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. 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 [157] 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 for electron shell completion allowed rapid prediction of molecule formulation by simple counting [180]. Resonance theory (Section 14.1) describes general bonding patterns as a combination of simple bonds (e.g, single and double bonds) [179]. The discrete behavior of DNA elucidated by Crick, Franklin, Watson, Wilkins and others [233, 236, 84] initiated the molecular biology revolution. Our objective here is to provide an introduction to some basic properties of protein-ligand interactions with the hope of stimulating further study of the discrete nature of molecular interactions in biology.

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. 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 2.3.

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 [49]. 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 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 is not a summary of finished work. 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 [157], "And thus the period of exploration begins."

Chapter 2

Challenges of protein models

We present here 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 begin by describing some of the 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 building blocks of chemistry are atoms. They can be characterized by the number of electrons, protons and neutrons of which they are composed. The atoms of primary interest in protein biochemistry are listed in Table 2.1.

Some comments are in order about Table 2.1. First of all, the number of neutrons can vary; we have listed what is known as the dominant isotope. Neutrons add mass but not charge. Other isotopes are important in various contexts; a hydrogen atom with an extra neutron is called deuterium. Atoms occur naturally in different isotopic forms, and the atomic 'weight' (properly, the 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; the rest mass of an electron is less that 0.00055 atomic mass units. In the column 'variation' we give the difference between the atomic 'weight' (the mass in atomic units) and the mass of the standard isotope's protons and neutrons. For chlorine, the atoms with 18 and 19 neutrons are nearly equally common. The given atomic masses are themselves only averages, and any particular set of atoms will vary in composition slightly; see the Periodic Table in [180].

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

Atom	Symbol	+/−	neutrons	outer	lacking	variation	radius
Hydrogen	H	1	1	1	1	+0.0008	
Carbon	C	6	6	4	4	+0.01	
Nitrogen	N	7	7	5	3	+0.007	
Oxygen	O	8	8	6	2	-0.0006	
Fluorine	F	9	10	7	1	-0.002	
Sodium	Na	11	12	1	7	-0.01	
Magnesium	Mg	12	12	2	6	+0.31	
Phosphorus	P	15	16	5	3	-0.03	
Sulfur	S	16	16	6	2	+0.06	
Chlorine	Cl	17	18	7	1	+0.45	
Potassium	K	19	20	1	7	+0.10	

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 'variation' give the difference between the observed atomic 'weight' (the mass in atomic units) of the naturally occurring isotopic distribution and the mass of the 'standard' isotope's protons and neutrons. The column 'radius' lists the Pauli exclusion radius.

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). This simple rule facilitates the determination of molecular bond formation.

The digital description of an atom is to be contrasted with the analog description of the Schrödinger equation (see Chapter 15). This equation describes the electron distribution, which is the key determinant of atomic interaction. We are forced to consider effects on this level in many cases, but operating at the atomic level has clear advantages.

There are other simple rules in chemistry that clarify bond formation, such as the resonance principle (Section 14.1). This rule 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 15).

We seek to illuminate rules like these in proteins; see Chapter 4 for an introduction to proteins.

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 observation that hydrophobic effects (Section 3.5) are involved 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 7) yields such a description, and we show that this can effect many important phenomena, including protein binding (Chapter 6) and the flexibility of the peptide bond (Chapter 14).

2.3 Eternal Struggle

The life of a protein in water is largely a struggle for the survival of its hydrogen bonds. The hydrogen bond (cf. Chapter 5) 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, but proteins must struggle to live with it [137].

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 [85], 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, 214, 234], 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 at 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 4.1.2). 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 under-wrapped by carbonaceous groups as **dehydrons** (Section 3.5.2) to simplify terminology. The name derives from the fact that these hydrogen bonds benefit energetically from being dehydrated.

2.4 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. Moreover, it has been recently observed that exposed hydrogen bonds appear to be sites of protein-protein interactions [73]. 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. Such forces also have analogies in personal, social and political interactions where forces based on detrimental circumstances cause a beneficial outcome. 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, or even non-polar materials. Changing the dielectric environment near an under-protected hydrogen bond makes it 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 8.1). 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 provide protection from water. It is intriguing that it arises from a defect which provides an opportunity to interact.

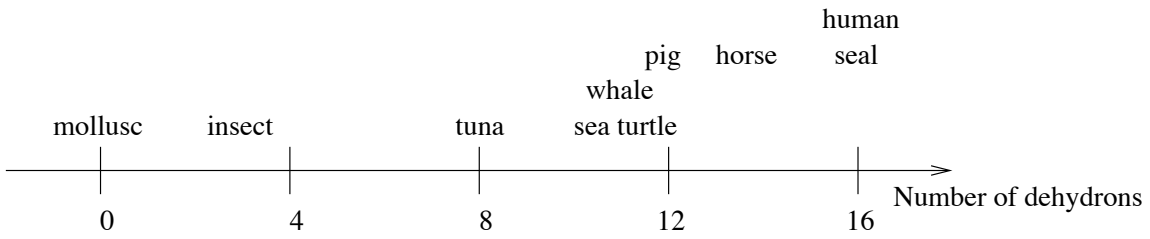


Figure 2.1: Number of dehydrons in the protein myoglobin found in various species [73].

2.5 Pchemomics

The term "omics" refers to the use of biological data-bases to extract new knowledge by large-scale statistical surveys. The term "cheminformatics" is an accepted moniker for the interaction of informatics and chemistry, so there is some precedent for combining terms like pchem (a.k.a., physical chemistry) with a term like 'omics.' We do not suggest the adoption of the (unpronounceable) term pchemomics, but it serves to suggest the particular techniques being combined in a unique way. An example of pchemomics is the early study of the hydrogen bond [132]. Indeed, the original study of the structure of the peptide bond (see section 8-4 of [179]) 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 [67]. We examine just one such result in Section 2.5.4; 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 [73].

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 [73].

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 [107] 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 [76] 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 [139]. 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.

2.5.1 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 [221], 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.

2.5.2 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?**

2.5.3 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 [221], 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 they 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 [143].

The function of DNA is largely to store sequence information, but proteins operate as three-dimensional widgets. All proteins have a three-dimensional representation, even if it is not one that forms into a stable, biologically viable, structure. The PDB is a curated database of such structures which 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 interpreting 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.

2.5.4 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 [132] shows clearly both the radial and the angular dependence of the hydrogen bond. Figure 8 of [243] 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 [162]. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations has also been reported [128].

An alternative method for modeling hydrogen bonds is to study their energetics via quantum mechanical calculations and to interpolate the resulting energy surfaces [209, 168].

Due to the primary importance of the hydrogen bond in protein structure, we will review what is known and not known in Chapter 5.

2.5.5 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 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. We may even 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.

2.6 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,

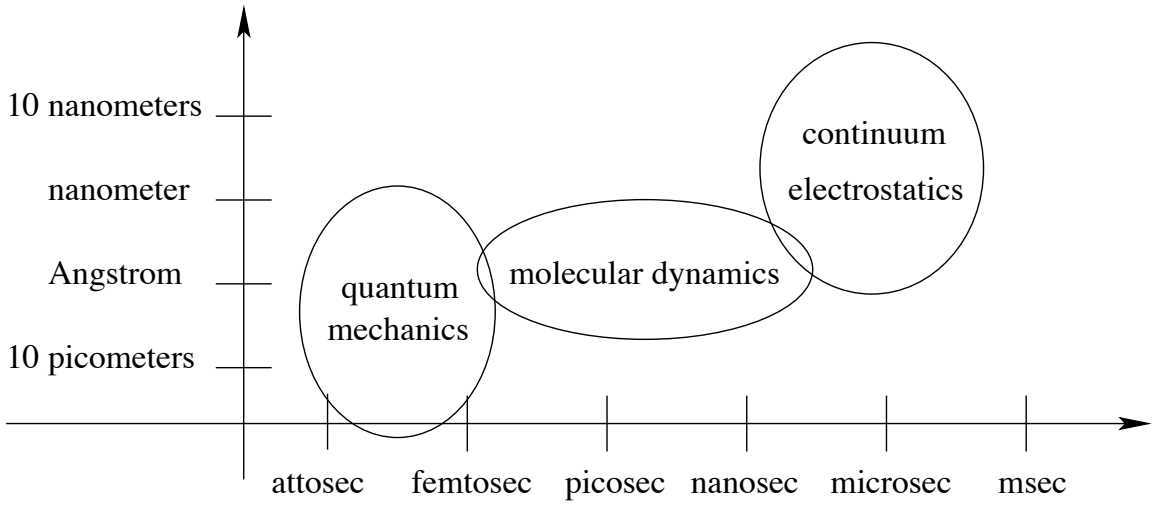


Figure 2.2: Spatial and temporal scales of biomolecular models. See the text for more details.

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 [105]. Scale separation often simplifies the interactions among the different scales. The differences often occur in both physical 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 15). 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 invoking 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 [46, 204]. 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 [111].

The lack of dramatic 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 [61]. The electric field is governed by the largest-scale model and causes a change in the smallest-scale model, forcing a re-structuring of the molecular model (Chapter 14).

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 14.4). On the other hand, they are still complicated enough that sufficiently long-time simulations, required for biological accuracy, are often prohibitive [7]. Electrostatic models hope to capture the expected impact of dielectric solvation, but there are limitations here as well. The dielectric coefficient of water is orders of magnitude larger than what would be found inside a large protein. This is a very large jump in a coefficient in a continuum model, and it is prudent to be cautious about any model with such large changes. It is clear that in the neighborhood of the jump in the coefficient, a more complex model might be required [204].

2.7 Exercises

Exercise 2.1 *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 [179].*

Exercise 2.2 *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 2.3 *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 4.1). 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 2.4 *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 2.5 *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 discretization with m degrees of freedom 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

Electronic 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 [179, 180] are still natural references.

The order of forces, or bonds, 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. Typical examples are shown in Figures 4.4–4.7 for aminoacid sidechains and Figure 14.1 for the peptide bond. Single lines represent single bonds and double (parallel) lines represent double bonds. 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. 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 [180, 179] allows the prediction of covalent bond formation through counting of electrons in the outer-most shell 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 14.1) 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 14.

3.1.2 Ionic bonds

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. Ionic bonds involve the direct attraction of electrons in one atom to the positive charge of another.

The potential for the electrostatic interaction between two charged molecules, separated by a distance r, is simply

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

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 simply $-r^{-1}$.

We will see that different 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 15.7. 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 [221]. The detailed structure of hydrogen bonds in biology is still being investigated [128, 162, 214, 234]. 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 5.2 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 5.

3.1.4 Cation-π interactions

Aromatic residues (phenylalanine, tyrosine and tryptophan: see Section 4.5.5) 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 [89, 244, 44]. 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 13.

3.2 Interactions involving dipoles

Many interactions can be modeled as dipole-dipole interactions, e.g., between water molecules. More generally, the use of partial charges (cf. Table 13.1) represents many interactions as dipole-dipole interactions. Forces between molecules with fixed dipoles are often called Keesom forces [87]. 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.

3.2.1 Single-file dipole-dipole interactions

In the single-file orientation, the base dipole has a positive charge at $(\epsilon, 0, 0)$ and a negative charge at $(-\epsilon, 0, 0)$; the other dipole is displaced on the x-axis at a distance r: a positive charge at $(r+\epsilon, 0, 0)$ and a negative charge at $(r-\epsilon, 0, 0)$ (cf. Figure 3.1). The potential due to the base dipole at a

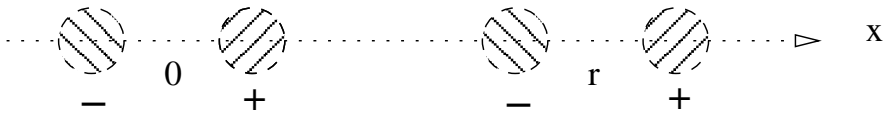


Figure 3.1: Single-file dipole-dipole configuration.

distance $r >> \epsilon$ along the x-axis is

$$\begin{aligned} V(r) =& \frac{1}{r-\epsilon} - \frac{1}{r+\epsilon} = \frac{(r+\epsilon)-(r-\epsilon)}{(r-\epsilon)(r+\epsilon)} \\ =& \frac{2\epsilon}{(r-\epsilon)(r+\epsilon)} = \frac{2\epsilon}{r^2-\epsilon^2} \approx 2\epsilon r^{-2} = \mu r^{-2}, \end{aligned} \tag{3.2}$$

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

$$\lim_{r\to\infty} g(r)/f(r) = 1. \tag{3.3}$$

In (3.2), $f(r) = 1/(r^2-\epsilon^2)$ and $g(r) = r^{-2}$, so that $g(r)/f(r) = 1-\epsilon^2/r^2$, and thus (3.3) follows. Moreover, we can get a quantitative sense of the approximation: the approximation in (3.2) is 99% accurate for $r \geq 10\epsilon$, and even 75% accurate for $r \geq 2\epsilon$.

In the field of the dipole (3.2), 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 9.2.2.

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 (3.2). 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:

$$\frac{\mu}{(r+\epsilon)^2} - \frac{\mu}{(r-\epsilon)^2}\,. \tag{3.4}$$

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

$$\begin{aligned} \frac{\mu}{(r+\epsilon)^2} - \frac{\mu}{(r-\epsilon)^2} =& -\mu\frac{(r+\epsilon)^2-(r-\epsilon)^2}{(r+\epsilon)^2(r-\epsilon)^2} \\ =& -\mu\frac{4r\epsilon}{(r+\epsilon)^2(r-\epsilon)^2} \approx -4\mu\epsilon r^{-3} = -2\mu^2 r^{-3}. \end{aligned} \tag{3.5}$$

The negative sign indicates that there is an attraction between the two dipoles in the configuration Figure 3.1.

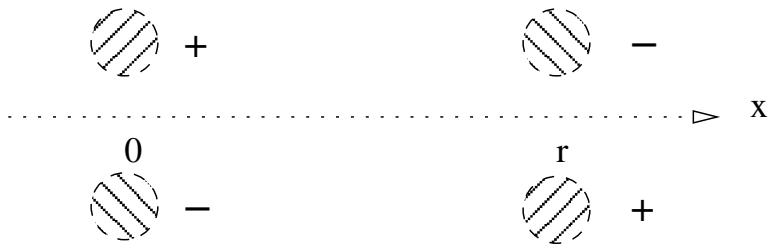


Figure 3.2: Parallel dipole-dipole configuration.

The electric force field $\mathbf{F}$ is the gradient of the potential ∇V. For V defined by (3.2), only the x-component of ∇V is non-zero along the x-axis, by symmetry. Differentiating (3.2), we find that for $r >> \epsilon$ along the x-axis,

$$\begin{aligned} F_x(r,0,0) =& -(r-\epsilon)^{-2} + (r+\epsilon)^{-2} \\ =& \frac{-(r+\epsilon)^2 + (r-\epsilon)^2}{(r-\epsilon)^2(r+\epsilon)^2} = \frac{-4\epsilon r}{(r-\epsilon)^2(r+\epsilon)^2} \\ \approx& -4\epsilon r^{-3} = -2\mu r^{-3}. \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} -\frac{2\mu}{(r+\epsilon)^3} + \frac{2\mu}{(r-\epsilon)^3} =& 2\mu \frac{(r+\epsilon)^3 - (r-\epsilon)^3}{(r+\epsilon)^3(r-\epsilon)^3} \\ =& 2\mu \frac{6r^2\epsilon + 2\epsilon^3}{(r+\epsilon)^3(r-\epsilon)^3} \approx 6\mu^2 r^{-4}, \end{aligned} \tag{3.7}$$

which is equal to the derivative of the potential (3.5) as we would expect.

3.2.2 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 $(r,-\epsilon,0,0)$ and a negative charge at $(r,+\epsilon,0,0)$ (cf. Figure 3.2).

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

$$V(x,y) = \frac{1}{\sqrt{(y-\epsilon)^2 + x^2}} - \frac{1}{\sqrt{(y+\epsilon)^2 + x^2}} \tag{3.8}$$

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

$$\begin{aligned}\left(\frac{1}{\sqrt{(2\epsilon)^2+r^2}}-\frac{1}{r}\right)-\left(\frac{1}{r}-\frac{1}{\sqrt{(2\epsilon)^2+r^2}}\right) &= -2\left(\frac{1}{r}-\frac{1}{\sqrt{(2\epsilon)^2+r^2}}\right)\\ &= -2\frac{\sqrt{(2\epsilon)^2+r^2}-r}{r\sqrt{(2\epsilon)^2+r^2}} = -2\frac{\sqrt{(2\epsilon/r)^2+1}-1}{r\sqrt{(2\epsilon/r)^2+1}}\\ &\approx -\frac{(2\epsilon/r)^2}{r} = -\mu^2 r^{-3}.\end{aligned} \tag{3.9}$$

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

The potential $V(x, y)$ in (3.8) 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 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 $f(\epsilon, r)$:

$$\frac{\partial V}{\partial x}(x, y) = \frac{-x}{((y-\epsilon)^2+x^2)^{3/2}} + \frac{x}{((y+\epsilon)^2+x^2)^{3/2}} \tag{3.10}$$

for general y. Choosing $y = \pm\epsilon$, (3.10) simplifies to

$$\begin{aligned}\frac{\partial V}{\partial x}(r, \pm\epsilon) &= \mp\, r^{-2} \pm \frac{r}{((2\epsilon)^2+r^2)^{3/2}} = \mp r^{-2}\left(1-\frac{1}{((\mu/r)^2+1)^{3/2}}\right)\\ &= \mp\,\frac{((\mu/r)^2+1)^{3/2}-1}{r^2\,((\mu/r)^2+1)^{3/2}} \approx \mp\frac{3\mu^2}{2r^4},\end{aligned} \tag{3.11}$$

for large r/ϵ. The net force of the field (3.11) on the two oppositely charged particles on the right side of Figure 3.2 is thus $3\mu^2 r^{-4}$, consistent with what we would find by differentiating (3.9) with respect to r.

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

$$\frac{\partial V}{\partial y}(r, y) = \frac{\epsilon - y}{((y-\epsilon)^2+r^2)^{3/2}} + \frac{\epsilon + y}{((y+\epsilon)^2+r^2)^{3/2}}. \tag{3.12}$$

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

$$\frac{\partial V}{\partial y}(r, \pm\epsilon) = \frac{\mu}{(\mu^2+r^2)^{3/2}} \approx \mu r^{-3}, \tag{3.13}$$

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 3.2 cancel this effect. So there is no net force on the dipole in the direction of the y-axis.

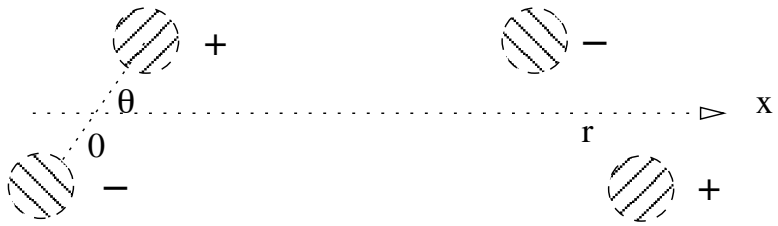


Figure 3.3: General θ-dependent dipole-dipole configuration.

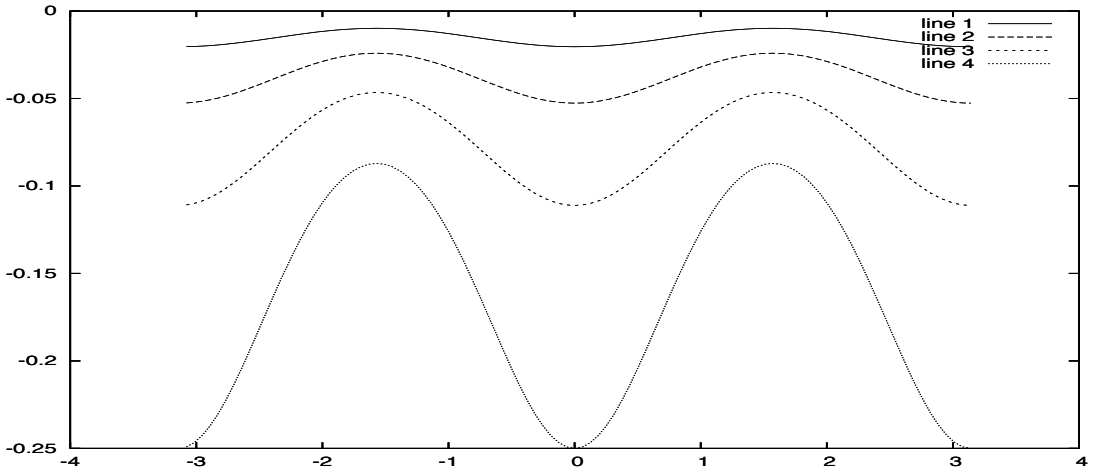


Figure 3.4: Potential energy variation $v(\rho, \theta)$ as defined in (3.17) (vertical axis) of dipoles as a function of θ (horizontal axis) for the configurations shown in Figure 3.3 for $\rho = 0.02$ (top), $0.05, 0.1, 0.2$ (bottom), where ρ is defined in (3.16).

