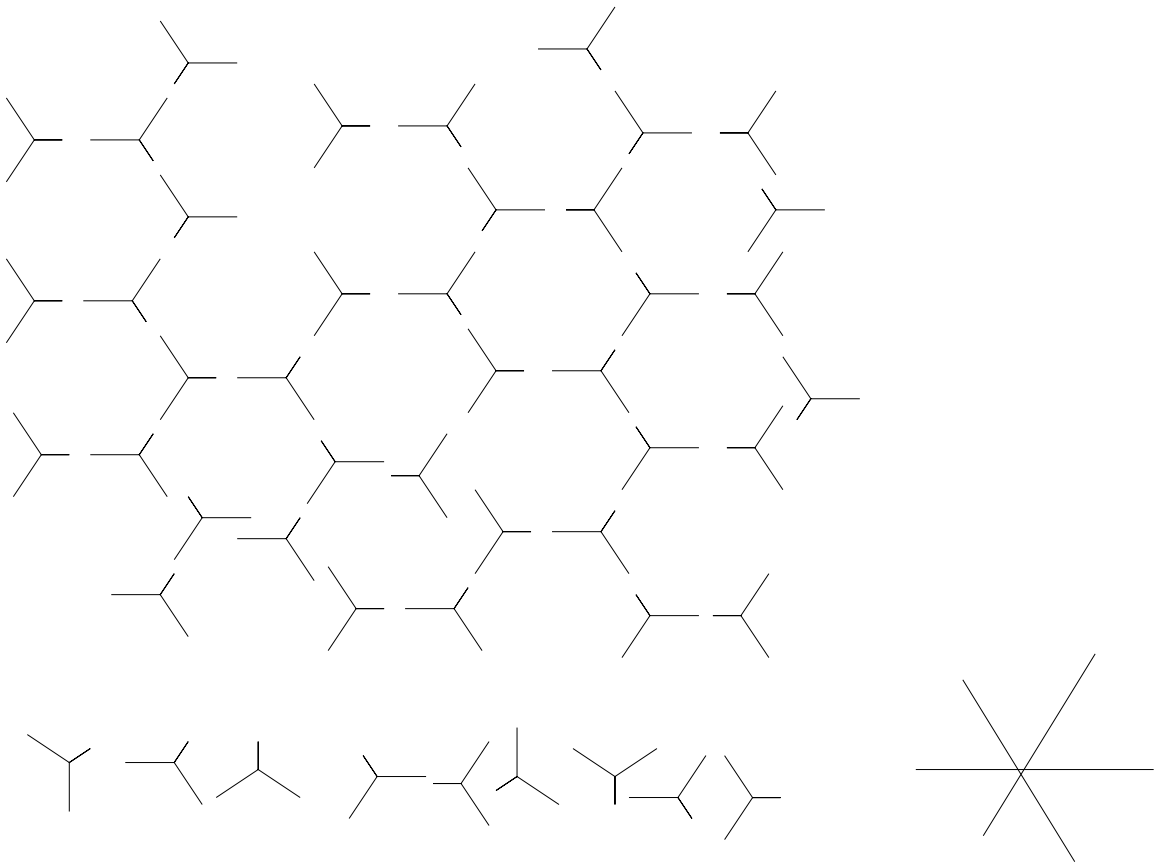


Figure 21.4: A two-D water model.

Chapter 22

Disorder in Protein Structure

Disorder is a common feature of proteins [49, 83, 100, 168, 195, 196, 205].

It can range from a small amount of flexibility to complete disorder. It has been argued that complete disorder is the common case for the vast majority of possible amino acid sequences [97]. Thus biologically relevant proteins represent a highly selective sampling.

One example of disorder is exhibited by prions [131].

The review [49] lists several programs that predict disorder based on protein sequence.

Chapter 23

Geckos' feet

It has recently been discovered that the feet of geckos are able to attach to surfaces by a type of van der Waals interaction [15]. It has also been determined that water plays a significant role in the effectiveness of binding [193]. Artificial 'feet' have been constructed with similar, or better, binding properties [164, 438].

Chapter 24

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 24.1, 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

Chapter 19

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	[241]
London dispersion	1930	
alpha/beta structure	1936	
hydrophobic force	1954–present	[95]
cation-π interaction	1996	[101]
dehydron	2003	[137, 126, 139]

Table 24.1: A brief history of the development of understanding of the principle bonds that are significant for protein structure and interaction.

Chapter 18

Chapter 21

Chapter 17

Chapter 10

Chapter 23

Chapter 25

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 **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 **moity** is a portion of a whole, usually with a defined property or structure.

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 **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.

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] Noam Agmon. Tetrahedral displacement: The molecular mechanism behind the Debye relaxation in water. *Journal of Physical Chemistry*, 100(1):1072–1080, 1996.

[2] C. J. Allègre, G. Manhès, and Christa Göpel. The age of the earth. *Geochimica et Cosmochimica Acta*, 59(8):1445–1456, 1995.

[3] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford Science Publications, Clarendon Press, Oxford, 1989.

[4] 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.

[5] 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.

[6] 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.

[7] G. Apic, J. Gough, and S. A. Teichmann. An insight into domain combinations. *Bioinformatics*, 17 (Suppl.1):S83–89, 2001.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] K. Atteson. The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25:251–278, Jun 1999. 10.1007/PL00008277.

[13] Kevin Atteson. *The performance of neighbor-joining algorithms of phylogeny reconstruction*, pages 101–110. LNCS Volume 1276. Springer, 1997. 10.1007/BFb0045077.

[14] R. Aurora and G. D. Rose. Helix capping. *Protein Science*, 7(1):21–38, 1998.

[15] Kellar Autumn, Metin Sitti, Yiching A. Liang, Anne M. Peattie, Wendy R. Hansen, Simon Sponberg, Thomas W. Kenny, Ronald Fearing, Jacob N. Israelachvili, and Robert J. Full. Evidence for van der Waals adhesion in gecko setae. *Proceedings of the National Academy of Sciences, USA*, 99(19):12252–12256, 2002.

[16] 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.

[17] 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.

[18] Franc Avbelj and Robert L. Baldwin. Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of phi. *Proceedings of the National Academy of Sciences, USA*, 100(10):5742–5747, 2003.

[19] Franc Avbelj and Robert L. Baldwin. Origin of the neighboring residue effect on peptide backbone conformation. *Proceedings of the National Academy of Sciences, USA*, 101(30):10967–10972, 2004.

[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] 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.

[25] 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.

[26] 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.

[27] 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.

[28] E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, 44:97–179, 1984.

[29] R. L. Baldwin. In search of the energetic role of peptide hydrogen bonds. *Journal of Biological Chemistry*, 278(20):17581–17588, 2003.

[30] R.L. Baldwin. Energetics of protein folding. *Journal of molecular biology*, 371(2):283–301, 2007.

[31] Robert L. Baldwin. Protein folding: Making a network of hydrophobic clusters. *Science*, 295:1657–1658, 2002.

[32] 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.

[33] Hans-Jurgen Bandelt and Andreas Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7:309–343, Sep 1986.

[34] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[35] A. L. Barabasi. *Linked: The New Science of Networks*. Perseus, New York, 2002.

[36] Jean-Pierre Barthelemy and Alain Guenoche. *Trees and Proximity Representations*. John Wiley & Sons, New York, 1991.

[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] Arturo Becerra, Luis Delaye, Sara Islas, and Antonio Lazcano. The very early stages of biological evolution and the nature of the last common ancestor of the three major cell domains. *Annual Review of Ecology, Evolution, and Systematics*, 38(1):361–379, 2007.

[42] Arieh Y. Ben-Naim. *Hydrophobic Interactions*. Springer, 1980.

[43] 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.

[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] 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.

[46] A. A. Bogan and K. S. Thorn. Anatomy of hot spots in protein interfaces. *Journal of Molecular Biology*, 280:1–9, 1998.

[47] A. A. Bogan and K. S. J. 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] Clay Bracken, Lilia M. Iakoucheva, Pedro R. Romero, and A. Keith Dunker. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.*, 14(5):570–576, 2004.

[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] 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.

[52] 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.

[53] William Fuller Brown, Jr. Dielectrics. In *Handbuch der Physik*, volume 14, pages 1–154. Springer, 1956.

[54] David Bryant. On the uniqueness of the selection criterion in neighbor-joining. *Journal of Classification*, 22:3–15, Jun 2005. 10.1007/s00357-005-0003-x.

[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] Christian J. Burnham and Sotiris S. Xantheas. Development of transferable interaction models for water. III. Reparametrization of an all-atom polarizable rigid model (TTM2–R) from first principles. *Journal of Chemical Physics*, 116(4):1500–1510, 2002.

[63] A. Buzzell. Action of urea on Tobacco Mosaic Virus II. The bonds between protein subunits. *Biophysical Journal*, 2(2P1):223–233, 1962.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] Eric Cancès, Claude Le Bris, and Yvon Maday. *Mèthodes mathèmatiques en chimie quantique.* Springer, 2006.

[69] Eric Cancès, M. Defranceschi, W. Kutzelnigg, Claude Le Bris, and Yvon Maday. Computational quantum chemistry: a primer. In Ph. Ciarlet and C. Le Bris, editors, *Handbook of numerical analysis. Volume X: special volume: computational chemistry*, pages 3–270. Elsevier, 2003.

[70] E. Candès and L. Demanet. The curvelet representation of wave propagators is optimally sparse. *Comm. Pure Appl. Math.*, 58(11):1472–1528, 2005.

[71] Pinak Chakrabarti and Joöl 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] Jianping Chen, Xi Zhang, and Ariel Fernández. Molecular basis for specificity in the druggable kinome: sequence-based analysis. *Bioinformatics*, 23(5):563–572, 2007.

[74] Dmitry A. Cherepanov. Force oscillations and dielectric overscreening of interfacial water. *Physical Review Letters*, 93:266104, 2004.

[75] Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283–1287, 2006.

[76] T. Clackson and J. A. Wells. A hot spot of binding energy in a hormone-receptor interface. *Science*, 267:383–386, 1995.

[77] C. Colovos and T. O. Yeates. Verification of protein structures: Patterns of nonbonded atomic interactions. *Protein Science*, 2(9):1511–1519, 1993.

[78] 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.

[79] David B. Cook, editor. *Handbook of Computational Quantum Chemistry*. Oxford Univ. Press, 1998.

[80] François-Xavier Coudert, Rodolphe Vuilleumiear, and Anne Boutin. Dipole moment, hydrogen bonding and IR spectrum of confined water. *ChemPhysChem*, 7(12):2464–2467, 2006.

[81] Christopher J. Cramer. *Essentials of Computational Chemistry*. Wiley, second edition, 2004.

[82] Thomas E. Creighton. *Proteins: Structures and molecular properties*. W. H. Freeman, 1993.

[83] Alejandro Crespo and Ariel Fernández. Induced disorder in protein–ligand complexes as a drug-design strategy. *Molecular Pharmaceutics*, 2008.