3.2.3 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 away from the x-axis at an angle θ, as shown in Figure 3.3. 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

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

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

$$\begin{aligned} V(r,\theta) = \frac{1}{\sqrt{r^2 + (2\epsilon\sin\theta)^2}} - \frac{1}{r + 2\epsilon\cos\theta} - \left(\frac{1}{r - 2\epsilon\cos\theta} - \frac{1}{\sqrt{r^2 + (2\epsilon\sin\theta)^2}}\right) \\ = \frac{2}{\sqrt{r^2 + (2\epsilon\sin\theta)^2}} - \frac{1}{r + 2\epsilon\cos\theta} - \frac{1}{r - 2\epsilon\cos\theta} \\ = \frac{2}{\sqrt{r^2 + (2\epsilon\sin\theta)^2}} - \frac{2r}{r^2 - (2\epsilon\cos\theta)^2} \\ = \frac{2}{r}\left(\frac{1}{\sqrt{1 + \rho\sin^2\theta}} - \frac{1}{1 - \rho\cos^2\theta}\right) := \frac{2}{r}v(\rho,\theta), \end{aligned} \tag{3.15}$$

where the (nondimensional) parameter ρ is defined by

$$\rho = (2\epsilon/r)^2. \tag{3.16}$$

This expression

$$v(\rho,\theta) = \frac{1}{\sqrt{1 + \rho\sin^2\theta}} - \frac{1}{1 - \rho\cos^2\theta} \tag{3.17}$$

in (3.15) has a maximum when $\theta = 0$ and a minimum when $\theta = \pi/2$. A plot of v in (3.17) is shown in Figure 3.4 for various values of ρ. When ρ is small, the expression (3.17) tends to the limit

$$\begin{aligned} v(\rho,\theta) \approx \frac{1}{1 + \frac{1}{2}\rho\sin^2\theta} - \frac{1}{1 - \rho\cos^2\theta} \\ \approx \left(1 - \tfrac{1}{2}\rho\sin^2\theta\right) - \left(1 + \rho\cos^2\theta\right) = -\tfrac{1}{2}\rho\left(1 + \cos^2\theta\right). \end{aligned} \tag{3.18}$$

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 3.8.

3.2.4 Different dipoles

So far, we considered dipoles with identical charges and charge distributions (separations). Here we consider a single-file configuration as in Figure 3.1, but with the dipole on the right consisting of charges $\pm q$ separated by a distance δ, as depicted in Figure 3.5. We consider the potential energy of the right-hand dipole in the potential field (3.2) of the left dipole. Similar to (3.5), we find

$$\frac{\mu q}{(r+\delta)^2} - \frac{\mu q}{(r-\delta)^2} = -\mu q\frac{4r\delta}{(r+\delta)^2(r-\delta)^2} \approx -4\mu q\delta r^{-3} = -2\mu\nu r^{-3}, \tag{3.19}$$

where $\nu = 2q\delta$ is the strength of the dipole on the right. Notice that the expression (3.19) is symmetric in the two dipole strengths μ and ν.

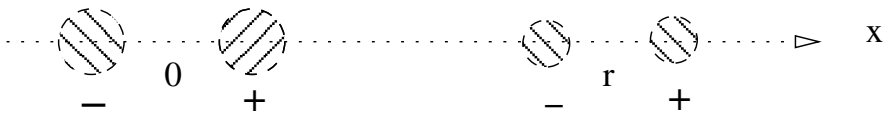


Figure 3.5: Single-file dipole-dipole configuration with different dipole strengths.

3.3 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 **??**, 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.

Debye forces and London dispersion forces [87] involve induced dipole-dipole interactions, which we will study using the results derived in Section **??**. The most significant example is the London dispersion force [87] 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 15.5.2.

The Lennard-Jones model

$$V(r) = cr^{-12} - c'r^{-6}. \tag{3.20}$$

is commonly used to model van der Waals interactions. 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 3.4. In Section 3.4.2, we consider a second, but defer to Section 15.5.2 a detailed discussion. The repulsive term r^{-12} is a convenient model, whereas other terms are more accurate [42] (cf. Section 15.7).

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.

3.4 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 [87]. More subtly, two polarizable molecules can induce dipoles in each other, via what are called London dispersion forces [87].

3.4.1 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

$$\mu_i \approx \alpha \mathbf{F} \tag{3.21}$$

Figure 3.6: 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.

for small $\mathbf{F}$, where α is the polarizability. This is depicted visually in Figure 3.6, 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 had 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 (3.21) will be deduced by a perturbation technique for small $\mathbf{F}$ from the concepts in Section 15.5.

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 3.5. This takes some justification, but it will be addressed in Chapter 9. There is ambiguity in the representation in that only the product $q\delta$ matters: $\mu = q\delta$.

We derived in (3.6) that the electric force field due to a fixed dipole μ_f has magnitude

$$F_x = 2\mu_f r^{-3}, \tag{3.22}$$

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 (3.21) and (3.22), we conclude that the strength of the induced dipole is

$$\mu_i \approx 2\alpha\mu_f r^{-3}. \tag{3.23}$$

From (3.19), we know that the potential energy of the two dipoles is

$$V(r) \approx -2\mu_f\mu_i r^{-3} \approx -4\alpha\mu_f^2 r^{-6}, \tag{3.24}$$

in agreement with the Lennard-Jones model in (3.20).

3.4.2 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 15.5), they can become polarized. 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 3.6. 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.

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 (3.22): $F_x = 2\mu r^{-3}$. 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 (3.19) remains valid for the potential energy of the induced dipoles:

$$V \approx -2\mu^2 r^{-3}. \tag{3.25}$$

But how big is the induced dipole μ in expression (3.25)? We saw in (3.23) 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

$$V \approx cr^{-9}, \tag{3.26}$$

which is quite different from the Lennard-Jones model.

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 (3.25) 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 (3.21) and (3.22), 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

$$r = \sqrt[3]{2\alpha}. \tag{3.27}$$

Note that this is dimensionally correct, since the units of the polarizability α are the same as volume. Thus using the two equations (3.21) and (3.22) 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 15.5.3 that this value of r can be interpreted in a mathematical, if not physical, sense.

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

We will derive in (15.56) a result that confirms the basic dependence of the dipole strength on r, namely

$$\mu \approx cr^{-3}, \tag{3.28}$$

where an expression for the constant c will be made explicit. Thus, at least for r sufficiently large, the expression (3.25) appears to be the correct asymptotic behavior.

The above arguments could be interpreted in the following way. For very large r, the potential of the induced dipole is very small, comparable to r^{-9}. As the induced field $\mathbf{F}$ increases, the polarizability will saturate, and μ will tend to a limit. For example, this might take a form similar to

$$\mu(r) \approx \frac{c_1}{1 + c_2 r^{-3}}. \tag{3.29}$$

Thus the potential energy varies from (3.26) for large r to cr^{-3} for smaller values of r. For this reason, using the intermediate exponent of 6, approximating the behavior of the potential energy by r^{-6}, may be a reasonable compromise.

3.4.3 Lennard-Jones potentials

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

$$V(r) := \epsilon \left(\left(\frac{\rho}{r}\right)^{12} - 2\left(\frac{\rho}{r}\right)^{6} \right). \tag{3.30}$$

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 [25]. 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 Hydrophobic interactions

There is another force that is crucial in biology, sometimes said to be more important than even the hydrogen bond force [120]. It is called the **hydrophobic force** [21], which derives from the **hydrophobic effect** [220]. This effect is one of the central topics of our study. However, the hydrophobic effect has many manifestations in protein behavior.

There is a simple view of how hydrophobic forces work. There are certain molecules that are hydrophobic (cf. Section 3.5.2 and Chapter 7), 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 4.4.4) since hydrophobic side chains take up more volume in water. Recent results show how a hydrophobic force can arise through a complex interaction between polarizable (e.g., hydrophobic) molecules and (polar) water molecules [49, 50]. These arguments are compelling, but they suggest a nonspecific interaction. Indeed, hydrophobic attraction lead to nonspecific binding [77].

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.

3.5.1 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 [62]. More generally, solvent mediation can alter any electrostatic force via dielectric effects (Chapter 16).

One type of solvent effect that is expressed on the quantum level is the rigidity of the peptide bond (Chapter 14) which requires an external field to select one of two resonant states. Such a field can be due to hydrogen bonds with the 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 [61].

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). We will see that this can be related to a gross structural change in protein conformation that has implications for drug delivery [75].

Changes in dielectric properties of the environment can have a substantial impact on any elec-

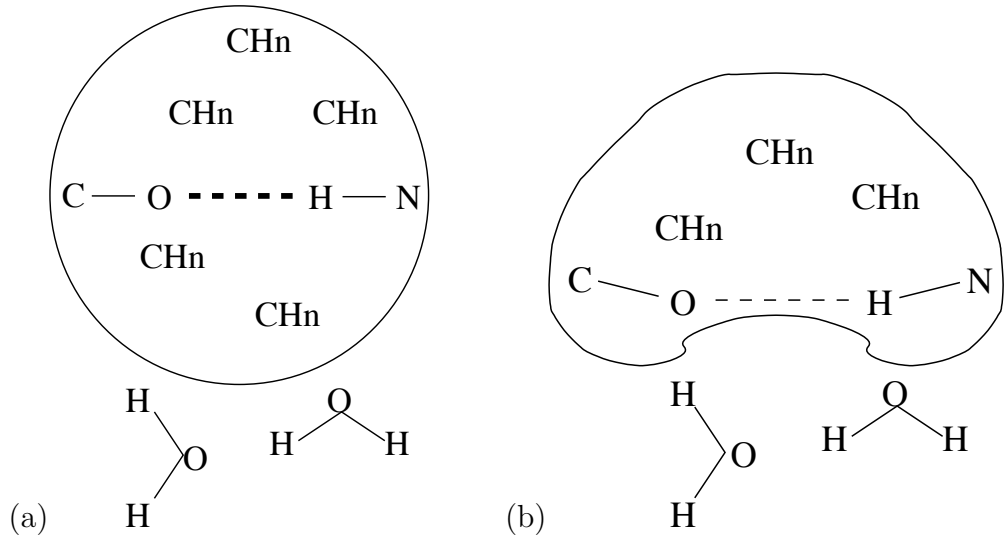


Figure 3.7: (a) Well wrapped hydrogen bond (b) Underwrapped hydrogen bond.

trical property. But rather than try to address this by a general model, 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.

3.5.2 Dehydrons

In [73], 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 (tentative) 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.

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 [21, 220]. 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 [71]. The concept of wrapping of a hydrogen bond is depicted informally in Figure 3.7. A well wrapped hydrogen bond (Figure 3.7(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 3.7(b)) allows a closer approach by water to the hydrogen bond, and this tends to disrupt the bond, allowing the distance between the groups to increase and the bond to weaken.

It is possible to identify dehydrons as **under wrapped hydrogen bonds** (**UWHB**) by simply counting the number of hydrophobic side chains in the vicinity of a hydrogen bond. This approach is reviewed in Section 7.2. More accurately, a count of all (nonpolar) carbonaceous groups gives a more refined estimate (Section 7.3). However, it is possible to go further and quantify a force

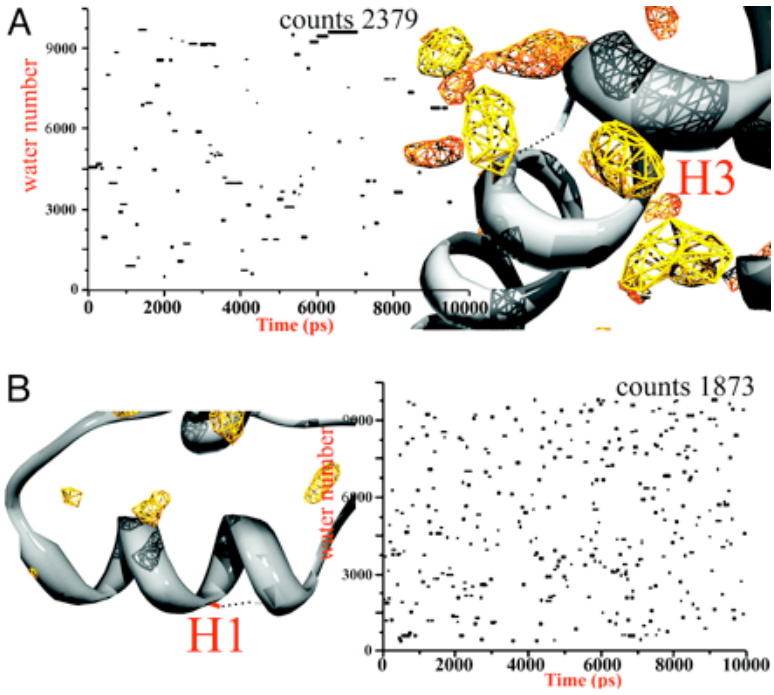


Figure 3.8: Dynamics of water near hydrogen bonds, reproduced from Fig. 5 in [45]. (A) Hydrogen bond (H3) is well wrapped. (B) Hydrogen bond (H1) is underwrapped.

associated with dehydrons which provides a more refined measure of the effect geometry [73] of the wrappers (Section 7.5).

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 7). The HIV-1 capsid protein P24 complexed with antibody FAB25.3 provides a dramatic example [73].

3.5.3 Dynamics of dehydrons

The extent of wrapping changes the nature of hydrogen bond. Hydrogen bonds that are not protected from water do not persist [45]. Figure 5 of [45] shows the striking difference of water residence times for well wrapped and underwrapped hydrogen bonds. Private communication with the authors of [45] 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 3.9 are quite different, confirming our prediction based on Figure 3.7 that the coupling of the hydrogen bond characteristics with the water environment would be different.

The H-bond R208–E212 depicted in Fig. 5(A) [45] is well wrapped whereas V189–T193 depicted in Fig. 5(B) is a dehydron (see Fig 3a in [69] 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 [45].

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 3.7(b) when water is near, to one more like that depicted in Figure 3.7(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 3.9. 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. 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.

3.5.4 Simulated dynamics

The data in Figure 3.9 can be interpreted via a simulation which is depicted in Figure 3.10. This figure records the distribution of positions for a random walk subject to a restoring force defined by

$$x_{i+1} = x_i + \Delta t(f_i + \phi(x_i)) \tag{3.31}$$

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

$$\phi(x) = (0.1/x)^{12} - (0.1/x)^6. \tag{3.32}$$

The particular time step used in Figure 3.10 is $\Delta t = 0.02$; the simulation was initiated with $x_1 = 0.1$ and carried out for 10^5 steps.

The simulation (3.31) 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 (3.32). Such a system exhibits a distribution with an exponential decay, as verified in Figure 3.10 by comparison with a least-squares fit of the logarithm of the data to a straight line.

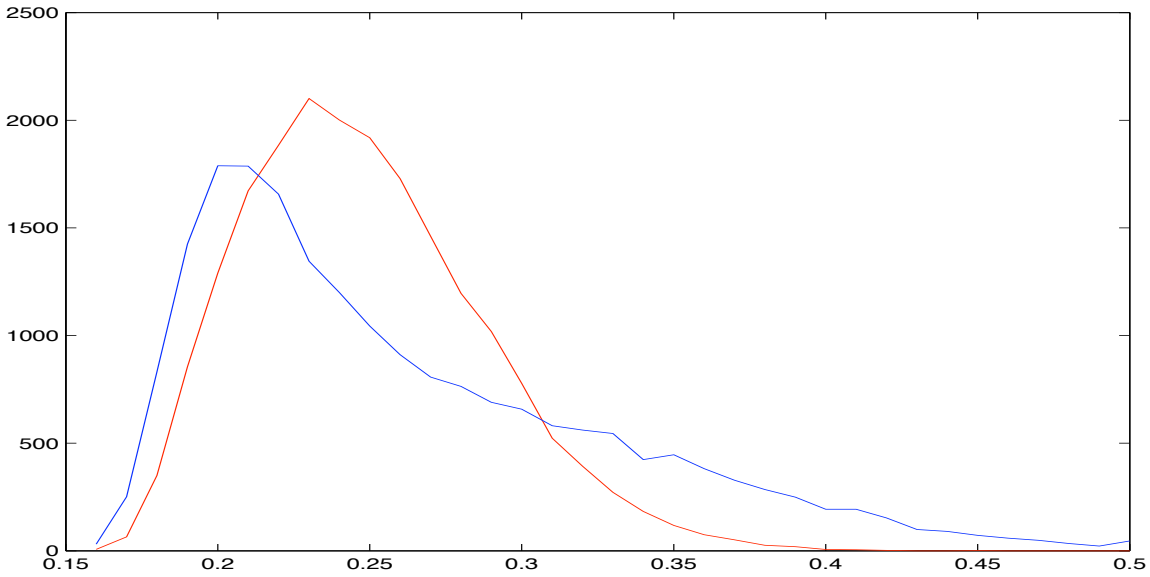


Figure 3.9: Distribution of bond lengths for two hydrogen bonds formed in a structure of the sheep prion [45]. The horizontal axis is measured in nanometers, whereas the vertical axis represents numbers of occurrences taken from a simulation with $20,000$ data points with bin widths of 0.1 Ångstrom. The distribution for the well-wrapped hydrogen bond (H3) is depicted with a solid line, whereas the distribution for the underwrapped hydrogen bond (H1) is depicted with a dotted line.

3.5.5 Stickiness of dehydrons

Desolvation of an underwrapped hydrogen bond can occur when a ligand binds nearby, as depicted in Figure 3.11. 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 8.

3.6 Exercises

Exercise 3.1 *Show that the approximation in (3.2) 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 (3.5) is still correct if we use the exact form in (3.2) instead of the approximation* $V(r) \approx \mu r^{-2}$.

Exercise 3.4 *Prove that (3.7) is still correct if we use the exact form in (3.6),* $F_x(r,0,0) = -4\epsilon r(r-\epsilon)^{-2}(r+\epsilon)^{-2}$, *instead of the approximation* $F_x(r,0,0) \approx -2\mu r^{-3}$.

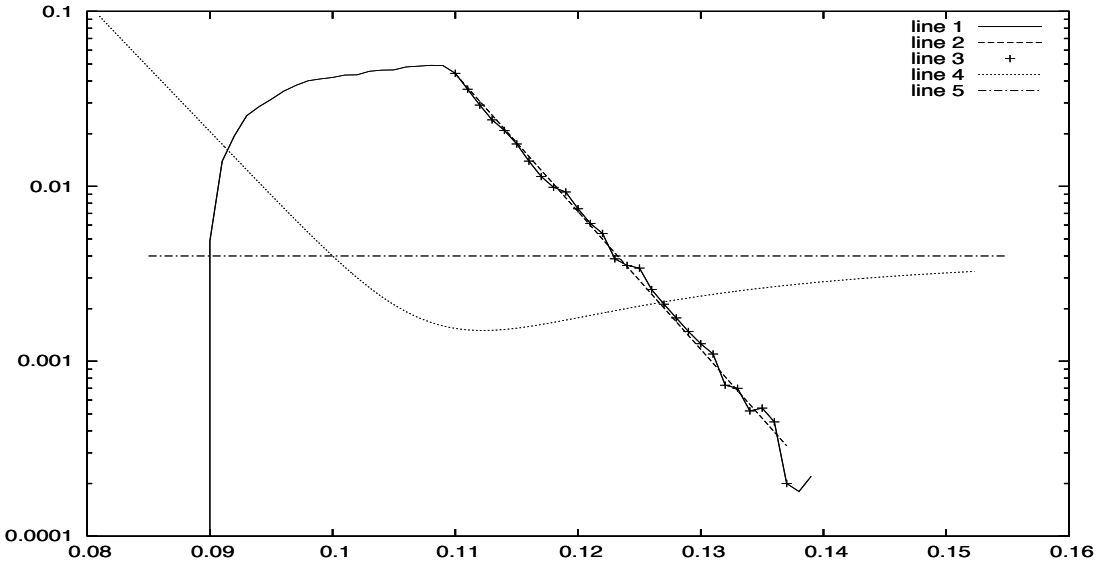


Figure 3.10: Simulation of a random walk with a restoring force. Shown is the distribution of values x_i defined in (3.31) for 10^5 time steps i, starting with $x_1 = 0.1$, scaled by a factor of 10^{-3}. Also shown is a graph of $\phi + 0.03$ where ϕ is the potential (3.31). The dot-dashed horizontal 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.