[84] Peter B. Crowley and Adel Golovin. Cation-π interactions in protein-protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 59:231–239, 2005.

[85] Matthew Davies, Christopher Toseland, David Moss, and Darren Flower. Benchmarking pK_a prediction. *BMC Biochemistry*, 7(1):18, 2006.

[86] 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.

[87] 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.

[88] 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.

[89] P. Debye. *Polar Molecules*. Dover, New York, 1945.

[90] 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.

[91] George D. Demetri. Structural reengineering of imatinib to decrease cardiac risk in cancer therapy. *The Journal of Clinical Investigation*, 117(12):3650–3653, 2007.

[92] 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.

[93] Cyril Deremble and Richard Lavery. Macromolecular recognition. *Current Opinion in Structural Biology*, 15(2):171–175, 2005.

[94] Florin Despa and R. Stephen Berry. The origin of long range attraction between hydrophobes in water. *Biophysical Journal*, 92:373–378, 2007.

[95] Florin Despa, Ariel Fernández, and R. Stephen Berry. Dielectric modulation of biological water. *Physical Review Letters*, 93:269901, 2004.

[96] Richard Desper and Olivier Gascuel. The minimum evolution distance-based approach to phylogenetic inference. In Olivier Gascuel, editor, *Mathematics of evolution and phylogeny*, pages 1–32. World Scientific, 2005.

[97] Christopher M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[98] Matthew Dobson and Mitchell Luskin. Iterative solution of the quasicontinuum equilibrium equations with continuation. *Journal of Scientific Computing*, 37(1):19–41, 2008.

[99] S. L. Dong, Y. Wang, A. I. Kolesnikov, and J. C. Li. Weakened hydrogen bond interactions in the high pressure phase of ice: Ice II. *Journal of Chemical Physics*, 109(1):235–240, 1998.

[100] Zsuzsanna Dosztanyi, Veronika Csizmok, Peter Tompa, and Istvan Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–3434, 2005.

[101] 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.

[102] John Doyle. Beyond the sperical cow. *Nature*, 411:151–152, May 2001.

[103] 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.

[104] Johannes J. Duistermaat. *Fourier Integral Operators*. Springer, Berlin, 1996.

[105] 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.

[106] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1999.

[107] Sean R. Eddy. Profile hidden Markov models. *Bioinformatics*, 14(9):755–763, 1998.

[108] Kord Eickmeyer, Peter Huggins, Lior Pachter, and Ruriko Yoshida. On the optimality of the neighbor-joining algorithm. *Algorithms for Molecular Biology*, 3(1):5, 2008.

[109] 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.

[110] A. Elgsaeter, B. T. Stokke, A. Mikkelsen, and D. Branton. The molecular basis of erythrocyte shape. *Science*, 234:1217–1223, 1986.

[111] Isaac Elias and Jens Lagergren. *Fast Neighbor Joining*, pages 1263–1274. LNCS Volume 3580. Springer, 2005. 10.1007/11523468_102.

[112] Jerald L. Ericksen. *Introduction to the Thermodynamics of Solids*. Springer, 1998.

[113] 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.

[114] 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.

[115] M. Fandrich, M. A. Fletcher, and C. M. Dobson. Amyloid fibrils from muscle myoglobin. *Nature*, 410(6825):165–166, Mar 2001.

[116] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, second edition, 2003.

[117] Enrico Fermi. *Thermodynamics*. Dover, 1956.

[118] Ariel Fernández. Cooperative walks in a cubic lattice: Protein folding as a many-body problem. *J. Chem. Phys.*, 115:7293–7297, 2001.

[119] Ariel Fernández. Intramolecular modulation of electric fields in folding proteins. *Phys. Lett. A*, 299:217–220, 2002.

[120] Ariel Fernández. Buffering the entropic cost of hydrophobic collapse in folding proteins. *Journal of Chemical Physics*, 121:11501–11502, 2004.

[121] Ariel Fernández. Keeping dry and crossing membranes. *Nature Biotechnology*, 22:1081–1084, 2004.

[122] Ariel Fernández. Direct nanoscale dehydration of hydrogen bonds. *Journal of Physics D: Applied Physics*, 38:2928–2932, 2005.

[123] Ariel Fernández. What factor drives the fibrillogenic association of beta-sheets? *FEBS Letters*, 579:6635–6640, 2005.

[124] Ariel Fernández. *Transformative concepts for drug design*. Springer, 2010.

[125] 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.

[126] 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.

[127] 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.

[128] Ariel Fernández and M. Boland. Solvent environment conducive to protein aggregation. *FEBS Letters*, 529:298–303, 2002.

[129] 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.

[130] 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.

[131] 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.

[132] Ariel Fernández and Sridhar Maddipati. A priori inference of cross reactivity for drug-targeted kinases. *Journal of Medicinal Chemistry*, 49(11):3092–3100, 2006.

[133] 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.

[134] 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.

[135] 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.

[136] Ariel Fernández, Angela Sanguino, Zhenghong Peng, Eylem Ozturk, Jianpeng Ma, David Maxwell, Aleksander Shavrin, Juri Gelovani, William Bornmann, and Gabriel Lopez-Berestein. Molecularly engineered specificity in anti-cancer activity: Cell-line and molecular assays translating a structural discriminator. 2006.

[137] 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.

[138] Ariel Fernández and L. Ridgway Scott. Adherence of packing defects in soluble proteins. *Physical Review Letters*, 91:18102(4), 2003.

[139] Ariel Fernández and L. Ridgway Scott. Dehydron: a structurally encoded signal for protein interaction. *Biophysical Journal*, 85:1914–1928, 2003.

[140] Ariel Fernández and L. Ridgway Scott. Under-wrapped soluble proteins as signals triggering membrane morphology. *Journal of Chemical Physics*, 119(13):6911–6915, 2003.

[141] Ariel Fernández and L. Ridgway Scott. Modulating drug impact by wrapping target proteins. *Expert Opinion on Drug Discovery*, 2:249–259, 2007.

[142] 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.

[143] 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.

[144] 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.

[145] 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.

[146] 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.

[147] Herman Feshbach. Unified theory of nuclear reactions. *Annals of Physics*, 5:357–390, 1958.

[148] 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.

[149] 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.

[150] 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.

[151] 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.

[152] 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.

[153] Rosalind Franklin and R. G. Gosling. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature*, 172:156–157, 1953.

[154] Felix Franks. *Water: a matrix for life*. Royal Society of Chemistry, 2000.

[155] J.S. Franzen and R.E. Stephens. The effect of a dipolar solvent system on interamide hydrogen bonds. *Biochemistry*, 2(6):1321–1327, 1963.

[156] 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.

[157] 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.

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] O Gascuel. A note on Sattath and Tversky's, Saitou and Nei's, and Studier and Keppler's algorithms for inferring phylogenies from evolutionary distances. *Mol Biol Evol*, 11(6):961–963, 1994.

[163] Olivier Gascuel and Mike Steel. Neighbor-joining revealed. *Mol Biol Evol*, 23(11):1997–2000, 2006.

[164] Liehui Ge, Sunny Sethi, Lijie Ci, Pulickel M. Ajayan, and Ali Dhinojwala. Carbon nanotube-based synthetic gecko tapes. *Proceedings of the National Academy of Sciences*, 104(26):10792–10795, 2007.

[165] 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.

[166] Robin Giles. *Mathematical foundations of thermodynamics*. Macmillan, 1964.

[167] 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.

[168] G. Goh, A.K. Dunker, and V. Uversky. Protein intrinsic disorder toolbox for comparative analysis of viral proteins. *BMC genomics*, 9(Suppl 2):S4, 2008.

[169] S. A. Goudschmidt and G. H. Uhlenbeck. Spinning electrons and the structure of spectra. *Nature*, 117(2938):264–265, 1926.

[170] 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.

[171] 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.

[172] 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.

[173] Nick V. Grishin. Fold change in evolution of protein structures. *Journal of Structural Biology*, 134(2-3):167 – 185, 2001.

[174] 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.

[175] 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.

[176] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, January 1997.

[177] George Hagedorn. High order corrections to the time-independent Born-Oppenheimer approximation. I. Smooth potentials. *Ann. Inst. Henri Poincaré*, 47:1–16, 1987.

[178] George Hagedorn. High order corrections to the time-independent Born-Oppenheimer approximation. II. Diatomic Coulomb systems. *Comm. Math. Phys.*, 116:23–44, 1988.

[179] Bertil Halle. Protein hydration dynamics in solution: a critical survey. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1448):1207–1224, 2004.

[180] 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.

[181] 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.

[182] Yehouda Harpaz, Mark Gerstein, and Cyrus Chothia. Volume changes on protein folding. *Structure*, 2:641–649, 1994.

[183] J. P. Helfrich. Dynamic laser light scattering technology for the molecular weight and hydrodynamic radius characterization of proteins. *Pharmaceutical Laboratory*, 1:34–40, 1998.

[184] 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.

[185] 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.

[186] Nora E. Hill, Worth E. Vaughan, A. H. Price, and Mansel Davies. *Dielectric properties and molecular behaviour*. van Nostrand, London, 1969.

[187] U. Hobohm, M. Scharf, R. Schneider, and C. Sander. Selection of representative protein data sets. *Protein Science*, 1(3):409–417, 1992.

[188] B. Honig and A. S. Yang. Free energy balance in protein folding. *Adv. Protein Chem.*, 46:27–58, 1995.

[189] 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.

[190] 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.

[191] Sven Hovmöller, Tuping Zhou, and Tomas Ohlson. Conformations of amino acids in proteins. *Acta Crystallographica Section D*, 58(5):768–776, 2002.

[192] 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.

[193] Gerrit Huber, Hubert Mantz, Ralph Spolenak, Klaus Mecke, Karin Jacobs, Stanislav N. Gorb, and Eduard Arzt. Evidence for capillarity contributions to gecko adhesion from single spatula nanomechanical measurements. *Proceedings of the National Academy of Sciences, USA*, 102(45):16293–16296, 2005.

[194] 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.

[195] L. M. Iakoucheva and K. A. Dunker. Order, disorder, and flexibility: prediction from protein sequence. *Structure*, 11:1316–1317, 2003.

[196] Lilia M. Iakoucheva, Predrag Radivojac, Celeste J. Brown, Timothy R. O'Connor, Jason G. Sikes, Zoran Obradovic, and A. Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucl. Acids Res.*, 32(3):1037–1049, 2004.

[197] Jacob Israelachvili. *Intermolecular and surface forces*. Academic Press, second edition, 1991.

[198] Jacob Israelachvili and Hakan Wennerstrom. Role of hydration and water structure in biological and colloidal interactions. *Nature*, 379:219–225, 1996.

[199] Joël Janin, S. Miller, and Cyrus Chothia. Surface, subunit interfaces and interior of oligomeric proteins. *Journal of Molecular Biology*, 204:155–164, 1988.

[200] 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.

[201] George A. Jeffrey. *An introduction to hydrogen bonds*. Oxford Science Publications, Clarendon Press, Oxford, 1997.