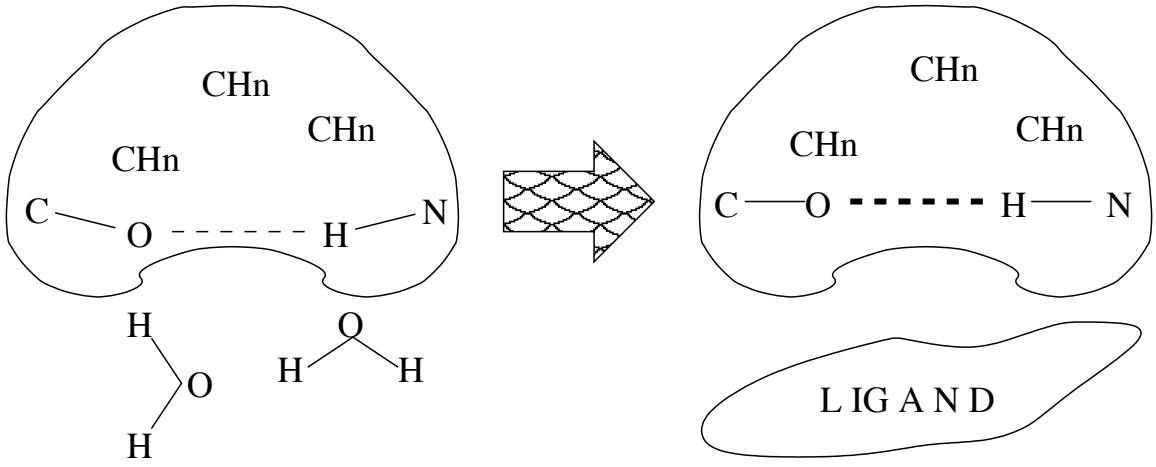


Figure 3.11: Cartoon showing dehydration due to ligand binding and the resulting strengthening of an underwrapped hydrogen bond.

Exercise 3.5 *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 3.6 *Consider the expression in (3.15). Prove that, for any $\rho < 1$, it has a maximum when $\theta = 0$ and a minimum when $\theta = \pi/2$.*

Exercise 3.7 *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 3.8 *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 [108]. This can be done with one distance parameter and three angular parameters.)*

Exercise 3.9 *Describe the orientation of the dipoles that corresponds to $\theta = 2\pi$ in Figure 3.4.*

Chapter 4

Protein basics

Proteins are sequences of amino acids which are covalently bonded along a "backbone." 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, 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.

It is not our intention to provide a complete introduction to the structure of proteins. Instead, we suggest consulting texts [43, 184] 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 chains of amino acid residues whose basic unit can be considered to be the peptide group shown in Figure 4.1. The **trans** form (a) of the peptide bond is the most common state, but the **cis** form (b) has a small but significant occurrence [98, 175].

The chain is repeated many times, up to many hundreds of **backbone** C_α carbons. The residues R^i come from the different amino acids and will be described in more detail in Section 4.1.2. See Figure 4.4 for some of the smaller residues. A cartoon of a peptide sequence is depicted in Figure 4.2.

The peptide chain is joined at the double bond indicated between the N and the O in Figure 4.1. 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. The **N-terminus**, or **N-terminal end**, has an NH_2 group instead of just N, and nothing else attached. In the standard numbering scheme, this is the beginning of the chain. The **C-terminus**, or **C-terminal end**, has a CO_2H group instead of just CO, and nothing else attached. In the standard numbering scheme, this is the end of the chain.

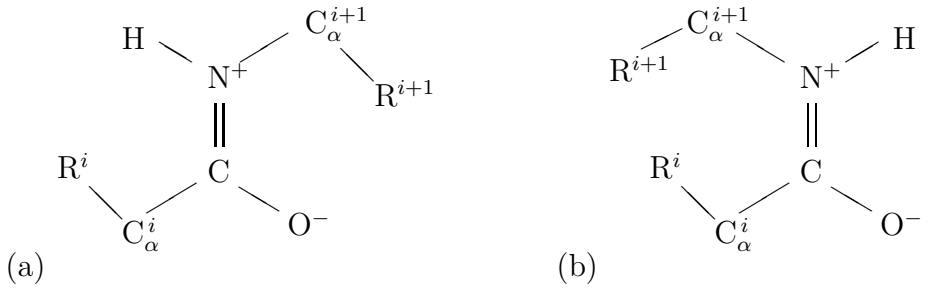


Figure 4.1: The rigid state of the peptide bond: (a) trans form, (b) cis form. The double bond between the central carbon and nitrogen keeps the peptide bond planar. Compare Figure 14.1.

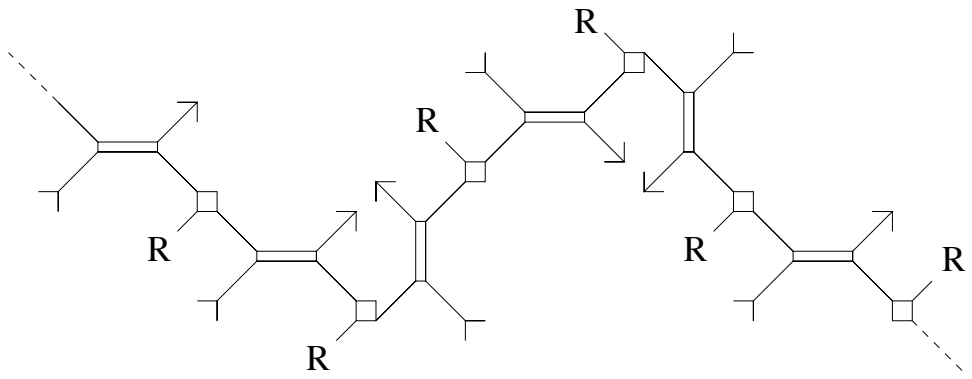


Figure 4.2: Cartoon of a peptide sequence where all of the peptides are in the trans form (cf. Figure 4.1). The small boxes represent the C-alpha carbons, the arrow heads represent the amide groups NH, the arrow tails represent the carbonyl groups CO, and the thin rectangular boxes are the double bond 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.

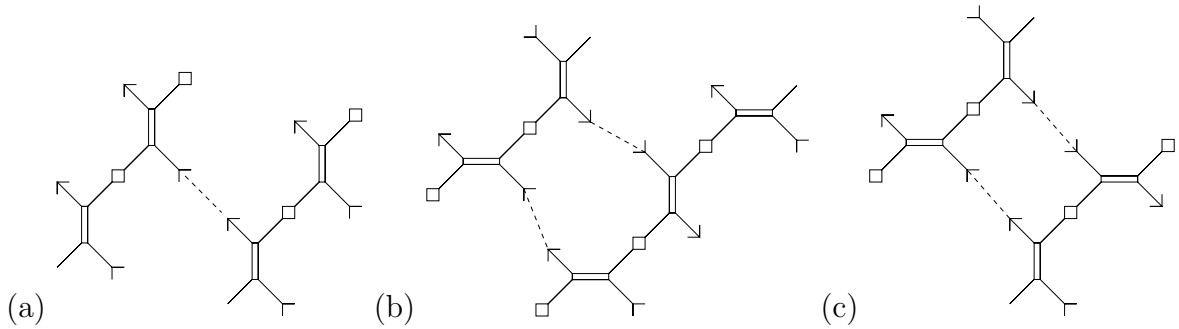


Figure 4.3: The hydrogen bond (dashed line) configuration in (a) α-helix, (b) antiparallel β-sheet, and (c) parallel β-sheet. The amide (N-H) groups are depicted by arrow heads and the carbonyl (O-C) groups are depicted by arrow tails.

4.1.1 Hydrogen bonds and secondary structure

The representation of proteins as a linear sequence of amino acid residues depicted in Figure 4.2 is called the **primary structure**. Proteins have a hierarchy of structure, the next being **secondary structure** consisting of two primary types: alpha-helices and beta-sheets (a.k.a., α-helices and β-sheets).

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. The hydrogen bond arrangement is depicted in Figure 4.3(a) between two such peptide groups.

Beta sheets represent different hydrogen bond arrangements, as depicted in Figure 4.3: (b) is the anti-parallel arrangement and (c) is the parallel. Both structures are essentially flat, in contrast to the helical structure in (a).

4.1.2 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 6 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—4.7.

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 "classical" amino acids commonly found.

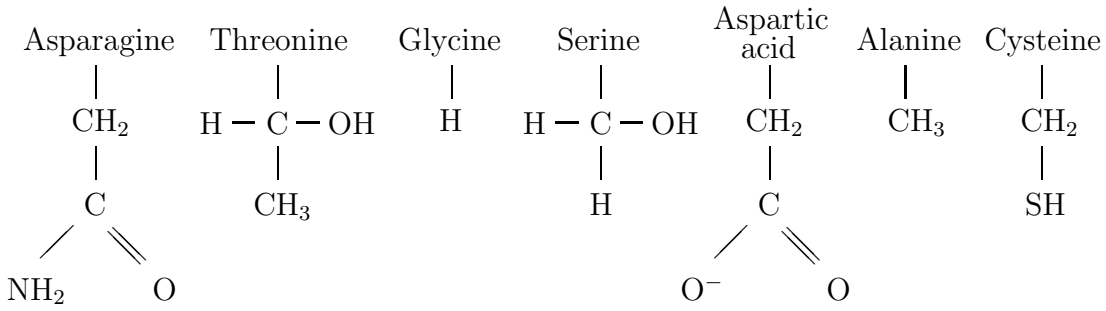


Figure 4.4: Periodic table of amino acid sidechains (residues). Not shown is the C_α carbon (see Figure 4.1), located at the top of the residue where the name appears. The smallest, and most likely to be involved in protein-ligand interactions, ordered from the left (asparagine).

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.

Full name of residue	three letter	single letter	water preference	sidechain type	nonpolar CH_n groups	intrinsic pK_a
Alanine	Ala	A	phobic	small	1	NA
Arginine	Arg	R	amphi	positive	2	12
Asparagine	Asn	N	philic	polar	1	NA
Aspartate	Asp	D	philic	negative	1	3.9-4.0
Cysteine	Cys	C	philic	polar	1	9.0-9.5
Glutamine	Gln	Q	amphi	polar	2	NA
Glutamate	Glu	E	amphi	negative	2	4.3-4.5
Glycine	Gly	G	NA	tiny	0	NA
Histidine	His	H	philic	positive	1	6.0-7.0
Isoleucine	Ile	I	phobic	aliphatic	4	NA
Leucine	Leu	L	phobic	aliphatic	4	NA
Lysine	Lys	K	amphi	positive	3	10.4-11.1
Methionine	Met	M	amphi	polar	3	NA
Phenylalanine	Phe	F	phobic	aromatic	7	NA
Proline	Pro	P	phobic	cyclic	2	NA
Serine	Ser	S	philic	polar	0	NA
Threonine	Thr	T	amphi	polar	1	NA
Tryptophan	Trp	W	amphi	aromatic	7	NA
Tyrosine	Tyr	Y	amphi	aromatic	6	10.0-10.3
Valine	Val	V	phobic	aliphatic	3	NA

Table 4.2: A taxonomy of amino acids. The code for water interaction is: phobic, hydrophobic; philic, hydrophilic; amphi, amphiphilic. Values of pK_a for ionizable residues are taken from Table 1.2 of [43].

residues are either charged or polar. But the converse of this relationship (that is, hydrophobic implies not charged or polar) is false, and there are few such general correlations. For example, the aromatic residues are among the most hydrophobic even though they are polar, cf. Section 4.5.5.

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.

4.1.3 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 6 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

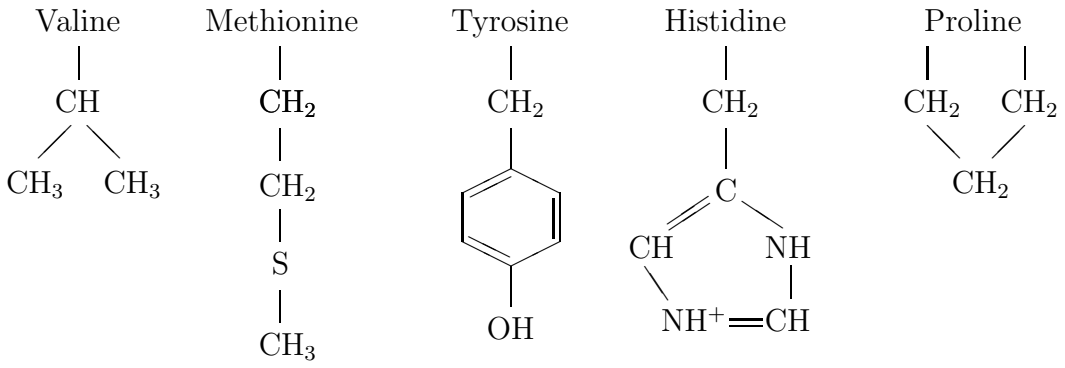


Figure 4.5: Periodic table of amino acid sidechains (sidechains only shown). Not shown is the C_α carbon (see Figure 4.1), located at the top of the residue where the name appears. The middle ground in terms of interactivity.

nearby amino acids. Such protection is provided by the hydrophobic effect. We discuss in Chapter 18 some details regarding the hydrophobic effect [21, 220], 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 [184], 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 4.5.1) 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—4.7). 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

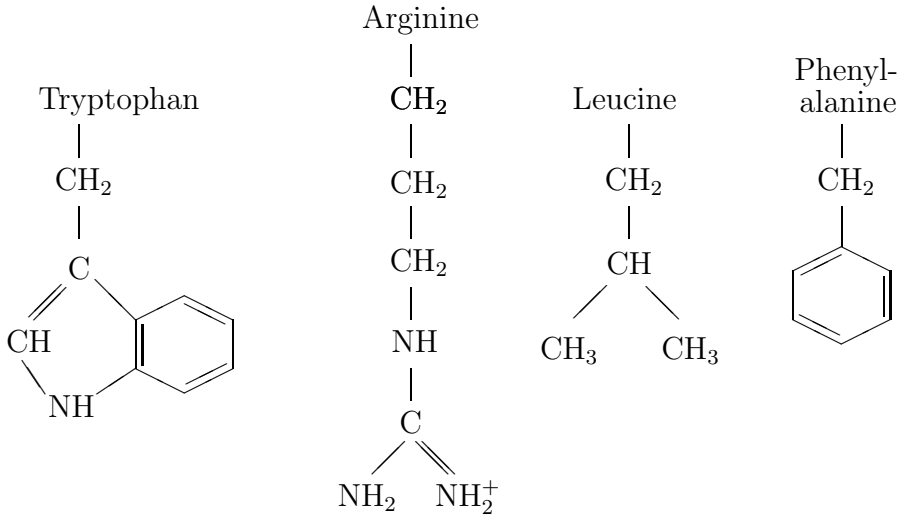


Figure 4.6: Periodic table of amino acid sidechains (sidechains only shown). Not shown is the C_α carbon (see Figure 4.1), located at the top of the residue where the name appears. The less likely to be interactive.

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 6. 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.

4.2.1 Salt Bridges

Certain sidechains are charged, as indicated in Table 4.2. 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 (Section 3.1.2) 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, but inside a protein core they would not likely prevail as they could instead be solvated by several water molecules.

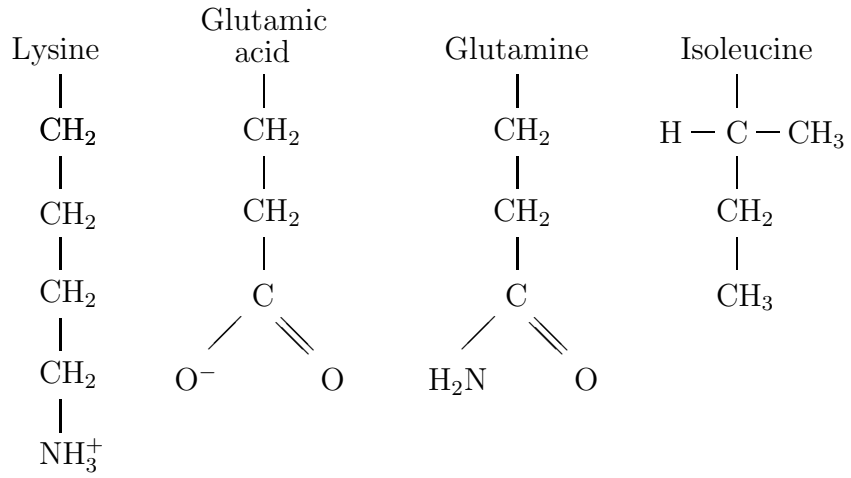


Figure 4.7: Periodic table of amino acid sidechains. The amino acids (sidechains only shown) least likely to be involved in interactions. Not shown is the C_α carbon (see Figure 4.1), located at the top of the residue where the name appears.

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. 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 [91, 172]. 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 resulting protein. A change in pH can cause the ends of some sidechains to be modified, as we discuss in Section 4.6.

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 13.1 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., PTR for

phosphotyrosine (phosphorylated tyrosine) and TPO for phosphothreonine (phosphorylated threonine).

4.4 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.

4.4.1 Conformational geometry of proteins

We recall the basic ingredients of the peptide group from Figure 4.1. 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:

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

are the $\alpha, \beta, \gamma, \delta, \epsilon, \eta$ carbons (denoted in plain text in the PDB by CA, CB, CG, CD, CE, 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.

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

$$\mathcal{B} = C_\alpha(x_{i+1}) - C_\alpha(x_i). \tag{4.1}$$

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

$$\mathcal{S}(x) = C_\beta(x) - C_\alpha(x), \tag{4.2}$$

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

$$\cos\,\theta(x_i, x_{i+1}) = \frac{\mathcal{S}(x_i) \cdot \mathcal{S}(x_{i+1})}{|\mathcal{S}(x_i)||\mathcal{S}(x_{i+1})|}, \tag{4.3}$$

where $\mathcal{B}$ is defined in (4.1), 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 [141], 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 [141], in keeping with the planar nature of sheets.

The distribution of the θ angle peaks roughly at 44, 82 and 167 degrees [141]. The peptide bond makes it difficult for θ to be much less than 50 degrees, thus the smaller peak corresponds to a motif where the side chains align as closely as possible. A small number of these occur in beta sheets, but the majority of them constitute an independent motif whose properties deserve further study.

The different structural motifs have characteristic sidechain compositions [8, 141]. For the larger values of θ, hydrophobic residues are found in most pairs; β-sheets have alternating hydrophobic and hydrophilic pairs [141]. By contrast [141], 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 [8].

We also recall the standard main-chain dihedral angles. Given a sequence of four main chain atoms a_i, let $[a_1, a_2, a_3, a_4]$ denote the dihedral (or torsion) angle between the planes defined by the points a_1, a_2, a_3 and a_2, a_3, a_4. Then the ψ, ω and ϕ angles are defined by

$$\begin{aligned} \psi(x_i) =& [N(x_i), C_\alpha(x_i), C(x_i), N(x_{i+1})] \\ \omega(x_{i+1}) =& [C_\alpha(x_i), C(x_i), N(x_{i+1}), C_\alpha(x_{i+1}] \\ \phi(x_{i+1}) =& [C(x_i), N(x_{i+1}), C_\alpha(x_{i+1}), C(x_{i+1})]. \end{aligned} \tag{4.4}$$

In Chapter 14 we study the effect of a polar environment on the flexibility of ω.

4.4.2 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 4.2). This correlation has recently been observed to have significant predictive power [82].

The conformations of ϕ_i, ψ_i are typical of different secondary structures, such as α-helix or β-sheet. The Ramachandran plot depicts the distributions of angles that are commonly adopted (cf. Exercise 4.6).

4.4.3 Sidechain rotamers

The sidechains are not rigid, so the geometric description of a sidechain requires more information than ϕ, ψ and so forth. Libraries of angular orientations of the different segments have been developed [146]. 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 4.8, the three primary

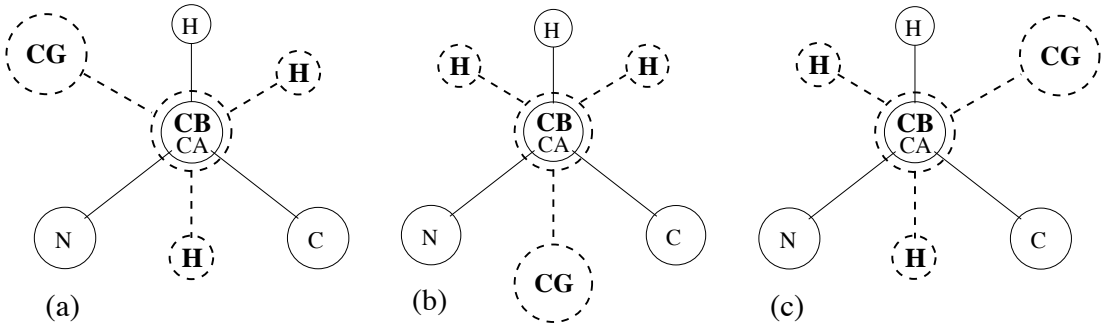


Figure 4.8: The primary sidechain rotamer conformations (a) gauche+, (b) gauche-, and (c) trans, corresponding to χ_1 values of (a) -60 degrees, (b) +60 degrees and (c) 180 degrees. The view is oriented so that the C_α and C_β atoms are aligned perpendicular to the plane of the page. The closer (and 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.

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 [142].

4.4.4 Volume of side chains

The sizes of amino acid side chains varies significantly. But there is also a more subtle size issue that is solvent dependent. In 1975, Chothia initiated a study of the size of sidechains and the change in size in the core of proteins. This was an early use of datamining in the PDB. That study was later revisited [99], and subsequent studies have further refined estimates of sidechain volume, including sizes of individual atom groups [223].