[202] George A. Jeffrey. Hydrogen-bonding: An update. *Crystallography Reviews*, 9:135–176, Jan 2003. 10.1080/08893110310001621754.

[203] 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.

[204] S. Ji. The linguistics of DNA: words, sentences, grammar, phonetics, and semantics. *Annals of the New York Academy of Sciences*, 870:411–417, 1999.

[205] Yumi Jin and Roland L. Dunbrack Jr. Assessment of disorder predictions in CASP6. *Proteins: Structure, Function, and Bioinformatics*, 61(S7):167–175, 2005.

[206] G. P. Johari. An estimate for the Gibbs energy of amorphous solid waters and differences between the low-density amorph and glassy water. *Journal of Chemical Physics*, 112:8573–8580, 2000.

[207] S. Jones, A. Marin, and J. M. Thornton. Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, 13:77–82, 2000.

[208] Susan Jones and Janet M. Thornton. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences, USA*, 93:13–20, 1996.

[209] Susan Jones and Janet M. Thornton. Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*, 272:121–132, 1997.

[210] Susan Jones and Janet M. Thornton. Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*, 272:133–143, 1997.

[211] 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.

[212] 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.

[213] 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.

[214] W. Kauzmann. Some factors in the interpretation of protein denaturation. In *Advances in Protein Chemistry*, volume 14, pages 1–63. Academic Press, 1959.

[215] F.N. Keutsch, J.D. Cruzan, and R.J. Saykally. The water trimer. *Chemical reviews*, 103(7):2533–2578, 2003.

[216] A. I. Khinchin. *Mathematical Foundations of Statistical Mechanics*. Dover, 1949.

[217] 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.

[218] 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.

[219] Sangtae Kim and Seppo J. Karrila. *Microhydrodynamics : Principles and Selected Applications*. Dover Publications, 2005.

[220] K. C. Klauer and J. D. Carroll. A mathematical programming approach to fitting general graphs. *Journal of Classification*, 6:247–270, Dec 1989. 10.1007/BF01908602.

[221] Karl Klauer. Ordinal network representation: Representing proximities by graphs. *Psychometrika*, 54:737–750, Sep 1989. 10.1007/BF02296406.

[222] M. Klein, A. Martinez, R. Seiler, and X. P. Wang. Born-Oppenheimer expansion for polyatomic molecules. *Comm. Math. Phys*, pages 607–639, 1992.

[223] 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.

[224] I.M. Klotz. Solvent water and protein behavior: View through a retroscope. *Protein Science*, 2(11), 1993.

[225] 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.

[226] 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.

[227] Peter A. Kollman and Leland C. Allen. The theory of the hydrogen bond. *Chemical Reviews*, 72(3):283–303, 1972.

[228] E. V. Koonin, Y. I. Wolf, and G. P. Karev. The structure of the protein universe and genome evolution. *Nature*, 420:218–223, 2002.

[229] 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.

[230] A. A. Kornyshev and A. Nitzan. Effect of overscreening on the localization of hydrated electrons. *Zeitschrift für Physikalische Chemie*, 215(6):701–715, 2001.

[231] 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.

[232] 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.

[233] Lawrence M. Krauss. *Fear of Physics*. Basic Books, 1994.

[234] S. Krishnaswamy and Michael G. Rossmann. Structural refinement and analysis of mengo virus. *Journal of Molecular Biology*, 211:803–844, 1990.

[235] 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.

[236] Jer-Lai Kuo, James V. Coe, Sherwin J. Singer, Yehuda B. Band, and Lars Ojamae. On the use of graph invariants for efficiently generating hydrogen bond topologies and predicting physical properties of water clusters and ice. *Journal of Chemical Physics*, 114(6):2527–2540, 2001.

[237] 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.

[238] M. Laing. No rabbit ears on water. *J. Chem. Ed.*, 64:124–128, 1987.

[239] L. D. Landau and E. M. Lifshitz. Fluid mechanics. *Course of theoretical physics*, 6, 1987.

[240] 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.

[241] 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.

[242] Christian Laurence and Michel Berthelot. Observations on the strength of hydrogen bonding. *Perspectives in Drug Discovery and Design*, 18:3960, 2000.

[243] 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.

[244] 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.

[245] 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.

[246] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol*, 104:59–107, 1976.

[247] Michael Levitt and Max F. Perutz. Aromatic rings act as hydrogen bond acceptors. *Journal of Molecular Biology*, 201:751–754, 1988.

[248] 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.

[249] 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.

[250] J. Li and D. K. Ross. Evidence for two kinds of hydrogen bond in ice. *Nature*, 365:327–329, 1993.

[251] Laurent Limozin and Erich Sackmann. Polymorphism of cross-linked actin networks in giant vesicles. *Physical Review Letters*, 89(16):168103, Oct 2002.

[252] C. R. Linder, B. M. E. Moret, L. Nakhleh, and T. Warnow. Network (reticulate) evolution: biology, models, and algorithms. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany A. Jung, and Teri E. Klein, editors, *Biocomputing 2004, Proceedings of the Pacific Symposium, Hawaii, USA, 6-10 January 2004*, pages –. World Scientific, 2004.

[253] 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.

[254] 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.

[255] 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.

[256] 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.

[257] 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.

[258] 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.

[259] 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.

[260] Loredana Lo Conte, Cyrus Chothia, and Joël Janin. The atomic structure of protein-protein recognition sites. *Mol. Biol.*, 285:2177–2198, 1999.

[261] 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.

[262] Per-Olov Löwdin. Expansion theorems for the total wave function and extended Hartree-Fock schemes. *Rev. Mod. Phys.*, 32(2):328–334, 1960.

[263] 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.

[264] 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.

[265] Jennifer C. Ma and Dennis A. Dougherty. The cation-π interaction. *Chemical Reviews*, 97(5):1303–1324, 1997.

[266] 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.

[267] S. Maddipati and A. Fernández. Feature-similarity protein classifier as a ligand engineering tool. *Biomolecular engineering*, 23(6):307–315, 2006.

[268] 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.

[269] A. Magalhaes, B. Maigret, J. Hoflack, J. A. N. F. Gomes, and H. A. Scheraga. Contribution of unusual arginine-arginine short-range interactions to stabilization and recognition in proteins. *J. Protein Chem.*, 13(2):195–215, 1994.

[270] 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.

[271] Vladimir Makarenkov. T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, 17(7):664–668, 2001.

[272] 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.

[273] 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.

[274] 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.

[275] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, 2002.

[276] R. B. Martin. Localized and spectroscopic orbitals: Squirrel ears on water. *J. Chem. Ed.*, 65:668–670, 1988.

[277] A. Martinez. Eigenvalues and resonances of polyatomic molecules in the Born-Oppenheimer approximation. In Erik Balslev, editor, *Schrödinger Operators, The Quantum Mechanical Many-Body Problem*, pages 145–152, Berlin, 1992. Springer-Verlag.

[278] 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.

[279] C. Matzler and U. Wegmuller. Dielectric properties of freshwater ice at microwave frequencies. *Journal of Physics D: Applied Physics*, 20(12):1623–1630, 1987.

[280] Barry McCoy and Tai Tsun Wu. *The Two-Dimensional Ising Model*. Harvard University Press, 1973.

[281] Peter McCullagh. Marginal likelihood for distance matrices. *Statistica Sinica*, 19:631–650, 2009.

[282] I. K. McDonald and Janet M. Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238:777–793, 1994.

[283] 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.

[284] 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.

[285] Carver Mead and Lynn Conway. *Introduction to VLSI Systems*. Addison-Wesley, 1979.

[286] 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.

[287] 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.

[288] Radu Mihaescu, Dan Levy, and Lior Pachter. Why neighbor-joining works. *Algorithmica*, 2008. 10.1007/s00453-007-9116-4.

[289] Radu Horia Mihaescu. *Distance Methods for Phylogeny Reconstruction*. PhD thesis, UC Berkeley, 1996.

[290] 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.

[291] C. Millot and A.J. Stone. Towards an accurate intermolecular potential for water. *Molecular Physics*, 77(3):439–462, 1992.

[292] 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.

[293] 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.

[294] 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.

[295] 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.

[296] 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.

[297] 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.

[298] J. Moro, J. V. Burke, and M. L. Overton. On the Lidskii-Lyusternik-Vishik perturbation theory for eigenvalues with arbitrary Jordan structure. *SIAM J. Matrix Anal. Appl.*, 18:793–817, 1997.

[299] 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.

[300] 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.

[301] 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.

[302] 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.

[303] 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.

[304] 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.

[305] 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.

[306] 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.

[307] Irene M. A. Nooren and Janet M. Thornton. Diversity of proteinprotein interactions. *EMBO Journal*, 22:3486–3492, 2003.

[308] 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.

[309] Marian Novotny and Gerard J. Kleywegt. A survey of left-handed helices in protein structures. *Journal of Molecular Biology*, 347(2):231 – 241, 2005.

[310] 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.

[311] Yanay Ofren and Burkhard Rost. Analysing six types of protein-protein interfaces. *Journal of Molecular Biology*, 335:377–387, 2003.

[312] 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.

[313] T. Ooi. Thermodynamics of protein folding: effects of hydration and electrostatic interactions. *Advan. Biophys.*, 30:105–154, 1994.

[314] D. Ouroushev. Electric field of dielectric cylinder with given surface charge density immersed in electrolite. *Journal of Physics A: Mathematical and General*, 31(16):3897–3902, 1998.

[315] David R. Owen. *A first course in the mathematical foundations of thermodynamics*. Springer, 1984.

[316] 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.

[317] 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.

[318] Samir Kumar Pal, Jorge Peon, and Ahmed H. Zewail. Biological water at the protein surface: Dynamical solvation probed directly with femtosecond resolution. *Proceedings of the National Academy of Sciences, USA*, 99(4):1763–1768, 2002.

[319] 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.

[320] Wolfgang Pauli. The connection between spin and statistics. *Phys. Rev.*, 58(8):716–722, Oct 1940.

[321] Wolfgang Pauli. Exclusion principle and quantum mechanics. In *Nobel Lectures, Physics 1942-1962*, pages 27–43, Amsterdam, 1964. Elsevier Publishing Company.

[322] Wolfgang Pauli. *Thermodynamics and the Kinetic Theory of Gases*. Dover, 2000.

[323] Linus Pauling. *Nature of the Chemical Bond*. Cornell Univ. Press, third edition, 1960.

[324] Linus Pauling. *General Chemistry*. Dover, 1970.

[325] Linus Pauling and E. Bright Wilson. *Introduction to Quantum Mechanics with Applications to Chemistry*. Dover, 1985.

[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] 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.

[328] J. K. Percus. *Kinetic Theory and Statistical Mechanics*. New York University Courant Institue of Mathemtatical Sciences, New York, 1970.

[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] Max Planck. *Treatise On Thermodynamics*. Dover, 2008.

[337] 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.

[338] 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.

[339] 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.

[340] 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.

[341] Paul Popelier. *Atoms in Molecules*. Prentice-Hall, 2000.