One of the significant conclusions [99] 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. 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 which are 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 in proteins at all. It may be that their role in the electrostatic balance of the protein is their reason for existence. On the other hand, if proteins had only hydrophobic cores, they would be harder to

unfold. Both of these effects may contribute to the reason for having charged and polar residues in protein cores.

4.5 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 [43]. But there are some special properties of sidechains that deserve special mention here for emphasis.

4.5.1 Glycine (and Alanine)

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

4.5.2 Proline

Proline is unique because it connects twice to the backbone. This causes a special rigidity not found with other residues. 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 [80, 208].

4.5.3 Arginine

The uniqueness of arginine is highlighted by the fact that its residue is the guanidinium ion. Guanidinium, like urea (a.k.a. carbamide), has the property that it can **denature** proteins, that is, cause them to unfold. How this is achieved, for either denaturant, is not fully understood. 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 7.5). How or why this might have a special effect is not clear.

It is very difficult to form natural water structures around an arginine [153]. There are two NH groups (in the two NH2 groups at the end) in which the N-H vectors are nearly parallel. Model building shows that it is very hard for waters attached to these hydrogens to cohabitate. There is a similar difficulty with the NH and NH2 groups, where there are N-H vectors nearly parallel. One can think of the planar structure of the terminal CN3H5 group as like a knife blade that cuts through water structures.

One property of arginine is that polyarginine is the polypeptide most able to cross a cell membrane without the help of a transporter molecule [158], and compounds rich in Arg have similar behavior [252].

4.5.4 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.

4.5.5 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 [52]. There is a doughnut of positive charge centered in the plane of the carbon and hydrogen atoms, and the hole of the doughnut contains disks of negative charge on either side of the main positive ring (see Figure 2A of [52]). 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 [167] and other behavioral impact [186]. 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) sidechain.

4.5.6 Remembering code names

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

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

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

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

The 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.

[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.

4.6 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 [237] 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 10.4). The number of water molecules in the simulation in [237] 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 [43]. The pH value (Section 10.4) 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 [43] 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, below pH four, Asp would be more likely to be protonated, so that one of the terminal oxygens would instead be an OH group. In this case, it would be appropriate to refer to the residue as aspartic acid. Similarly, for pH below 4.4, Glu would more likely have an OH terminal group, and be called glutamic acid. For simplicity, we refer to the residues in their form that is most common at physiological values of pH.

4.7 Exercises

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

Exercise 4.2 *In typical peptide bonds, the ω angle is constrained to so that the peptide bond is planar (cf. Figure 14.1). In this case, there is a relationship imposed between θ, ϕ and ψ. Determine what this relationship is.*

Exercise 4.3 *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*

$$\mu(p_i, p_{i+1}, \ldots, p_{i_\ell}) = \sum_{j=0}^{\ell} \mathcal{C}(p_{i+j}) \left(j - \tfrac{1}{2}\ell\right) \tag{4.5}$$

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? (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 4.4 *Consider the definition of macrodipole introduced in Exercise 4.3. Explain why the α-helical polypeptide $Glu_{20}Ala_{20}$ would be more stable than $Ala_{20}Glu_{20}$.*

Exercise 4.5 *Determine the Ramachandran plot 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.*

Exercise 4.6 *Determine the Ramachandran plot for a set of proteins. That is, plot the ϕ_i and ψ_i angles for all peptides in the set. Instead of using the designation in the PDB file (as helix, sheet or turn), use the software DSSPcont [4] and use a different symbol or color for the different classes.*

Chapter 5

Hydrogen bonds

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 studied extensively [16, 132], they are not yet fully understood and are still actively studied [112, 113]. Secondly, in most PDB files, hydrogens are not listed at all, due to the difficulty of locating them by typical imaging techniques. We describe how their locations can be inferred starting in Section 5.3. We begin by reviewing some of the main results.

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.

5.1 Hydrogen bond theory

Hydrogen bonds differ based on the heavy atoms that are involved. The variation in bond distance and strength is illustrated in Table 5.1 which has been extracted from [124]. 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 [124]. In some sources (including Wikipedia), the electronegativity of the constituents is given as the key factor. But according to [124], "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."

There is a strong angular dependence for the energy of the hydrogen bond [168]. One might hope that modeling the hydrogen bond as a simple dipole-dipole interaction (Section 9.2.1) would be sufficient to capture the angular dependence. But a purely partial-charge (i.e., dipole-dipole) 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" [124], as we confirm in Section 9.2.1.

A model based only on atom distances has been proposed [168], 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

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 5.1: 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" [124].

and with additional data derived from ab initio quantum chemistry calculations.

The accurate computation of the most basic hydrogen bond, the water dimer, has been of recent interest [123], even though this computation has been carried out for several decades [124]. The fact that this simple interaction is still studied is an indicator of the difficulty of determining information about general hydrogen bonds.

One question to ask about hydrogen bonds is whether the hydrogens take on a symmetric position between the donor and acceptor, or whether it favors one side (donor) over the other. The answer is: yes and no [187]. Both situations arise in nature, and there is an intriguing bifurcation between the two states, as depicted by the caricature in Figure 5.1. Depicted is a curve that was fit [187] 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. The dashed parts of the curves indicate where data has been found in both states. 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 5.1 is 3.0Å, thus clearly in the asymmetric regime (actually off the chart in Figure 5.1).

5.2 Types of hydrogen bonds

As indicated in Table 5.1, 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.

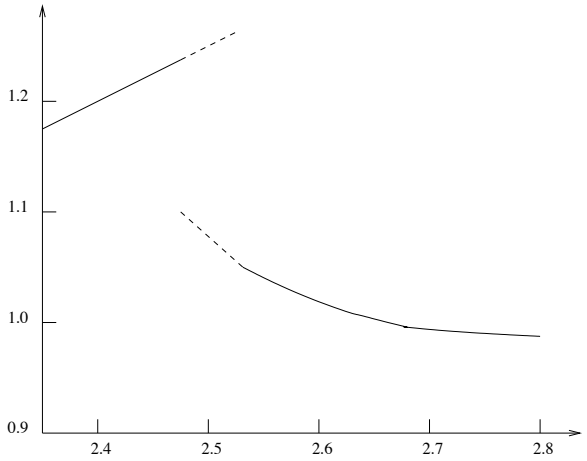


Figure 5.1: Cartoon of the bifurcation of O-H..O hydrogen bonds from a symmetric arrangement to an asymmetric arrangement, based on Figure 4 of [187]. The horizontal axis is the O-O distance and the vertical coordinate is the O-H distance (both in Ångstroms). The upper-left segment is the symmetric arrangement.

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. In addition, many of the standard sidechains can act as donors or acceptors, as listed in Table 5.2. 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. In Chapter 12, we will see that mainchain-mainchain are more much more common. By simple counts in a database of 1547 nonredundant structures, the number of M-M bonds is nearly four times the number of mainchain-sidechain (M-S and S-M) bonds combined, and it is seven times the number of sidechain-sidechain bonds. On the other hand, one finds a significant number of potential sidechain-water hydrogen bonds in many PDB files. These include apparent water bridges [185, 243]. 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 [113].

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 5.2: Donors and acceptors for sidechain hydrogen bonds. *If a Cys is involved in a disulfide bridge, it cannot be a hydrogen bond donor.

5.3 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. 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 4.1) can be estimated by a simple formula. The C-O vector and the N-H vector are very nearly parallel, so one can simply take

$$H = N + |C - O|^{-1}(C - O) \tag{5.1}$$

since the N-H distance is approximately one Ångstrom. We leave as an exercise (Exercise 5.1) to make the small correction suggested by the figure on page 282 in [179].

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 5.2. The angle between the C-N and the C-O vectors is very close to 120 degrees [160], so the C-O vector and one of the N-H vectors are very nearly parallel. So one can again take

$$H^1 = N + |C - O|^{-1}(C - O) \tag{5.2}$$

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

$$H^2 = N + \tfrac{1}{2}\left(|O - C|^{-1}(O - C) + |N - C|^{-1}(N - C)\right) \tag{5.3}$$

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 5.3: 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.

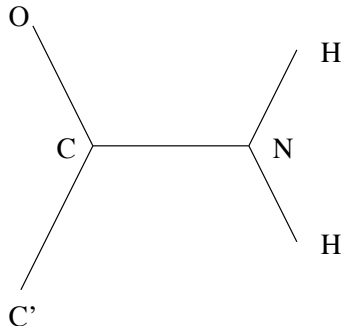


Figure 5.2: Hydrogen placement for Asn and Gln. Shown is the terminal group of atoms for the sidechains. The atom marked C' denotes the preceding carbon in the sidechain, viz., CB for Asn and CG for Gln.

We leave as an exercise (Exercise 5.2) to make the small corrections suggested by the Figure 13 in [160].

The position of hydrogens can be modeled by the bond lengths and angles given in [160]. A program called HBPLUS [155] was developed based on this information to provide hydrogen positions in a PDB format.

Most hydrogens can be located uniquely. In particular, the Appendix in [160] 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 5.3. The one which makes the stronger H-bond with an acceptor is presumably the one that is adopted.

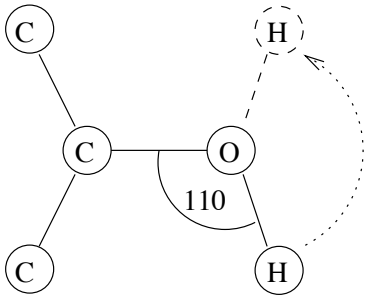


Figure 5.3: Hydrogen placement for Tyr. Both positions are possible for the terminal hydrogen.

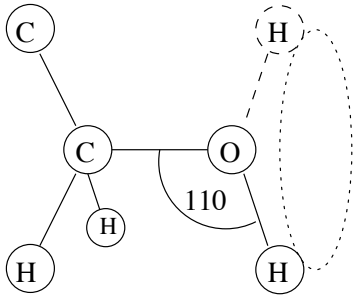


Figure 5.4: Hydrogen placement for Ser and Thr: anywhere on the dotted circle. A Cys sidechain not in a disulfide bond would be similar, with O replaced by S.

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 5.4. A Cys sidechain that is not engaged in a disulfide bond would be similar, with the oxygen in Figure 5.4 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 14.1) 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 9.2.1). According to this metric, Thr30 is the preferred donor.

5.4 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 5.6) based on those used in [155].

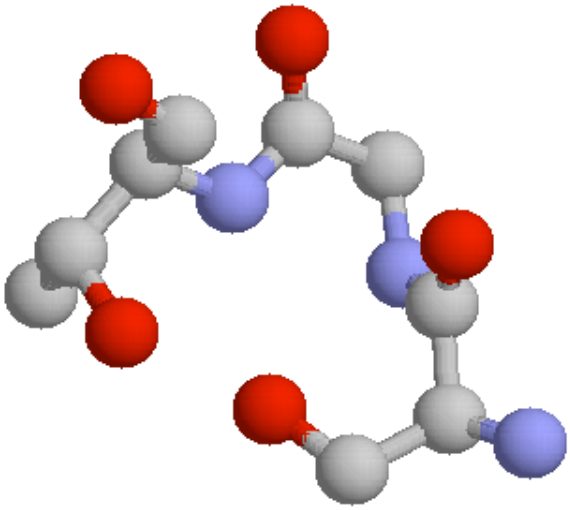


Figure 5.5: Ambiguous hydrogen placement for serine-28 (lower right)—threonine-30 (upper left) sidechain-sidechain hydrogen bond involving the terminal O-H groups; from the B chain in the PDB file 1C08. The sidechain of isoleucine-29 has been omitted but the backbone atoms are shown connecting the two residues. Only the oxygen atoms in the terminal O-H groups are shown.

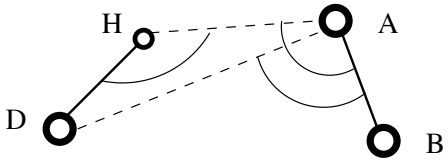


Figure 5.6: Geometric model for hydrogen bonds: D is the donor atom, H the hydrogen, A the acceptor, B acceptor antecedent (i.e. an atom one covalent bond away from the acceptor).

1. Distance between donor and acceptor $|D - A| < 3.5$Å
2. Distance between hydrogen and acceptor $|H - A| < 2.5$Å
3. Angle of donor-hydrogen-acceptor $\angle DHA > 90°$
4. Angle of donor-acceptor-acceptor antecedent $\angle DAB > 90°$
5. Angle of hydrogen-acceptor-acceptor antecedent $\angle HAB > 90°$

5.5 Carboxyl-carboxylate hydrogen bonds

Under suitable conditions, the terminal groups of Asp and Glu can be come 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 [239]. These are referred to as carboxyl-carboxylate hydrogen bonds. Although these bonds would be expected in low pH environments [200], they have been found to be critical elements of ion channels [163]. 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.

5.6 Exercises

Exercise 5.1 *Refine the formula (5.1) 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 [179].*

Exercise 5.2 *Refine the formulas (5.2) and (5.3) 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 [160].*

Exercise 5.3 *Use the improved model for the energy of a hydrogen bond in [168] 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 5.4 *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.*

Chapter 6

Determinants of protein-protein interfaces

We now turn to a key question: what factors are most influential in protein-ligand binding? We review attempts to answer this question both to give a sense for 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, but their physical basis, often attributed to favorable pairwise interactions, remains an active topic of research [193, 39, 24, 115, 145, 231, 86, 159]. Hydrophobic-polar mismatches at protein-protein interfaces are all too common and difficult to properly account for. The prediction and rationalization of binding sites for soluble proteins require that we quantify pairwise energy contributions, and concurrently, 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 [231, 79, 166, 68, 12, 65, 59].

It is well known that water removal from hydrophobic patches on the protein surface results in a high thermodynamic benefit [193, 39, 24, 115, 145, 231, 86, 159], due to an entropic gain by the solvent. Thus, hydrophobic patches might become suitable binding regions provided a geometric match on the binding partner is obtained. However, such patches are rare: most protein surfaces have the expected high ratios (typically 7:1 to 10:1) of hydrophilic to hydrophobic residues [193, 39, 24, 115, 145, 231, 86, 159]. Furthermore, even if overexposed hydrophobic patches become involved in associations, the resulting interface often presents hydrophobic-polar mismatches [215].

At the simplest level, one would expect 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 [243]. 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. 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 [243] 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 3.11.

One factor that complicates the picture of protein-protein interactions is the appearance of water molecules which appear to play a structural role, as opposed to simply mediating interactions via dielectric effects. In the protein interfaces studied in [243], 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.

6.1 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 6.5. The data here is drawn primarily from [29, 78, 90].

The site specificity of protein-protein interactions has been widely studied due to its central biological significance [47, 90, 106, 115, 116, 117]. Hydrophobic residues such as Leu and Val are more abundant at protein-ligand interfaces. As a result, the removal of water surrounding hydrophobic residues on the protein surface has been assumed to be a driving force for association [71, 213]. But it is also true that such residues are more abundant over-all (see Table 6.2).

The first question [78] 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 [29].

The second question [90] 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 [90] is a natural metric. That is, two residues are defined [90] as interacting if their C_β coordinates differ by at most 6Å (with a similar scheme to include 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 [142].

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 6.1, 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.

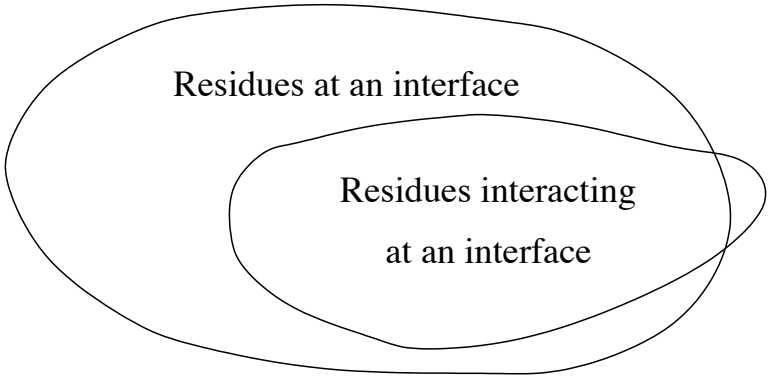


Figure 6.1: Cartoon showing possible relationship between two datasets.

The distribution of amino acid composition in proteins displays evolutionary trends [29], 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.

6.2 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 6.3, we present evidence that supports this case with suitable clarifications. However, [78] 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 [78] 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. 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

$$f_i = \frac{o_i}{\sum_{j=1}^{N} o_j} \tag{6.1}$$

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 (6.1).

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

$$r_i = f_i/g_i \tag{6.2}$$

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 [78], the following approach was taken.

Define a normalized **relative propensity** via

$$R_i = \frac{r_i}{\sum_{j=1}^{N} r_j}. \tag{6.3}$$

These relative propensities sum to one, so we can think of them like ordinary frequencies. Similarly, we multiply by 100 in (6.1) 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 [78]. In this case, the sum of the relative propensities (in percentage units) is one hundred, so the mean is five. In Table 6.1, data from [78] 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 6.1 was explained in [78] 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 6.1, 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 5.2) to be more likely at interfaces. But these residues are uniformly distributed in Table 6.1, 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 [78], namely, the amount of wrapping a residue can provide.

As noted in [78], the seven residues in Figure 4.4, 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 4.5), 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 6.1 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 6.1. It is presumed [78] 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

3-letter code	1-letter code	Nonpolar Carbons	Interface Rel. Prop.	Dehydron Rel. Prop.	Hydro-pathy
Asn	N	1	+1.28	+1.63	-3.5
Thr	T	1	+1.10	+1.41	-0.7
Gly	G	0	+0.99	+1.42	-0.4
Ser	S	0	+0.60	+0.80	-0.8
Asp	D	1	+0.34	+0.76	-3.5
Ala	A	1	+0.29	+0.6	1.8
Cys	C	1	+0.25	+0.24	2.5
Val	V	3	+0.20	-0.31	4.2
Met	M	3	+0.10	+0.10	1.9
Tyr	Y	6	+0.10	+0.10	-1.3
His	H	1	-0.25	-0.25	-3.2
Pro	P	3	-0.25	-0.25	-1.6
Trp	W	7	-0.33	-0.4	-0.9
Arg	R	2	-0.35	-0.4	-4.5
Leu	L	4	-0.35	-1.10	3.8
Phe	F	7	-0.40	-0.40	2.8
Lys	K	3	-0.42	-0.38	-3.9
Glu	E	2	-0.50	-0.11	-3.5
Gln	Q	2	-0.62	-0.6	-3.5
Ile	I	4	-0.70	-0.92	4.5

Table 6.1: Amino acids ranked according to their likelihood of being found at protein-protein interfaces. The second column indicates the number of carbon groups in the side chain. Interface and dehydron relative propensity (Rel. Prop.) is given as $R_i - 5$ as in (6.3). 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. [133] is included for reference.

Res. Code	Pairing Rel. Prop.	Pairing Rel. Freq.	Pairing Freq. [90]	Total Abundance [29]	Interface Rel. Prop.	Rim/Core freq. [37]
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 6.2: Amino acids which occur in pairs at interfaces and their relative abundances. Primary data is taken from the indicated references. Relative Propensity is defined in (6.3) and Relative Frequency is defined in (6.2). Interface Relative Propensity from Table 6.1 is included for comparison.

is exposed; Ala is the penultimate under-wrapper and may also exhibit some of the polar qualities of Gly (cf. Section 4.5.1).

6.3 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 [90] 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. In this setting, some dominant residues are indeed hydrophobic, although it is pointed out in [90] 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 6.2 the residues and their relative propensities, as defined in (6.3), in decreasing order.

Two of the residues in Table 6.2 with greatest relative propensity, namely Trp and Pro, are distinctively hydrophobic, as we might expect. However, these are also two of the most unique residues, as discussed in Section 4.5. Moreover, other high-ranking residues are as found in Table 6.1. The differences between this table and Table 6.1 reflect the fact that we are now asking about residues which are in proximity of a specific residue and thus may be interacting in some direct way.

Since Table 6.2 does not provide relative abundances directly, we need to say how these have been derived. The fundamental data in Table 6.2 is Table II on page 93 in [90], 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

$$F_i = 100 \frac{\sum_{j=1}^{20} C_{ij}}{\sum_{i,j=1}^{20} C_{ij}} \tag{6.4}$$

to be the amino acid pairing frequency, shown in the column entitled 'Pairing Freq. [90]' in Table 6.2.

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 [29] which are reproduced in the column entitled 'Total Abundance [29]' in Table 6.2. 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 6.2 represents the ratio of F_i, defined in (6.4), to the abundances reported in [29].

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 [90].

6.4 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

$$O_{ij} = \frac{C_{ij}}{f_i f_j} \tag{6.5}$$

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.

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 [90], 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 [29] 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 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 4.5.2 and 4.5.5). 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 [189].

The following four pairs 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 is a salt-bridge, and the second is a charge-polar interaction known as a cation-π interaction [89, 244, 44] (see Section 13.1) based on the special polarity of aromatic residues (Section 4.5.5). The latter two pairs are charged and polar residues as well. The next four pairs in ranking of odds ratio are Cys-Ser (1.20), Asp-Arg (1.19), Met-Met (1.16) and Cys-His (1.15). These show a similar mix of polar 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 [90] are much smaller than found for alpha helices or beta sheets [142]. The top thirty values for the odds ratios for amino acid pairs with $\theta < 50$ (Section 4.4.1) are all greater than two, with the highest being 3.75 [142]. 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 [142]. 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.

When we add the further analysis in [37] which differentiated the prevalence of core versus rim residues in protein interfaces, the picture is clarified. In [37], 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 6.2. There is a strong correlation between being charged or polar and preferring the rim, as indicated in Table 6.3.

Similarly, it is noteworthy that the variance in relative propensities is much greater for pairs of interacting residues at interfaces (Table 6.2) than it is for all (unrestricted) residues at interfaces (Table 6.1). 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 [90], the typical configuration of Arg-Trp is pictured, and similar polar pairings are highlighted, such as Lys-Lys (odds ratio 0.81).

Res. Code	Rel. Prop.	Rel. Freq.	Pair Freq.[90]	Total Abundance [29]	Rim/Core freq. [37]	Homodimer Rim/Core [14]
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 6.3: Amino acids which occur in pairs at interfaces and their relative abundances. Primary data is taken from the indicated references.

6.5 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 [90]. However, the methodology is quite similar to that of the earlier paper [222], 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 [90] is different in definition from exposed and interior residues in the interface in [222], although similar in spirit. Figure 3B of [222] 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 [14] which is very similar in spirit to [37], the difference being the use of homodimers for the study of interfaces. In Table 6.3, we present this data, with the residues reordered to give the rim/core preferences in order for the data in [37] to facilitate comparison with the data in [14]. 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 [14].

The dissection trilogy is completed in [15] in which an attempt is made to determine aminoacid distributions for "nonspecific" interactions. This is intended to be a proxy for any surfaces which might bind however briefly to other protein surfaces. The dataset is determined by looking at crystal contact surfaces in the PDB. We leave as an exercise to compare the data for these surfaces with the other data presented here. See [15] for a comparison with the data in [37] and [14].

Protein-ligand interfaces differ in function, and interfaces with different function can have different composition. In [115], 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 [171].

In [23], 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 [89] (see Section 13.1). The special polar nature of the aromatic residues

(Section 4.5.5) provides the opportunity for interaction with positively charged (cation) residues (Lys, Arg, His). The cation-π motifs play a special role in protein interfaces [44, 244]. The cation-π interaction also has a significant role in α-helix stabilization [207].

A study of the role of evolution on protein interface composition can be found in [33]. In [97, 148], 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 [169, 170].

Identification of individual sidechains that may play the role of 'anchors' in protein-ligand recognition is studied in [190] 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 [40] as a way to discriminate between correct structures and incorrect ones. In [40], 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.

6.6 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.

6.7 Exercises

Exercise 6.1 *Compare the data for the surfaces in [14, 15, 37] by constructing a table analogous to Table 6.3.*

Exercise 6.2 *The aminoacid frequencies for different datasets constitute probability distributions on the set of aminoacids. Different datasets have different distributions. In [15], the distributions for*

nonspecific interaction surfaces are compared with the distributions for other surfaces [37, 14]. 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 6.3 *The frequency of location at interfaces provides a linear ranking (Table 6.1) 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.*

Chapter 7

Wrapping electrostatic bonds

For a protein structure to persist in water, its electrostatic bonds must be shielded from water attack [71, 79, 183, 224]. This can be achieved through wrapping by nonpolar groups (such as CH_n, $n = 1, 2, 3$) in the vicinity of electrostatic bonds to exclude surrounding water [71]. Such desolvation enhances the electrostatic contribution and stabilizes backbone hydrogen bonds [17]. In a nonbonded state, exposed polar amide and carbonyl groups which are well wrapped are hindered from being hydrated and more easily return to the bonded state [45], as depicted in Figure 3.7.

The thermodynamic benefit associated with water removal from pre-formed structure makes under-wrapped proteins adhesive [65, 72, 74]. As shown in [71], under-wrapped hydrogen bonds (UWHB's) are determinants of protein associations. In Section 8.1, 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 enhanced 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 this 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 will be discussed at more length in Chapter 6. As noted in Section 4.5.1, 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 representation 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 7.3) 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

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

Table 7.1: Electronegativity scale [180, 197] of principal atoms in biology. The 'outer electrons' row lists the number of electrons needed to complete the outer shell.

pedagogical reasons, although we would not recommend using it in general.

In our first 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. In Section 7.2.2, we show that this definition is sufficient to give some insight into protein aggregation and 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 7.3. We will show in Section 8.1 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 7.5. In Section 7.5, we will review a more sophisticated technique that incorporates the geometry of nonpolar groups as well as their number to assess the extent of protection via dielectric modulation.

7.1 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 3.2). Similarly, we have noted that certain sidechains, such as glutamine, are polar, even though there is no apparent charge difference in relevant molecules. Here we explain how such polarity can arise due to more subtle differences in charge distribution.

7.1.1 Electronegativity scale

The key to understanding the polarity of certain molecules is the **electronegativity scale** [180, 197], part of which is reproduced in Table 7.1. Atoms with similar electronegativity tend to form nonpolar groups, such as CH_n and $C-S$. Atomic pairs with 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.

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 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 7.1.2, we will see that molecular dynamics codes assign such partial charges. The electronegativity difference for C-O is 0.91, and for N-H it is 0.6. 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.

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. 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

$$\mathcal{E}(O) - \mathcal{E}(C) = 3.66 - 2.75 = 0.91 > 0.60 = 3.19 - 2.59 = \mathcal{E}(N) - \mathcal{E}(H). \tag{7.1}$$

Thus we conclude that the charge difference in the dipole representation of the carbonyl group $(C - O)$ is larger than the charge difference in the dipole representation of the amide $(N - H)$ group.

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 [180]. 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 7.1. More precisely, there is a nearly linear relationship between the electronegativity scale and the number of electrons in the outer shell. The value for hydrogen can be explained by realizing that the outer shell is half full, as it is for carbon.

The atoms with a complete outer shell (helium, neon, argon, etc.) are not part of the electronegativity scale, since they have no room to put 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 halfway from being full and empty of electrons.

7.1.2 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, $n = 1, 2, 3$ are not polar, at least in appropriate contexts. The typically symmetric arrangement of hydrogens also decreases the polarity of a carbonaceous group, at least when the remaining $4-n$ atoms bonded to it are other carbons or atoms of similar electronegativity.

If a carbon is not covalently attached exclusively to carbon or hydrogen then it is likely polarized and carries a partial charge. Thus, C_α carbons 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.

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	CB1, CB2, CG, CD1
Leucine	Leu	L	CB, CG, CD1, CD2
Lysine	Lys	K	CB, CG, CD
Methionine	Met	M	CB
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, CE1, CE2, CZ3, CH2
Tyrosine	Tyr	Y	CB, CG, CD1, CD2, CE1, CE2
Valine	Val	V	CB, CG1, CG2

Table 7.2: PDB codes for nonpolar carbonaceous groups.

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

Table 7.3: Sample PDB codes and nonpolar carbonaceous groups for some nonstandard amino acids and other compounds.

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

Table 7.4: Partial charges from the Gromos force field for polar and negatively charged amino acids.

The case CH_n with $n = 0$ is not encountered in biology unless the carbon is attached to at least one heteroatom.

To illustrate the polarity of the atoms not listed in Table 7.2, we present the partial charges of the remaining atoms as utilized in the Gromos code in Table 7.4 and Table 7.5. In Table 13.1, 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 (7.1).

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 7.2 have zero partial charge.

7.2 Counting residues

In [69], "under-wrapped" was defined in relation to an average native environment. The extent of hydrogen-bond desolvation was defined by the number of residues ρ_R with at least two *nonpolar* carbonaceous groups (CH_n, $n = 1, 2, 3$) whose β-carbon is contained in a specific desolvation domain. In Section 7.1.2, we explained how to determine the polarity of groups using the electronegativity

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 7.5: 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.

scale.

The C_α carbons in all residues 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. However, the CH groups in methionine attached to a sulfur are not polar. Similarly, a lone carbon that is attached to oxygens is also polar. Thus the seven residues listed in Figure 4.4 are eliminated from the group of wrappers.

7.2.1 Desolvation domain

The desolvation domain was chosen in [69] 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 7.1. The choice of the C_α carbons as the centers of the desolvation spheres is justified in Figure 7.2. 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 cutoff distance to evaluate interactions between nearby residues. C_α-carbons which are neighboring in protein sequence are about 3.8Å apart. The distance between other C_α-carbons is easily determined by datamining in the PDB (cf. Exercise 2.2).

An amide-carbonyl hydrogen bond was defined in [69] 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 5.4). At maximum density, water occupies a volume that corresponds to a cube of dimension just over 3.1Å on a side (cf. Section 10.7).

The average extent of desolvation, ρ_R, over all backbone hydrogen bonds of a monomeric struc-

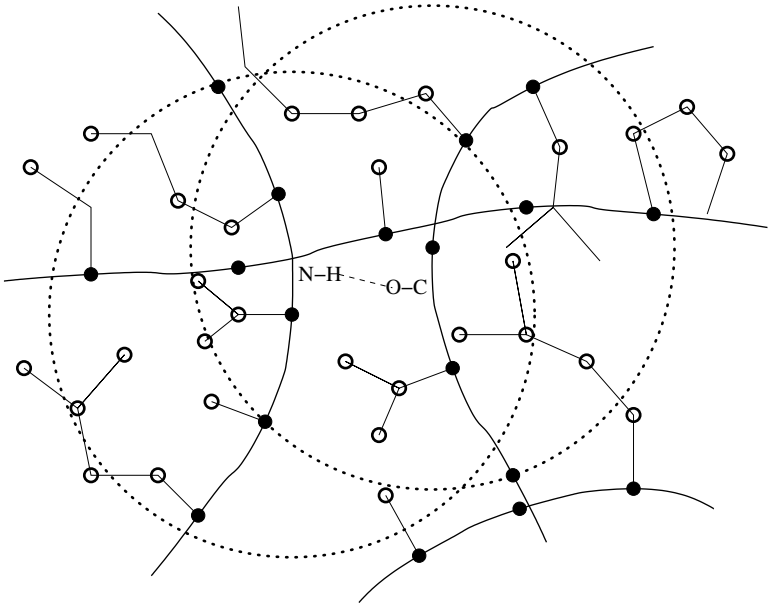


Figure 7.1: Caricature showing desolvation spheres with various side chains. The open circles denote the nonpolar carbonaceous groups, and the solid circles represent the C_α carbons. The hydrogen bond between the amide (N-H) and carbonyl (O-C) groups is shown with a dashed line. Glycines appear without 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.

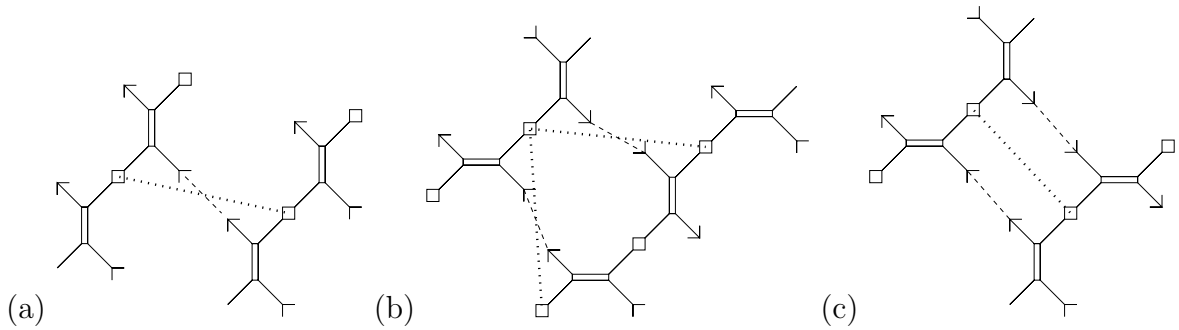


Figure 7.2: The hydrogen bond (dashed line) configuration in (a) α-helix, (b) antiparallel β-sheet, and (c) parallel β-sheet. A dotted line connects the C_α carbons (squares) that provide the centers of the spheres forming the desolvation domains in Figure 7.1. The amide (N-H) groups are depicted by arrow heads and the carbonyl (O-C) groups are depicted by arrow tails.

ture can be computed from any set of structures. In [69], a nonredundant sample of 2811 PDB-structures was examined. The average ρ_R over the entire sample set was 6.6. 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 $\sigma = 1.46$. 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 proteins according to their average extent of hydrogen-bond wrapping is shown in Fig. 5 in [69]. 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.

Figure 2b in [69] depicts the (three) UWHB's for the hemoglobin (Hb) β-subunit (PDB file 1BZ0, chain B). The UWHB's identified by the residue method appear to signal important binding regions [227]. The UWHB (Pro5,Ser9) is adjacent to Glu6 which in sickle cell anemia mutates to Val6 and is located at the Val6-(Phe85, Leu88) interface in the deoxyHbS fiber. The two UWHB's (Glu90,Asp94), (Glu90,Lys95) are associated with the β-FG corner involved in the quaternary $\alpha_1\beta_2$ interface. Thus it would appear that the residue method for defining UWHB's is effective at predicting important binding sites.

In Section 7.2.2, 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 virtues of the residue-based assessment of wrapping, we review additional results and predictions of [69].

7.2.2 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 [104, 129, 156]. Figure 3a of [69] depicts the (many) UWHB's for the human cellular prion protein (PDB file 1QM0) [188, 192, 246]. 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 [188, 192, 246]. 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 [188]. 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 [13, 48, 57, 93, 104, 156, 129] 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 [69]. A statistical analysis involving $\delta_{\max}$ [69] 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 [69] 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 [136]. Meizothrombin is formed during prothrombin activation, and is known to be involved in blood clotting [119] and is able to bind to procoagulant phospholipid membranes [182]. Plasminogen has been identified as being a significant factor in wound healing [195].

7.3 Counting nonpolar groups

A more refined measure of hydrogen-bond protection has been proposed based on the number of vicinal nonpolar groups [65, 71]. 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 7.1. 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_{PG}, is defined by the number of sidechain nonpolar groups (CH_n, $n = 1, 2, 3$) in the desolvation domain.

The distribution of wrapping for a large sample of non-redundant proteins is given in Figure 12.1 for a radius of 6Å for the definition of the desolvation domain. In [72], an UWHB was defined by the inequality $\rho_{PG} < 12$ for this value of the radius. Statistical inferences involving this definition of ρ_{PG} were found to be robust to variations in the range 6.4 ± 0.6 Å for the choice of desolvation radius [71, 79]. In Figure 7.3 the distribution of wrapping is presented for a particular PDB file.

7.3.1 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 7.3 the distribution of wrapping is presented for the antibody complex whose

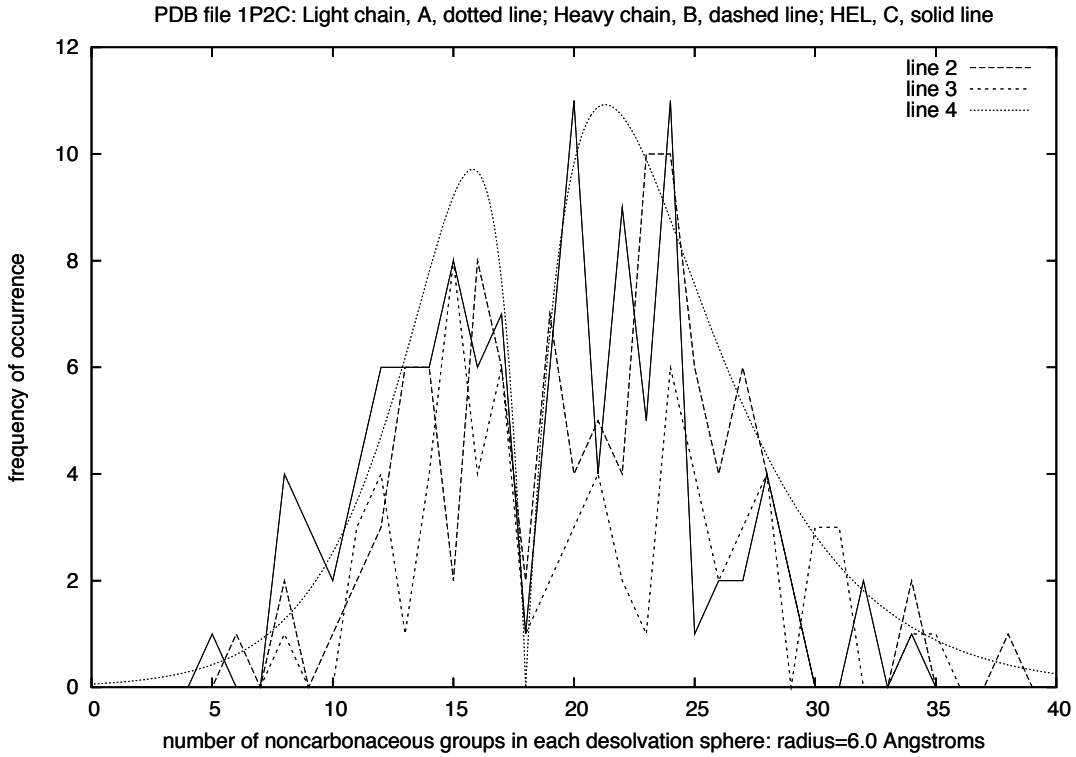


Figure 7.3: Distribution of wrapping for PDB file 1P2C. There are three chains: light, heavy chains of the antibody, and the antigen (HEL) chain. The desolvation radius is 6.0Å. Smooth curves (7.2) are added as a guide to the eye.

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

$$d_i(r) = a_i|r - r_0|e^{-|r-r_0|/w_i} \tag{7.2}$$

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 7.3 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.

7.4 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 the number of carbonaceous groups that are found inside the desolvation domain.

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 $\rho_{PG} = 15.9$. The residues in the former count represent at least two non-polar groups, so we would expect that $\rho_{PG} > 2\rho_R$. 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_{PG} and 7.0Å for ρ_R. Adjusting for this difference would make ρ_{PG} even larger, indicating an even greater discrepancy between the two measures. This implies that ρ_{PG} provides a much finer estimation of local hydrophobicity.

The structural analysis in [69] identified site mutations which might stabilize the part of the cellular prion protein believed to nucleate the cellular-to-scrapie transition. The (134, 159)-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 [69] that a factor that triggers the prion disease is the stabilization of the (134,159) 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 (134,159) 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 (134,159), 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 (134,159) 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 to the right of Thr in Figure 4.4 and is only polar via the backbone polarity. Subsequently, in Table 6.1, we will see that it is reasonable to assign a more refined notion of wrapping for different sidechains, but we do not pursue this here.

7.5 Defining dehydrons via geometric requirements

The enhancement of backbone hydrogen-bond strength and stability depends on the partial structuring, immobilization or removal of surrounding water. In this section we review an attempt [73] to quantify this effect using a continuous representation of the local solvent environment surrounding backbone hydrogen bonds [31, 65, 71, 79, 103, 173, 230]. 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 [65, 79]. 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 [65, 79].

We now describe an attempt to estimate ϵ as a function of the fixed positions $\{r_j \,:\, j = 1, \ldots, n_k\}$ of surrounding hydrophobic groups (in our case, such groups are CH_n, with $n = 1, 2, 3$). The simpler estimates of wrapping considered so far could fail to predict an adhesive site when it is produced by an uneven distribution of desolvators around a hydrogen bond, rather than an insufficient number of such desolvators. Based on the fixed atomic framework for the monomeric structure, we now identify Coulomb energy contributions from intramolecular hydrogen bonds that are most sensitive to local environmental perturbations by subsuming the effect of the perturbations as changes in ϵ.

Suppose that the carbonyl oxygen atom is at $\mathbf{r}_O$ and that the partner hydrogen net charge is at $\mathbf{r}_H$. The electrostatic energy contribution $E_{\text{COUL}}(k, \mathbf{r})$ for this hydrogen bond in a dielectric medium with dielectric permittivity $\epsilon(\mathbf{r})$ is approximated (see Chapter 16) by

$$E_{\text{COUL}}(\mathbf{r}) = \frac{-1}{4\pi\epsilon(\mathbf{r})} \frac{qq'}{|\mathbf{r}_O - \mathbf{r}_H|} \tag{7.3}$$

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 (7.3) 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:

$$\nabla_R(1/\epsilon(\mathbf{r})) = \frac{4\pi|\mathbf{r}_O - \mathbf{r}_H|}{qq'} \left(-\nabla_R E_{\text{COUL}}(\mathbf{r})\right) = \frac{4\pi|\mathbf{r}_O - \mathbf{r}_H|}{qq'} F(\mathbf{r}), \tag{7.4}$$

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 [65], 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 introduced [73] 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 7.1. In [73], a radius of 7Å was used. The set of vector positions of the n_k hydrophobic groups surrounding the hydrogen bond is extended from $\{\mathbf{r}_j \,:\, j = 1, 2, \ldots, n_k\}$ to $\{\mathbf{r}_j \,:\, j = 1, 2, \ldots, n_k; R\}$ 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

$$M_k = \max\left\{|\nabla_R(1/\epsilon)|_{R=R_o}| \;:\; R_o \in C\right\}. \tag{7.5}$$

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 at the surface of the desolvation domain.