[342] Christopher J. Preston. *Gibbs States on Countable Sets*. Cambridge, 2008.

[343] S. B. Prusiner. Prions. *Proceedings of the National Academy of Sciences, USA*, 95(23):13363–13383, Nov 1998.

[344] Sandra Pruzansky, Amos Tversky, and J. Carroll. Spatial versus tree representations of proximity data. *Psychometrika*, 47:3–24, Mar 1982. 10.1007/BF02293848.

[345] 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.

[346] Sergei Radaev and Peter Sun. Recognition of immunoglobulins by Fcγ receptors. *Molecular Immunology*, 38(14):1073–1083, 2002.

[347] 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.

[348] J. J. Ramsden. Review of optical methods to probe protein adsorption at solid-liquid interfaces. *J. Stat. Phys.*, 73:853–877, 1993.

[349] 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.

[350] 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.

[351] Dagmar Ringe. What makes a binding site a binding site? *Curr. Opin. Struct. Biol.*, 5:825–829, 1995.

[352] G. G. Roberts, editor. *Langmuir-Blodgett Films*. Plenum Press, New York, 1990.

[353] 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.

[354] 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.

[355] David Ruell. *Thermodynamic formalism : the mathematical structures of classical equilibrium statistical mechanics*. Addison-Wesley, 1978.

[356] David Ruell. *Statistical Mechanics: rigorous results*. Addison-Wesley, 1989.

[357] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425, 1987.

[358] 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.

[359] 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.

[360] R. T. Sanderson. Electronegativity and bond energy. *Journal of the American Chemical Society*, 105:2259–2261, 1983.

[361] 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.

[362] Shmuel Sattath and Amos Tversky. Additive similarity trees. *Psychometrika*, 42:319–345, Sep 1977. 10.1007/BF02293654.

[363] Aleister J. Saunders, Gregory B. Young, and Gary J. Pielak. Polarity of disulfide bonds. *Protein Science*, 2(7):1183–1184, 1993.

[364] Lindsay Sawyer and Michael N. G. James. Carboxyl-carboxylate interactions in proteins. *Nature*, 295:79–80, Jan 1982. 10.1038/295079a0.

[365] John A. Schellman. Fifty years of solvent denaturation. *Biophysical Chemistry*, 96(2-3):91 – 101, 2002.

[366] 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.

[367] 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.

[368] Erwin Schrodinger. *Statistical Thermodynamics*. Dover, 1989.

[369] Laurent Schwartz. *Théorie des distributions*. Hermann, Paris, 1966.

[370] Eric Schwegler, Matt Challacombe, and Martin Head-Gordon. Linear scaling of the Fock matrix. II. Rigorous bounds on exchange integrals and incremental Fock build. *J. Chem. Phys.*, 106:9708–9717, 1997.

[371] 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.

[372] J. Seelig and A. Seelig. Lipid conformation in model membranes and biological membranes. *Q. Rev Biophys*, 13(1):19–61, Feb 1980.

[373] 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.

[374] 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.

[375] 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.

[376] H. Shull and G. G. Hall. Atomic units. *Nature*, 184:1559–1560, Nov 1959. 10.1038/1841559a0.

[377] 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.

[378] 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.

[379] John C. Slater. *Solid-State and Molecular Theory: A Scientific Biography.* John Wiley & Sons, New York, 1975.

[380] John C. Slater. *The Calculation of Molecular Orbitals.* John Wiley & Sons, New York, 1979.

[381] G. R. Smith and M. J. Sternberg. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.*, 12:28–35, 2002.

[382] 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.

[383] 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.

[384] B. J. Stapley and T. P. Creamer. A survey of left-handed polyproline II helices. *Protein Sci.*, 8:587–595, Mar 1999.

[385] Thomas Steiner. Competition of hydrogen-bond acceptors for the strong carboxyl donor. *Acta Crystallographica Section B*, 57(1):103–106, Feb 2001.

[386] 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.

[387] DF Stickle, LG Presta, KA Dill, and GD Rose. Hydrogen bonding in globular proteins. *Journal of molecular biology*, 226(4):1143–1159, 1992.

[388] A. J. Stone. Distributed multipole analysis, or how to describe a molecular charge distribution. *Chemical Physics Letters*, 83(2):233 – 239, 1981.

[389] A. J. Stone. *The theory of intermolecular forces*. Oxford University Press, USA, 1997.

[390] A. J. Stone and M. Alderton. Distributed multipole analysis: methods and applications. *Molecular Physics*, 56(5):1047–1064, 1985.

[391] Matthew C. Strain, Gustavo E. Scuseria, and Michael J. Frisch. Achieving linear scaling for the electronic quantum Coulomb problem. *Science*, 271:51–53, 1996.

[392] Mark P. Styczynski, Kyle L. Jensen, Isidore Rigoutsos, and Gregory Stephanopoulos. BLOSUM62 miscalculations improve search performance. *Nat Biotech*, 26:274–275, Mar 2008. 10.1038/nbt0308-274.

[393] Kulandavelu Subramanian, Srinivasakannan Lakshmi, Krishnamurthy Rajagopalan, Gertraud Koellner, and Thomas Steiner. Cooperative hydrogen bond cycles involving O—H$\cdots\pi$ and C—H$\cdots$O hydrogen bonds as found in a hydrated dialkyne. *Journal of Molecular Structure*, 384(2-3):121 – 126, 1996.

[394] 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.

[395] 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.

[396] D. I. Svergun, S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin, and G. Zaccai. Protein hydration in solution: Experimental observation by x-ray and neutron scattering. *Proceedings of the National Academy of Sciences, USA*, 95(5):2267–2272, 1998.