The quantity M_k may be interpreted in physical terms as a measure of the maximum possible attractive force exerted on the test hydrophobic group by the pre-formed hydrogen bond. The only difficulty in estimating M_k is that it requires a suitable model of the dielectric permittivity ϵ as a function of the geometry of surrounding hydrophobic groups. We will consider the behavior of the dielectric permittivity more carefully in Chapter 16, but for now we consider a heuristic model used in [73].

The model in [73] for the dielectric may be written

$$\epsilon^{-1} = (\epsilon_o^{-1} - \epsilon_w^{-1})\Omega(\{\mathbf{r}_j\})\Phi(\mathbf{r}_H - \mathbf{r}_O) + \epsilon_w^{-1}, \tag{7.6}$$

where ϵ_w and ϵ_o are the permittivity coefficients of bulk water and vacuum, respectively, and

$$\Omega(\{\mathbf{r}_j\}) = \prod_{j=1,\ldots,n_k} \left(1 + e^{-|\mathbf{r}_O - \mathbf{r}_j|/\Lambda}\right)\left(1 + e^{-|\mathbf{r}_H - \mathbf{r}_j|/\Lambda}\right) \tag{7.7}$$

provides an estimate of the change in permittivity due to the hydrophobic effects of the carbonaceous groups. In [73], 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

$$\Phi(\mathbf{r}) = (1 + |\mathbf{r}|/\xi)\, e^{-|\mathbf{r}|/\xi}, \tag{7.8}$$

where $|\cdot|$ denotes the Euclidean norm and $\xi = 5$Å is a water dipole-dipole correlation length, approximates the effect of hydrogen bond length on its strength [73].

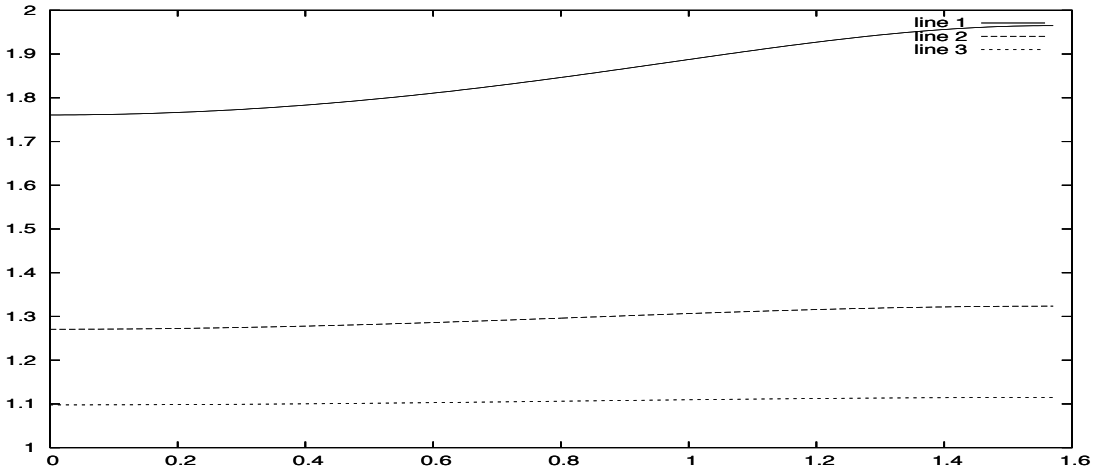


Figure 7.4: The function $\omega(x, y)$ plotted as a function of angle from the perpendicular bissector of the axis connecting $\mathbf{r}_H$ and $\mathbf{r}_O$, for three different values of the distance r from the midpoint between $\mathbf{r}_H$ and $\mathbf{r}_O$: $r = 1$ (solid line), $r = 2$ (dashed line), $r = 3$ (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 (7.7) as

$$\Omega(\{\mathbf{r}_j\}) = \prod_{j=1,\ldots,n_k} \omega(\mathbf{r}_j), \tag{7.9}$$

where the function ω is defined by

$$\omega(\mathbf{r}) = \left(1 + e^{-|\mathbf{r}_O - \mathbf{r}|/\Lambda}\right)\left(1 + e^{-|\mathbf{r}_H - \mathbf{r}|/\Lambda}\right). \tag{7.10}$$

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 7.4 as a function of angle from the perpendicular bissector of the axis connecting $\mathbf{r}_H$ and $\mathbf{r}_O$, for three different values of the distance r from the midpoint between $\mathbf{r}_H$ and $\mathbf{r}_O$.

We see that the deviation in ω provides a strong angular dependence on the dielectric coefficient in this model. Thus hydrophobes close to the plane bissecting 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. When $\Omega = 1$, we get $\epsilon = \epsilon_o$ reflecting the maximal amount of water exclusion possible. Bigger values of Ω correspond to the effect of underwrapping.

The computation of M_k involves computing the gradient of

$$\Omega(\{\mathbf{r}_1, \ldots, \mathbf{r}_{n_k}, R\}) = \omega(R)\Omega(\{\mathbf{r}_1, \ldots, \mathbf{r}_{n_k}\}) \tag{7.11}$$

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 (7.6).

A sensitivity threshold for hydrogen bonds was established in [73] 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 [102], were found to be highly sensitive in the sense that

$$M_k > \lambda/10, \tag{7.12}$$

where λ was defined to be

$$\lambda = \frac{\epsilon_o^{-1} - \epsilon_w^{-1}}{2\text{Å}}. \tag{7.13}$$

On the other hand, 91.6% of backbone hydrogen bonds were found to be relatively insensitive to water removal, namely,

$$0 < M_k < \lambda/100 \tag{7.14}$$

This remarkable separation in the (nearly bimodal) distribution of sensitivities led [73] to the definition of a dehydron as a backbone hydrogen bond satisfying (7.12).

7.6 Exercises

Exercise 7.1 *It was predicted [69] 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 7.2 *In Section 4.4.3, we noted that sidechains have different conformations. Determine the number of different rotameric states possible for each sidechain (hint: read [146]). Compare the number of rotameric degrees of freedom for the seven residues listed in Figure 4.4 with the remaining group of thirteen sidechains.*

Exercise 7.3 *In Figure 7.2(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.*

Chapter 8

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 contribute to this conclusion. One of them involves a mesoscopic measurement of the force associated with a dehydron [72]. A second presents data on the direct measurement of the dehydronic force using atomic force microscopy [63]. Another paper examines the effect of such a force on a deformable surface [74].

8.1 Surface adherence force

We defined the notion of an under-wrapped hydrogen bond by a simple counting method in Chapter 7 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.

8.1.1 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 4.4.4.

The volume change causes self-assembly of lipids and provides a substantial pressure that holds the surface together. The architecture of a lipid bilayer is extremely adaptive. For example, a curved surface can be formed simply by allocating more lipid to one side than the other. Moreover, it easily allows insertion of other molecules of complex shape but with other composition. Much of a cell membrane is lipid, but there are also proteins with various functions as well as other molecules such as cholesterol. However, a simple lipid bilayer provides a useful model biological surface.

8.1.2 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 [66] 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 [66] indicated that more dehydrons lead to more attachments, strongly suggesting that dehydrons are sticky. However, such a conclusion was only qualitative.

A more refined analysis of lipid bilayer experiments was able to quantify a force of attachment [72]. 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 [72], as shown in Table 8.1, 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 [72]. The particular surface was a Langmuir-Blodgett bilayer made of the lipid DLPC (1,2 dilauroyl-sn-glycero-3 phosphatidylcholine) [194].

8.2 A two-zone model

In [72], 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 of arrival is dependent on the particular experiment, so we only summarize the model results from [72]. The second zone is the fluid boundary layer close to the lipid bilayer, where binding can occur. In this layer, the probability 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

$$M = \Phi P(n_{UW}, n_W, T)N \tag{8.1}$$

where Φ 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 fraction Φ depends on details of the experimental design, so we focus on on the second term P.

8.2.1 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

$$f \approx \frac{n_{UW}}{n_{UW} + n_W}. \tag{8.2}$$

Then the probability of attachment at a dehydron is predicted by thermodynamics as

$$P(n_{UW}, n_W, T) = \frac{f e^{\Delta U/k_B T}}{(1-f)e^{\Delta V/k_B T} + f e^{\Delta U/k_B T}} \approx \frac{n_{UW} e^{\Delta U/k_B T}}{n_W e^{\Delta V/k_B T} + n_{UW} e^{\Delta U/k_B T}}, \tag{8.3}$$

with k_B = Boltzmann's constant. In [72], ΔV was assumed to be zero. In this case, (8.3) simplifies to

$$P(n_{UW}, n_W, T) = \frac{f e^{\Delta U/k_B T}}{(1-f) + f e^{\Delta U/k_B T}} \approx \frac{n_{UW} e^{\Delta U/k_B T}}{n_W + n_{UW} e^{\Delta U/k_B T}} \tag{8.4}$$

(cf. equation (2) of [72]). Note that this probability is lower if $\Delta V > 0$.

8.2.2 Diffusion zone model

The probability Φ in (8.1) of penetrating the bottom layer of the fluid is estimated in [72] by a model for diffusion via Brownian motion in the plane orthogonal to the flow direction. This depends on the molecular mass, m, the solvent bulk viscosity μ, and the **hydrodynamic radius** [198] or **Stokes radius** [100]. 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 [53, 130] is very accurate, since the variation in flow characteristics due to shape variation is quite small [198]. 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

$$\xi = 6\pi R\mu/m = F/mv \tag{8.5}$$

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

$$\alpha = \frac{m\xi^2 L^2}{2k_B T} = \frac{L^2(6\pi R\mu)^2/m}{2k_B T}, \tag{8.6}$$

which has units of energy in numerator and denominator. We have [2]

$$\begin{aligned}\Phi(v,R,m) &= \int_\Lambda \int_{\Omega\setminus\Lambda} \int_{[0,\tau]} \frac{\alpha L^{-2}}{\pi\Gamma(t)} e^{-\alpha L^{-2}|\mathbf{r}-\mathbf{r}_0|^2/\Gamma(t)}\, dt d\mathbf{r}_0 d\mathbf{r} \\ &= \int_{\tilde\Lambda} \int_{\tilde\Omega\setminus\tilde\Lambda} \int_{[0,L/v]} \frac{\alpha}{\pi\Gamma(t)} e^{-\alpha|\tilde{\mathbf{r}}-\tilde{\mathbf{r}}_0|^2/\Gamma(t)}\, dt d\tilde{\mathbf{r}}_0 d\tilde{\mathbf{r}}\end{aligned} \tag{8.7}$$

where $\mathbf{r}$ is the two-dimensional position vector representing the cell cross-section Ω, $|\mathbf{r}|$ denotes the Euclidean norm of $\mathbf{r}$, Λ is the 6Å×10^8Å cross-section of the bottom layer, and $\Gamma(t) = 2\xi t - 3 + 4e^{-\xi t} - e^{-2\xi t}$. The domains $\tilde\Lambda$ and $\tilde\Omega$ represents 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.

8.2.3 Model validity

The validity of the model represented by equations (8.1—8.7) 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 [72]) across the entire range of flow velocities v. This value is

$$\Delta U = 3.91 \pm 0.67\text{kJ/mole}. \tag{8.8}$$

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 (8.8) of the binding energy for a dehydron, an estimate was made [72] of the force

$$|F| = 7.78 \pm 1.5\text{pN} \tag{8.9}$$

exerted by the surface on a single protein molecule at a 6Å distance from the dehydron.

8.3 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 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 [63] 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

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 8.1: Six proteins and their hydrogen bond distributions. WWHB=well-wrapped hydrogen bonds.

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

The force-displacement curve provided by the AFM have similarities for both monolayers [63]. 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 [63] of

$$5.9 \pm 1.2\text{pN} \tag{8.10}$$

at a distance of 5Å is in remarkably close agreement with the estimate (8.9) of 7.78 ± 1.5pN at a distance of 6Å in [72]. Part of the discrepancy could be explained by the fact that in [72] 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 [72] would be reduced.

8.4 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 [205] of the inner membranes of cellular or subcellular compartments containing soluble proteins [74]. These vary from simple bag-like membranes [56] (e.g., erythrocytes, a.k.a. red blood cells) to highly invaginated membranes [227] (e.g., mitochondrial inner membranes). This raises the question of what might be causing the difference in membrane structure [126, 138, 164, 229].

Some evidence [74] 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 [74].

8.4.1 Protein adsorption

Morphology induction was tested in fluid phospholipid (DLPC) bilayers (Section 8.1) coating an optical waveguide [74]. 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 [191, 217]. 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.

8.4.2 Density of invaginations

The density of invaginations correlates with the extent of wrapping, ρ, of the soluble protein structure (Fig. 1, 2a in [74]). 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 [74]) shows that protein aggregation is a competing effect in the protection of solvent-exposed hydrogen bonds ([71, 65, 66, 60, 79]): for each protein there appears to be a concentration limit beyond which aggregation becomes more dominant.

8.5 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 [66]. The rate of change of curvature $\frac{\partial g}{\partial t}$ is an increasing function of the force f:

$$\frac{\partial g}{\partial t} = \phi(f) \tag{8.11}$$

for some increasing function ϕ. Note that $\phi(0) = 0$: 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 8.2 the development of a more complete model.

Assuming equilibrium is attained rapidly, we can assert that the force f is proportional to $p(g)$:

$$f = cp(g) \tag{8.12}$$

at least up to some saturation limit, which we discuss subsequently. If we wish to be conservative, we can assert only that $f = \psi(p(g))$ with ψ increasing. In any case, we conclude that f may be regarded as an increasing function of the curvature g, say

$$F(g) := \phi(\psi(p(g))). \tag{8.13}$$

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 $F(0) = 0$.

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 [74]). 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

$$\frac{\partial g}{\partial t} = F(g), \tag{8.14}$$

where F is the function in (8.13) 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:

$$\frac{\partial g}{\partial t} = sg. \tag{8.15}$$

Figure 4 in [74] 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 [74]) 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.

8.6 Exercises

Exercise 8.1 *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 8.2 *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 (8.13) that incorporate the effect of decreasing surface area, and show how it would lead to a model like (8.14) which would saturate (rather than grow exponentially forever), reflecting the crowding effect of the molecules on the lipid surface.*

Chapter 9

Electronic force details

In Section 3.2, we introduced the basic electronic interactions. Here we look at them in more detail. The basic electronic entities are groups of charges that are constrained to be together, such as dipoles. In Section 9.2.1 we study dipole-dipole interactions. In Section 9.2.2, 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 9.2.3.

There is a natural hierarchy of charged groups. These can be ranked by the rate of decay of their potentials, and thus by their globality. At the highest 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, involve 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.

9.1 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 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 $\left\{\mathbf{r} \in I\!R^3 \ : \ R-1 \leq |\mathbf{r}| < R\right\}$ 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

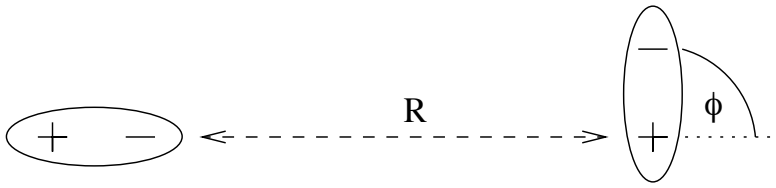


Figure 9.1: Dipole-dipole (in-line) interaction configuration.

do that they are uniformly distributed). Then the total force would be proportional to

$$\sum_{R=1}^{R_{\text{mac}}} R^{2-n} \tag{9.1}$$

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

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: $\mathcal{O}(1/R_{\text{cut}})$ 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 carefully handle electrostatic interactions to compute the forces accurately.

9.2 Modeling interactions among polar and charged residues

We have seen that certain bonds can be modeled by simple interactions by charge groups. For example, polar groups can be modeled simply by placing partial charges appropriately at atom centers, as described in Section 7.1.2. Here we investigate in detail the angular dependence of these models.

9.2.1 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 9.1. 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 of 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 $(\cos\phi, \sin\phi)$.

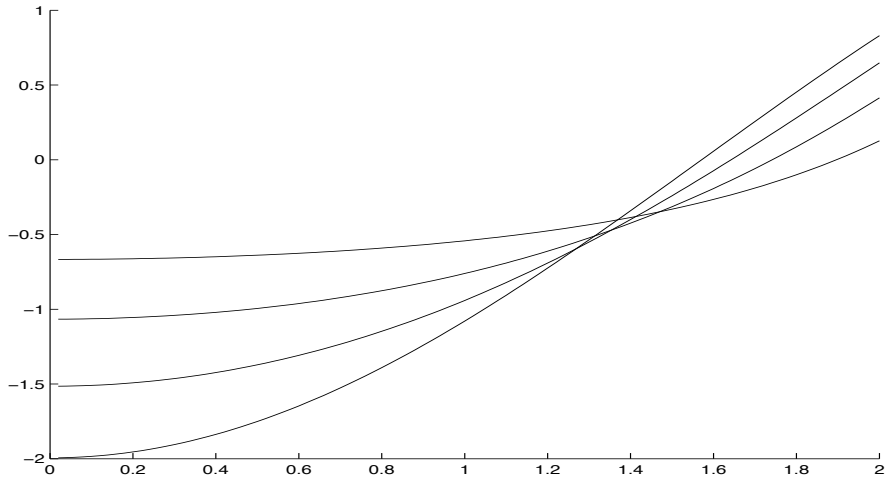


Figure 9.2: Dipole-dipole (in-line) interaction energy, scaled by R^3, for $R = 2, 4, 10, 1000$. Horizontal ϕ-axis measured in radians. The flattest curve corresponds to $R = 2$.

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

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R, 0)| &= \sqrt{(R + \cos\phi)^2 + \sin^2\phi} \\ &= \sqrt{1 + R^2 + 2R\cos\phi} \end{aligned} \tag{9.2}$$

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

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R - 1, 0)| &= \sqrt{(1 + R + \cos\phi)^2 + \sin^2\phi} \\ &= \sqrt{1 + (R+1)^2 + 2(R+1)\cos\phi} \end{aligned} \tag{9.3}$$

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

$$\begin{aligned} \frac{1}{R+1} - \frac{1}{R} + &\frac{1}{\sqrt{1 + R^2 + 2R\cos\phi}} \\ &- \frac{1}{\sqrt{1 + (R+1)^2 + 2(R+1)\cos\phi}} \end{aligned} \tag{9.4}$$

A plot of the interaction energy (9.4) is given in Figure 9.2 as a function of ϕ for various values of R. Since we know (cf. (3.5)) that the interaction energy will decay like R^{-3}, we have scaled the energy in Figure 9.2 by R^3 to keep the plots on the same scale. The value of $R = 1000$ indicates the asymptotic behavior; see Exercise 9.1 for the analytical expression. Indeed, there is little difference between $R = 100$ 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.

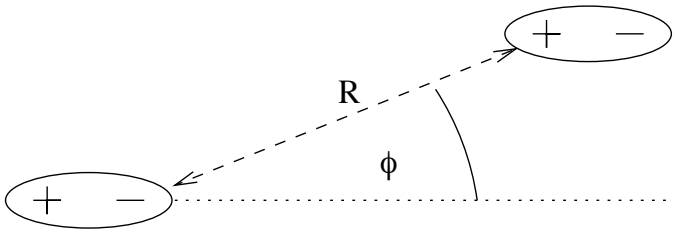


Figure 9.3: Dipole-dipole (parallel) interaction configuration.

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 9.3. 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 of the positive charge on the right and the negative charge on the left is R. Thus the charge centers of the dipole on the right are at $R(\cos\phi, \sin\phi)$ (positive charge) and $(1 + R\cos\phi, R\sin\phi)$ (negative charge).

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

$$|(1 + R\cos\phi, R\sin\phi)| = \sqrt{1 + R^2 + 2R\cos\phi}. \tag{9.5}$$

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

$$|(2 + R\cos\phi, R\sin\phi)| = \sqrt{4 + R^2 + 4R\cos\phi}. \tag{9.6}$$

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

$$-\frac{1}{R} + \frac{2}{\sqrt{1 + R^2 + 2R\cos\phi}} - \frac{1}{\sqrt{4 + R^2 + 4R\cos\phi}} \tag{9.7}$$

A plot of the interaction energy (9.7) is given in Figure 9.4 as a function of ϕ for various values of R. Since we know (cf. (3.5)) that the interaction energy will decay like R^{-3}, we have scaled the energy in Figure 9.2 by R^3 to keep the plots on the same scale. The value of $R = 1000$ indicates the asymptotic behavior; see Exercise 9.2 for the analytical expression. Indeed, there is little difference between $R = 100$ and $R = 1000$. The flatter curve is the smallest value of R (=2) and shows only limited angular dependence.

Two-parameter interaction configuration

Now we consider the effect of a dual angular orientation on the strength of interaction of two dipoles. Suppose we have two dipoles as indicated in Figure 9.5. The exact positions of the charges are as

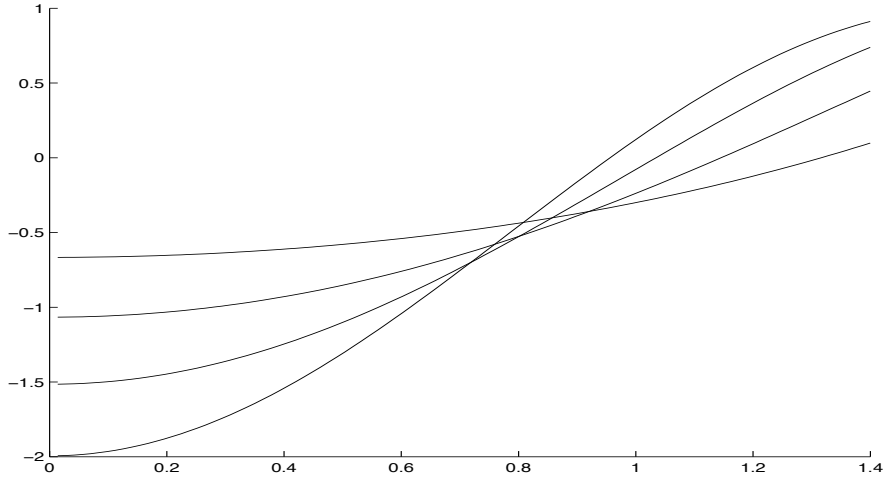


Figure 9.4: Dipole-dipole (parallel) interaction energy, scaled by R^3, for $R = 2, 4, 10, 1000$. Horizontal ϕ-axis measured in radians.

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 of 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

$$|(1 + R\cos\theta, R\sin\theta)| = \sqrt{1 + R^2 + 2R\cos\theta} \tag{9.8}$$

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

$$\begin{aligned} |R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi)| =& \sqrt{(R\cos\theta + \cos\phi)^2 + (R\sin\theta + \sin\phi)^2}| \\ =& \sqrt{(R^2 + 1 + 2R(\cos\theta\cos\phi + \sin\theta\sin\phi)}| \end{aligned} \tag{9.9}$$

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

$$\begin{aligned} |R(\cos\theta, \sin\theta) + (\cos\phi, \sin\phi) - (-1, 0)|^2 =& |(1 + R\cos\theta + \cos\phi, R\sin\theta + \sin\phi)|^2 \\ =& (1 + \cos\phi)^2 + 2R\cos\theta(1 + \cos\phi) + R^2 + 2R\sin\theta\sin\phi + \sin^2\phi \\ =& 2(1 + \cos\phi) + 2R\cos\theta(1 + \cos\phi) + R^2 + 2R\sin\theta\sin\phi \\ =& 2(1 + \cos\phi)(1 + R\cos\theta) + R^2 + 2R\sin\theta\sin\phi \end{aligned} \tag{9.10}$$

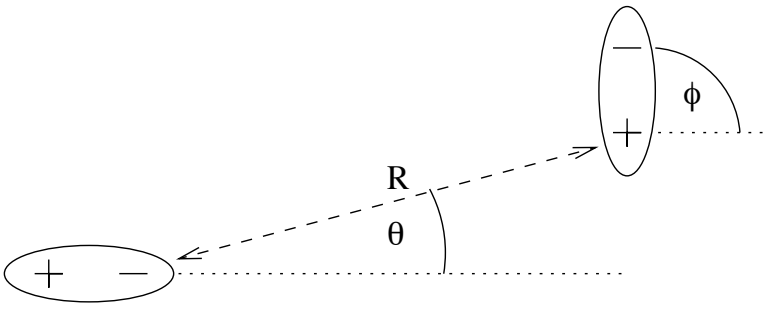


Figure 9.5: Dipole-dipole (two-angle) interaction configuration.

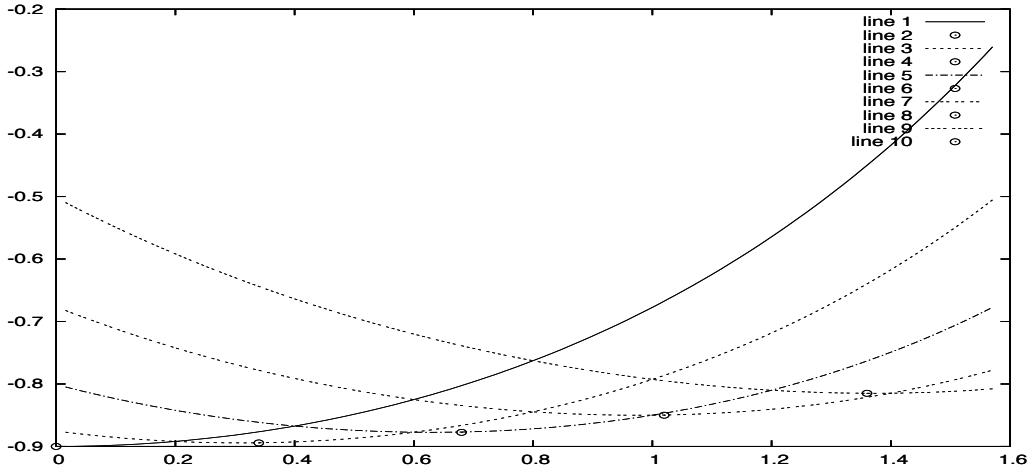


Figure 9.6: Dipole-dipole (two-angle) interaction energy, scaled by R^3, for $R = 3$, as a function of ϕ for various fixed values of $\theta = 0, 0.2, 0.4, 0.6, 0.8$. Minimum values of the energy are plotted as circles at the points $\phi = 1.7\theta$. Horizontal ϕ-axis measured in radians.

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

$$\begin{aligned} &-\frac{1}{R} + \frac{1}{\sqrt{1+R^2+2R\cos\theta}} \\ &-\frac{1}{\sqrt{R^2+2(1+\cos\phi)(1+R\cos\theta)+2R\sin\theta\sin\phi}} \\ &+\frac{1}{\sqrt{1+R^2+2R(\cos\theta\cos\phi+\sin\theta\sin\phi)}} \end{aligned} \tag{9.11}$$

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 a function of ϕ for various values of θ in Figure 9.6. 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 $\phi = \theta$ 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 (9.11) simplifies to

$$\begin{aligned} &-\frac{1}{R} + \frac{1}{\sqrt{1+R^2}} \\ &-\frac{1}{\sqrt{R^2 + 2(1+\cos\phi) + 2R\sin\phi}} \\ &+\frac{1}{\sqrt{1+R^2+2R\sin\phi}} \end{aligned} \tag{9.12}$$

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 (9.12) does not occur at $\phi = \pi$, due to the asymmetry of the expression around this value. We leave as Exercise 9.4 to plot (9.12) 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 9.7, 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 $R \to \infty$.

The energy, again scaled by R^3, at the optimal value of ϕ is plotted as a function of θ in Figure 9.8. 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).

9.2.2 Charge-dipole interactions

Charge-dipole interactions are simpler to analyze, and we have already anticipated their asymptotic strength in (3.2). On the other hand, this forms a very important class of interactions. Although mainchain-mainchain interactions do not involve such pairs, all of the three other interactions among sidechains and mainchains can occur. In addition, more complex interactions, such as

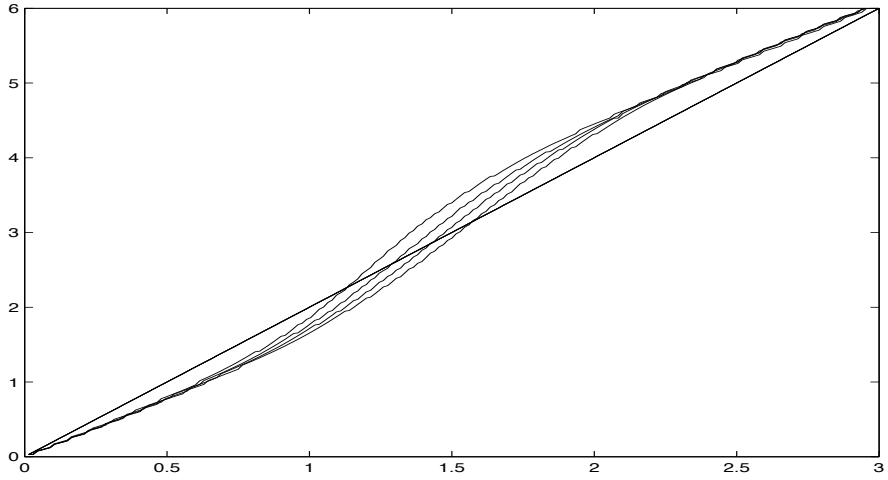


Figure 9.7: Optimal ϕ angle (minimum interaction energy) as a function of θ for the dipole-dipole (two-angle) interaction, for $R = 3, 5, 10, 1000$ (the left-most curve corresponds to $R = 3$, and they move to the right with increasing R). Horizontal θ-axis and vertical ϕ-axis are measured in radians. The line $\phi = 2\theta$ has been added as a guide.

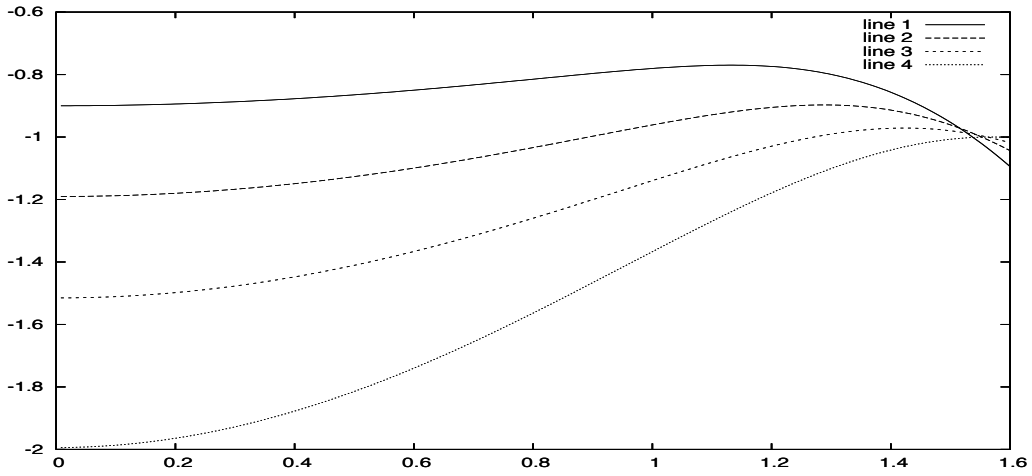


Figure 9.8: Dipole-dipole (two-angle) interaction energy minimum, scaled by R^3, for $R = 3, 5, 10, 1000$ (top to bottom), as a function of θ. Horizontal θ-axis measured in radians. Plotted is the energy at the optimal value of ϕ that minimizes the energy as a function of ϕ for fixed θ.

cation-π interactions (Section 13.1) 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. Assume that the negative charge of the dipole is at the origin and that the single charge is positive, located at $(r\cos\theta, r\sin\theta, 0)$. We choose scales such that the charges of the dipole are of unit size and the positive charge of the dipole is at $(-1, 0, 0)$. If a is the charge of the single charge, then the interaction energy of the system is

$$V(r,\theta) = -\frac{a}{r} + \frac{a}{\sqrt{(1+r\cos\theta)^2 + r^2\sin^2\theta}} \tag{9.13}$$

We leave as Exercise 9.8 to show that

$$V(r,\theta) \approx -\frac{a\cos\theta}{r^2} \tag{9.14}$$

for large r and fixed θ.

We will be interested in the force field that the dipole exerts on the charge as well. It is easier to compute the gradient of V in Cartesian coordinates (note that we can ignore the z direction in our computations):

$$V(x,y) = -\frac{a}{\sqrt{x^2+y^2}} + \frac{a}{\sqrt{(1+x)^2+y^2}} \tag{9.15}$$

To improve readability, we will use the notation $[x, y]$ to denote the vector with components x and y. Similarly, we will use $r = \sqrt{x^2+y^2}$ to reduce bookkeeping. Thus we find

$$\begin{aligned}
\nabla V(x,y) =& \frac{a[x,y]}{(x^2+y^2)^{3/2}} - \frac{a[1+x,y]}{((1+x)^2+y^2)^{3/2}} \\
=& \frac{a[x,y]\left((1+2x+r^2)^{3/2} - r^3\right)}{r^3(1+2x+r^2)^{3/2}} - \frac{a[1,0]}{(1+2x+r^2)^{3/2}} \\
=& a\frac{[x,y]\left((r^{-2}(1+2x)+1)^{3/2}-1\right) - [1,0]}{(1+2x+r^2)^{3/2}}.
\end{aligned} \tag{9.16}$$

This expression can be used to evaluate the force field on a charge in a dipole. For example, for $r = 2$ we find

$$\nabla V_{r=2}(\theta) = a\frac{[\cos\theta, \sin\theta]\left((\frac{5}{4}+\cos\theta)^{3/2}-1\right) - [\frac{1}{2}, 0]}{4(\frac{5}{4}+\cos\theta)^{3/2}}. \tag{9.17}$$

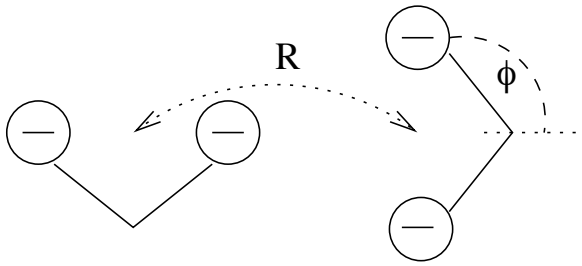


Figure 9.9: Charge-charge interaction configuration similar to what is found in an interaction between Asp and Glu.

We can approximate (9.16) for large r as

$$\begin{aligned}
\nabla V(x,y) \approx & a\frac{[x,y]\left(r^{-2}(\frac{3}{2}+3x)\right)-[1,0]}{(1+2x+r^2)^{3/2}} \\
=& a\frac{[x,y]\left(\frac{3}{2}+3x\right)-[r^2,0]}{r^2(1+2x+r^2)^{3/2}} \\
=& a\frac{[\frac{3}{2}x+2x^2-y^2,\frac{3}{2}y+3xy]}{r^2(1+2x+r^2)^{3/2}} \\
\approx& a\frac{[2x^2-y^2,3xy]}{r^5} \\
=& ar^{-3}[2\cos^2\theta-\sin^2\theta,3\cos\theta\sin\theta] \\
=& \tfrac{1}{2}ar^{-3}[1+3\cos 2\theta,3\sin 2\theta].
\end{aligned} \tag{9.18}$$

Finally, we recall that these calculations are fully valid in three dimensions, so we have derived expressions valid for all z as well. In all cases, the z-component of ∇V is zero.

9.2.3 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 9.9. 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 dipole pair (assuming unit charges) depend on the distances

$$\begin{aligned}
r_{++} =& |(\cos\phi,\sin\phi)-(R+1,0)| \\
r_{-+} =& |-(\cos\phi,\sin\phi)-(R+1,0)| \\
r_{+-} =& |(\cos\phi,\sin\phi)-(R-1,0)| \\
r_{--} =& |-(\cos\phi,\sin\phi)-(R-1,0)|
\end{aligned} \tag{9.19}$$

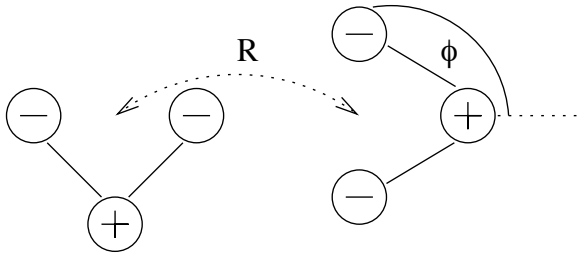


Figure 9.10: Charge-charge interaction configuration similar to what is found in an interaction between Asp and Glu, but with a more refined model.

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

$$\begin{aligned} r_{++}^2 &= (C - R - 1)^2 + S^2 = 1 - 2\cos\phi(R+1) + (R+1)^2 \\ r_{-+}^2 &= (C + R + 1)^2 + S^2 = 1 + 2\cos\phi(R+1) + (R+1)^2 \\ r_{+-}^2 &= (C - R + 1)^2 + S^2 = 1 - 2\cos\phi(R-1) + (R-1)^2 \\ r_{--}^2 &= (C + R - 1)^2 + S^2 = 1 + 2\cos\phi(R-1) + (R-1)^2 \end{aligned} \tag{9.20}$$

Thus we can write this succinctly as

$$r_{\pm_1\pm_2} = \sqrt{1 \pm_1 2\cos\phi(R \pm_2 1) + (R \pm_2 1)^2}. \tag{9.21}$$

Thus the energy (of repulsion) for the charge groups is

$$\frac{1}{r_{++}} + \frac{1}{r_{-+}} + \frac{1}{r_{+-}} + \frac{1}{r_{--}} \tag{9.22}$$

and we seek to find the value of ϕ that minimizes it. We leave as an exercise to plot the expression in (9.22) 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 given in Figure 9.10. We leave it as an exercise to investigate the minimum energy configuration. For example, we could assume a positive charge on the left at $(-R, -1)$ and on the right at $(\sin\phi, -\cos\phi)$.

9.3 General form of a charge group

The general form of a potential for a charged system can be written as a sum of point charge potentials

$$V(\mathbf{r}) = \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r} - \mathbf{r}_k|}, \tag{9.23}$$

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

$$W(\mathbf{r}) = 1/|\mathbf{r}| \tag{9.24}$$

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

$$(T_{\mathbf{x}}f)(\mathbf{r}) = f(\mathbf{r}-\mathbf{x}) \tag{9.25}$$

for any function f. Then we can interpret the expression (9.23) as

$$V = \sum_{k=1}^{K} q_k T_{\mathbf{r}_k} W. \tag{9.26}$$

In view of (9.26), we define the operator

$$\mathcal{D} = \sum_{k=1}^{K} q_k T_{\mathbf{r}_k}. \tag{9.27}$$

We will see that this corresponds to a difference operator when the net charge of the system is zero.

9.3.1 Asymptotics of general potentials

The decay of $V(\mathbf{r})$ for simple dipoles can be determined by algebraic manipulations as in Section 3.2. However, for more complex arrangements, determining the rate is quite complicated. Multipole expansions such as in Section 15.5.1 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

$$V_\epsilon(\mathbf{r}) = \sum_{k=1}^{K} \frac{q_k}{|\mathbf{r}-\epsilon\mathbf{r}_k|}. \tag{9.28}$$

There is a dual relationship between the asymptotics of $V_\epsilon(\mathbf{r})$ as $\mathbf{r}\to\infty$ and $\epsilon\to 0$, as follows:

$$V_\epsilon(\mathbf{r}) = |\mathbf{r}|^{-1} V_{\epsilon/|\mathbf{r}|}\left(|\mathbf{r}|^{-1}\mathbf{r}\right). \tag{9.29}$$

The proof just requires changing variables in (9.28):

$$V_\epsilon(\mathbf{r}) = \frac{1}{|\mathbf{r}|}\sum_{k=1}^{K} \frac{q_k}{|\mathbf{r}|^{-1}|\mathbf{r}-\epsilon\mathbf{r}_k|} = |\mathbf{r}|^{-1} V_{\epsilon/|\mathbf{r}|}\left(|\mathbf{r}|^{-1}\mathbf{r}\right). \tag{9.30}$$

In particular, we have the simplified form

$$V(\mathbf{r}) = V_1(\mathbf{r}) = |\mathbf{r}|^{-1} V_{|\mathbf{r}|^{-1}}\left(|\mathbf{r}|^{-1}\mathbf{r}\right) = \epsilon V_\epsilon(\omega), \tag{9.31}$$

where $\epsilon = |\mathbf{r}|^{-1}$ 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 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 (9.27), we define

$$\mathcal{D}_\epsilon = \sum_{k=1}^{K} q_k T_{\epsilon \mathbf{r}_k}, \tag{9.32}$$

and observe from (9.23) and (9.24) that

$$V_\epsilon = \mathcal{D}_\epsilon W. \tag{9.33}$$

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

$$\lim_{\epsilon \to 0} \epsilon^{-k} \mathcal{D}_\epsilon = \mathcal{D}_0 \tag{9.34}$$

where $\mathcal{D}_0$ is a differential operator of order k. The convergence in (9.34) is (at least) weak convergence, in the sense that for any smooth function f in a region $\Omega \subset I\!R^3$,

$$\lim_{\epsilon \to 0} \epsilon^{-k} \mathcal{D}_\epsilon f(\mathbf{x}) = \mathcal{D}_0 f(\mathbf{x}) \tag{9.35}$$

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

$$\lim_{\epsilon \to 0} \epsilon^{-k} V_\epsilon = V_0, \tag{9.36}$$

where the limiting potential is defined by

$$V_0(\mathbf{r}) = \mathcal{D}_0 W(\mathbf{r}). \tag{9.37}$$

Applying (9.31), (9.36), and (9.37), we conclude that

$$V(\mathbf{r}) \approx \frac{1}{|\mathbf{r}|^{k+1}} \mathcal{D}_0 W(|\mathbf{r}|^{-1}\mathbf{r}), \tag{9.38}$$

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

$$\epsilon^{-k} \mathcal{D}_\epsilon \phi(\mathbf{r}) = \mathcal{D}_0 \phi(\mathbf{r}) + \mathcal{O}(\epsilon) \tag{9.39}$$

in which case we can assert that

$$V(\mathbf{r}) = \frac{1}{|\mathbf{r}|^{k+1}} \mathcal{D}_0 W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-k-2}). \tag{9.40}$$

9.3.2 Application of (9.40)

Let us show how (9.40) 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

$$\mathcal{D}_\epsilon \phi(x, y, z) = \phi(x + \tfrac{1}{2}\epsilon, y, z) - \phi(x - \tfrac{1}{2}\epsilon, y, z). \tag{9.41}$$

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

$$\psi(x + \xi) - \psi(x - \xi) = 2\xi\psi'(x) + \tfrac{1}{3}\xi^3\psi^{(3)}(x) + \mathcal{O}(\xi^5). \tag{9.42}$$

Applying (9.42) to $\psi(x) = \phi(x, y, z)$, we have

$$\mathcal{D}_\epsilon \phi(x, y, z) = \epsilon \frac{\partial}{\partial x}\phi(x, y, z) + \mathcal{O}(\epsilon^3). \tag{9.43}$$

Taking limits, we see that

$$\epsilon^{-1}\mathcal{D}_\epsilon \to \frac{\partial}{\partial x} \tag{9.44}$$

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 3.2. More precisely, applying (9.40) we have

$$V(\mathbf{r}) = |\mathbf{r}|^{-2}\frac{\partial}{\partial x} W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-3}). \tag{9.45}$$

9.4 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 9.3. We begin with a simple case.