[397] Attila Szabo and Neil S. Ostlund, editors. *Modern Quantum Chemistry*. Dover, 1996.

[398] Charles Tanford. *Hydrophobic Effect*. Wiley, 1973.

[399] Charles Tanford and Jacqueline Reynolds. *Nature's Robots, a history of proteins*. Oxford Univ. Press, 2001.

[400] Robin Taylor and Olga Kennard. Hydrogen-bond geometry in organic crystals. *Accounts of Chemical Research*, 17(9):320–326, 1984.

[401] Colin J. Thompson. *Mathematical Statistical Mechanics*. Macmillan, 1972.

[402] Colin J. Thompson. *Classical Equilibrium Statistical Mechanics*. Oxfrd University Press, 1988.

[403] 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.

[404] 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.

[405] 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.

[406] Clifford Truesdell and S. Bharatha. *The concepts and logic of classical thermodynamics as a theory of heat engines, rigorously constructed upon the foundation laid by S. Carnot and F. Reech*. Springer, 1977.

[407] 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.

[408] 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.

[409] G. E. Uhlenbeck and L. S. Ornstein. On the theory of the brownian motion. *Phys. Rev.*, 36(5):823–841, Sep 1930.

[410] P.T. Van Duijnen and M. Swart. Molecular and atomic polarizabilities: Thole's model revisited. *J. Phys. Chem. A*, 102(14):2399–2407, 1998.

[411] 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.

[412] D. Vitkup, E. Melamud, J. Moult, and C. Sander. Completeness in structural genomics. *Nature Structural Biology*, 8:559–566, 2001.

[413] 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.

[414] D. Voet and J. G. Voet. *Biochemistry*. John Wiley & Sons, New York, 1990.

[415] 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.

[416] Timothy C. Wallstrom. Inequivalence between the Schrödinger equation and the Madelung hydrodynamic equations. *Physical Review A*, 49:1613–1617, 1994.

[417] 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.

[418] A. Warshel and A. Papazyan. Electrostatic effects in macromolecules: fundamental concepts and practical modeling. *Curr. Opin. Struct. Biol.*, 8:211–217, 1998.

[419] 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.

[420] Michael Waterman. *Introduction to Computational Biology*. Chapman & Hall/CRC Press, 1995.

[421] James D. Watson and Francis H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[422] 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.

[423] 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.

[424] Maurice Wilkins, A. R. Stokes, and H. R. Wilson. Molecular structure of deoxypentose nucleic acids. *Nature*, 171:738–740, 1953.

[425] 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.

[426] Stephen Willson. Unique reconstruction of tree-like phylogenetic networks from distances between leaves. *Bulletin of Mathematical Biology*, 68:919–944, May 2006. 10.1007/s11538-005-9044-x.

[427] 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.

[428] 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.

[429] 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.

[430] 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.

[431] 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.

[432] 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.

[433] Yuri I. Wolf, Igor B. Rogozin, Nick V. Grishin, and Eugene V. Koonin. Genome trees and the tree of life. *Trends in Genetics*, 18(9):472 – 479, 2002.

[434] 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.

[435] Yang Wu and Zhong-Zhi Yang. Atom-bond electronegativity equalization method fused into molecular mechanics. II. A seven-site fluctuating charge and flexible body water potential function for liquid water. *The Journal of Physical Chemistry A*, 108(37):7563–7576, 2004.

[436] Stefan Wuchty. Scale-free behavior in protein domain networks. *Molecular Biology and Evolution*, 18(9):1694–1702, Sep 2001.

[437] Dong Xu, Chung-Jung Tsai, and Ruth Nussinov. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Engineering*, 10(9):999–1012, 1997.

[438] H Yao, G. Della Rocca, P.R Guduru, and H Gao. Adhesion and sliding response of a biologically inspired fibrillar surface: experimental observations. *Journal of The Royal Society Interface*, 5(24):723–733, 2008.

[439] Wai-Ming Yau, William C. Wimley, Klaus Gawrisc, and Stephen H. White. The preference of tryptophan for membrane interfaces. *Biochemistry*, 37:14713–14718, 1998.

[440] Min Yi Shen and Karl F. Freed. Long time dynamics of Met-Enkephalin: Comparison of explicit and implicit solvent models. *Biophysical Journal*, 82(4):1791–1808, 2002.

[441] Anne D. Yoder and Ziheng Yang. Estimation of primate speciation dates using local molecular clocks. *Mol Biol Evol*, 17(7):1081–1090, 2000.

[442] Tsuyoshi Yokomizo, Masayoshi Nakasako, Toshimasa Yamazaki, Heizaburo Shindo, and Junich Higo. Hydrogen-bond patterns in the hydration structure of a protein. *Chemical Physics Letters*, 401:332–336, 2005.

[443] Natalya Yutin, Kira S. Makarova, Sergey L. Mekhedov, Yuri I. Wolf, and Eugene V. Koonin. The deep archaeal roots of eukaryotes. *Mol Biol Evol*, 25(8):1619–1630, 2008.

[444] 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.

[445] 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.

[446] 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.

[447] Ruhong Zhou, Xuhui Huang, Claudio J. Margulis, and Bruce J. Berne. Hydrophobic collapse in multidomain protein folding. *Science*, 305(5690):1605–1609, 2004.

[448] 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.

[449] 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.

[450] 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.

[451] 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.

[452] Matjaz Zorko and Ülo Langel. Cell-penetrating peptides: mechanism and kinetics of cargo delivery. *Advanced Drug Delivery Reviews*, 57(4):529–545, 2005.

Index