9.4.1 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 $(a + 1, 0, 0)$ and $(-a - 1, 0, 0)$. 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 $S = 2a + 1$. The partial charges for a benzene ring as modeled in Table 13.1 consist of three sets of such paired dipoles, arranged in a hexagonal fashion.

The potential for such a charge group can be estimated by algebraic means, as we did in Chapter 3, or we can utilize the technology of Section 9.3. We define

$$\mathcal{D}_\epsilon = T_{\epsilon(a+1,0,0)} - T_{\epsilon(a,0,0)} + T_{\epsilon(-a-1,0,0)} - T_{\epsilon(-a,0,0)}. \tag{9.46}$$

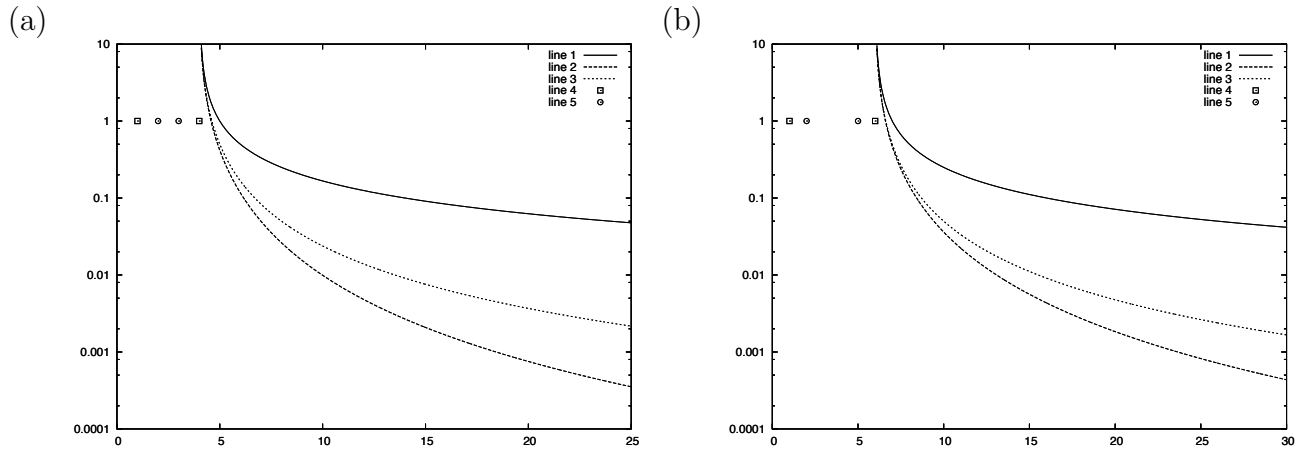


Figure 9.11: Comparison of single charge, dipole and quadrupole potentials. Dipole separation (a) two units and (b) four units. The locations of the negative charges are indicated by circles and the locations of the positive charges are indicated by squares. The upper solid line is the potential for a single positive charge 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.

In evaluating $\mathcal{D}_\epsilon\phi$, we may as well assume that ϕ is only a function of x. Applying (9.42) to ϕ and ϕ' we find that

$$\begin{aligned}\mathcal{D}_\epsilon\phi(x) =&\phi(x-\epsilon(a+1)) - \phi(x-\epsilon a) + \phi(x+\epsilon(a+1)) - \phi(x+\epsilon a)\\ =&\epsilon\phi'(x-\epsilon(a+\tfrac{1}{2})) - \epsilon\phi'(x+\epsilon(a+\tfrac{1}{2})) + \mathcal{O}(\epsilon^3)\\ =&\epsilon^2(2a+1)\phi''(x) + \mathcal{O}(\epsilon^3)\\ =&\epsilon^2 S\phi''(x) + \mathcal{O}(\epsilon^3),\end{aligned} \tag{9.47}$$

where S is the separation distance between the dipoles. Thus

$$\lim_{\epsilon\to 0}\epsilon^{-2}\mathcal{D}_\epsilon = (2a+1)\frac{\partial^2}{\partial x^2} = s\frac{\partial^2}{\partial x^2}. \tag{9.48}$$

Applying (9.40), we find

$$V(\mathbf{r}) = |\mathbf{r}|^{-3} s\frac{\partial^2}{\partial x^2}W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-4}) \tag{9.49}$$

for large $\mathbf{r}$, where W $(= 1/r)$ is defined in (9.24) and S is the separation distance between the dipoles.

The potential for opposing dipoles is depicted in Figure 9.11 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

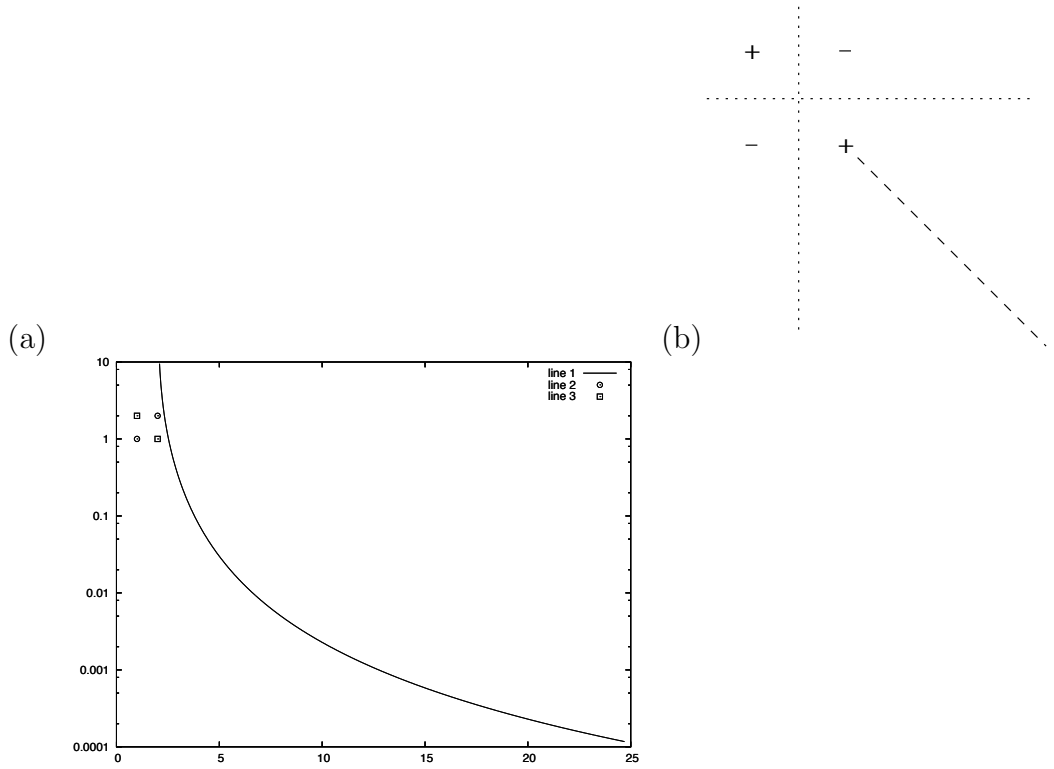


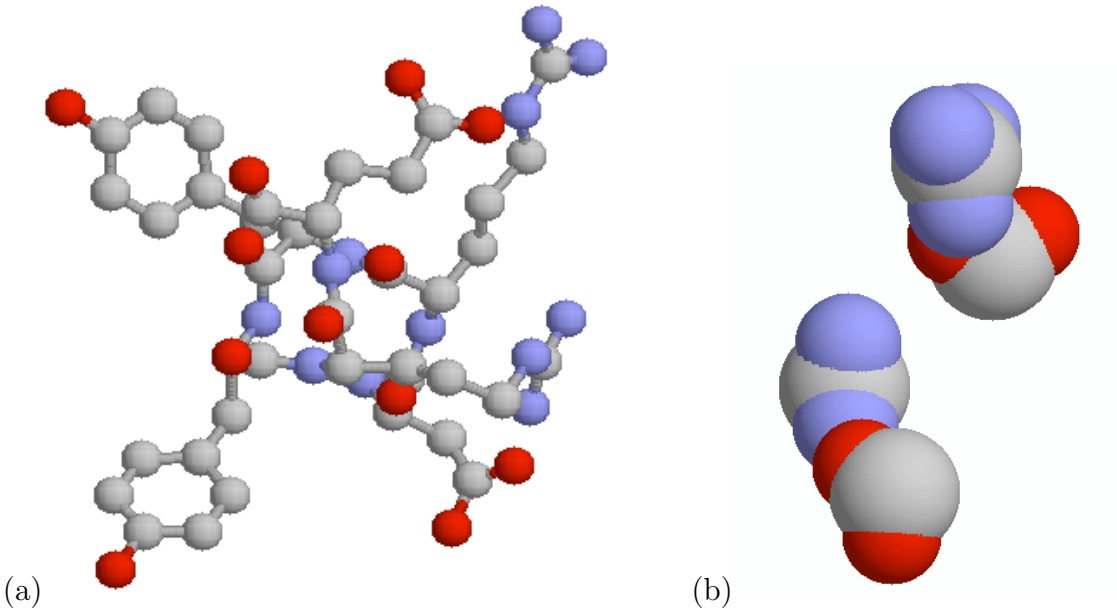
Figure 9.12: Four corner quadrupole potential. (a) The potential is plotted as a function of distance s along the line $(x(x), y(s)) = ((2 + s/\sqrt{2}, 1 - s/\sqrt{2})$ which emanates from the lower-right corner of the quadrupole. The locations of the negative charges are indicated by circles and the 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.

interpreted as Ångstroms, then the separation $S = 4$ (b) is roughly comparable to the partial charge model of a benzene ring (cf. Table 13.1) consisting of three sets of such paired dipoles.

9.4.2 Four-corner quadrupole

The four-corner arrangement provides a two-dimensional arrangement of opposing dipoles. 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 9.12. Note that it dies off a bit more rapidly than the potential for the opposing dipoles (cf. Figure 9.11). Defining

$$\mathcal{D}_\epsilon = \sum_{k=1}^{K} q_k T_{\epsilon r_k} \tag{9.50}$$



(a) (b)

Figure 9.13: 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.

and applying (9.42) twice, we see that

$$\begin{aligned}\mathcal{D}_\epsilon\phi(\mathbf{r}) =& \sum_{k=1}^{K} q_k\phi(\mathbf{r}-\epsilon\mathbf{r}_k) \\ =& 4\frac{\partial}{\partial x}\frac{\partial}{\partial y}\phi(0)\epsilon^2 + \mathcal{O}(\epsilon^3)\end{aligned} \tag{9.51}$$

Thus

$$V(\mathbf{r}) = |\mathbf{r}|^{-3}4\frac{\partial}{\partial x}\frac{\partial}{\partial y}W(|\mathbf{r}|^{-1}\mathbf{r}) + \mathcal{O}(|\mathbf{r}|^{-4}). \tag{9.52}$$

It is not hard to generalize these results to the case where the opposing charges form the four corners of any parallelogram.

9.4.3 Quadrupole example

An example of a (near) quadrupole is found in the human prion (PDB file 1I4M) in the motif DRYYRE. This is shown in Figure 9.13. 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 9.13(b) shows the skewness of the two opposing dipoles.

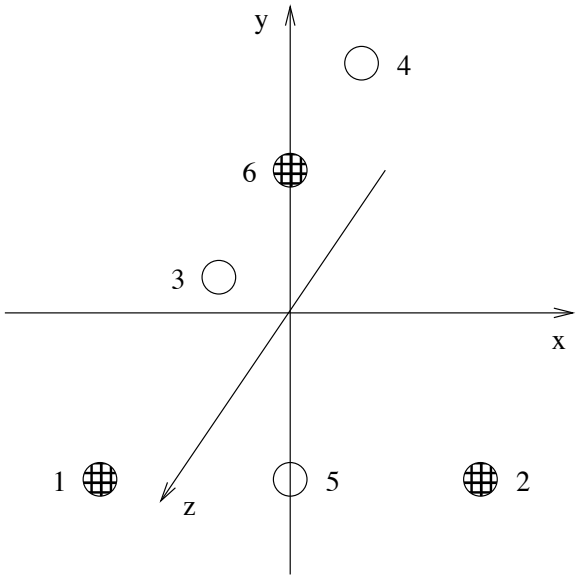


Figure 9.14: Configuration of charges in water model. Open circles indicate negative charge locations; shaded circles indicate locations of positive charge.

9.4.4 Water: dipole or quadrupole?

Water can be written as a combination of two dipoles, following the general pattern of Section 9.3. 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 write water as system with positive charges

$$q_1 = q_2 = a \quad \text{at} \quad \mathbf{r}_{1,2} = (\pm c, -1, 0) \tag{9.53}$$

and negative charges

$$q_3 = q_4 = -b \quad \text{at} \quad \mathbf{r}_{3.4} = (0, y^0, \pm d), \tag{9.54}$$

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 used by Tip5P [150], with $a = b$. We would like to show that this system is a dipole; by that, we mean two things, one of which is that it is *not* a quadrupole.

To discover the exact multipole nature of our water model encoded in (9.53) and (9.54), we modify it to form a quadrupole. We extend the system (9.53–9.54) to involve two more charges:

$$\begin{aligned} q_5 = -2a \quad &\text{at} \quad \mathbf{r}_5 = (0, -1, 0) \quad \text{and} \\ q_6 = 2b \quad &\text{at} \quad \mathbf{r}_6 = (0, y_0, 0). \end{aligned} \tag{9.55}$$

The configuration of charges is depicted in Figure 9.14.

The extended system is a quadrupole due to the cancellations leading to an expression such as (9.51). More precisely, note that the charges at locations 1, 2 and 5 correspond to a second difference stencil centered at point 5 for approximating

$$\frac{\partial^2\phi}{\partial x^2}(0,-1,0) \tag{9.56}$$

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

$$\frac{\partial^2\phi}{\partial z^2}(0,y^0,0) \tag{9.57}$$

(with suitable scaling). Therefore

$$\begin{aligned}\mathcal{D}_\epsilon\phi(0) =& \sum_{k=1}^{6} q_k\phi(\epsilon\mathbf{r}_k)\\ =& ac^2\epsilon^2\frac{\partial^2\phi}{\partial x^2}(0,-1,0) - bd^2\epsilon^2\frac{\partial^2\phi}{\partial z^2}(0,y^0,0)\epsilon^2 + \mathcal{O}(\epsilon^4),\end{aligned} \tag{9.58}$$

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 (9.55). We leave as Exercise 9.10 to show that this is a dipole provided $a = b$. Let V_Q denote the quadrupole potential associated with (9.58), and let V^W be the water potential using the model (9.53–9.54). Thus we have written the water potential as

$$V^W = V^D + V^Q \tag{9.59}$$

for an explicit dipole potential V^D, with charges at $\mathbf{r}_5$ and $\mathbf{r}_6$, and a quadrupole. Thus the water model (9.53–9.54) 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 $\angle HOH$.

9.5 Further results

We collect here some further results about electrostatic interactions.

9.5.1 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 10.6), 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 (**??**)).

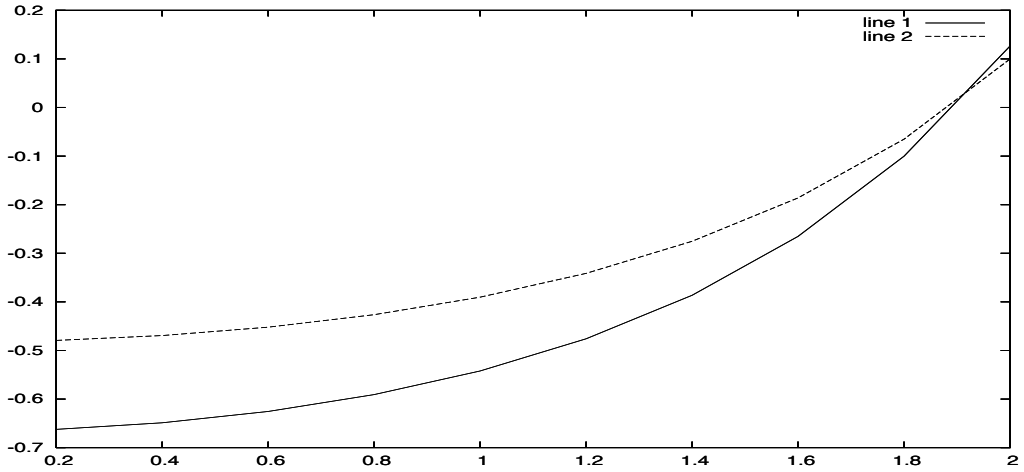


Figure 9.15: Dipole-dipole (in-line) interaction energy, scaled by R^3, for $R = 2$ for the two models. Horizontal ϕ-axis measured in radians. The flattest (dashed) curve corresponds to $\alpha = 0.8, \beta = 1.0, \gamma = 0.8, \delta = 0.2$, whereas the solid curve corresponds to $\alpha = 0.8, \beta = 1.0, \gamma = 0.8, \delta = 0.2$,

9.5.2 Modified dipole interaction

Since we found that 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 the question: if the hydrogen charge density is represented in a more complex way, will a stronger angular dependence appear? To address this question, we introduce a negative charge to represent the electron density beyond the hydrogen. The exact positions of the charges are as follows. The position of the negative charge on the right we take as the origin, and we assume the separation distance of 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 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

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R+\delta, 0)| =& \sqrt{(R-\delta+\cos\phi)^2 + \sin^2\phi} \\ =& \sqrt{1 + (R-\delta)^2 + 2(R-\delta)\cos\phi} \end{aligned} \tag{9.60}$$

The distance between the two positive charges is

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R, 0)| =& \sqrt{(R+\cos\phi)^2 + \sin^2\phi} \\ =& \sqrt{1 + R^2 + 2R\cos\phi} \end{aligned} \tag{9.61}$$

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

$$\begin{aligned} |(\cos\phi, \sin\phi) - (-R-1, 0)| =& \sqrt{(1+R+\cos\phi)^2 + \sin^2\phi} \\ =& \sqrt{1+(R+1)^2 + 2(R+1)\cos\phi} \end{aligned} \tag{9.62}$$

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

$$\begin{aligned} &\frac{\alpha}{R+1} - \frac{\beta}{R} + \frac{\gamma}{R-\delta} \\ &\qquad - \frac{\gamma}{\sqrt{1+(R-\delta)^2 + 2(R-\delta)\cos\phi}} \\ &+ \frac{\beta}{\sqrt{1+R^2+2R\cos\phi}} - \frac{\alpha}{\sqrt{1+(R+1)^2+2(R+1)\cos\phi}} \end{aligned} \tag{9.63}$$

A plot of the interaction energy (9.63) is given in Figure 9.15 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.

9.5.3 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 5.4. 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 9.16. 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 9.16(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 5.4) 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 9.16, 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

$$\frac{-1}{\sqrt{x_0^2 + (y_0 - \cos\theta)^2 + \sin^2\theta}} + \frac{1}{\sqrt{x_1^2 + (y_1 - \cos\theta)^2 + (z_1 - \sin\theta)^2}} \tag{9.64}$$

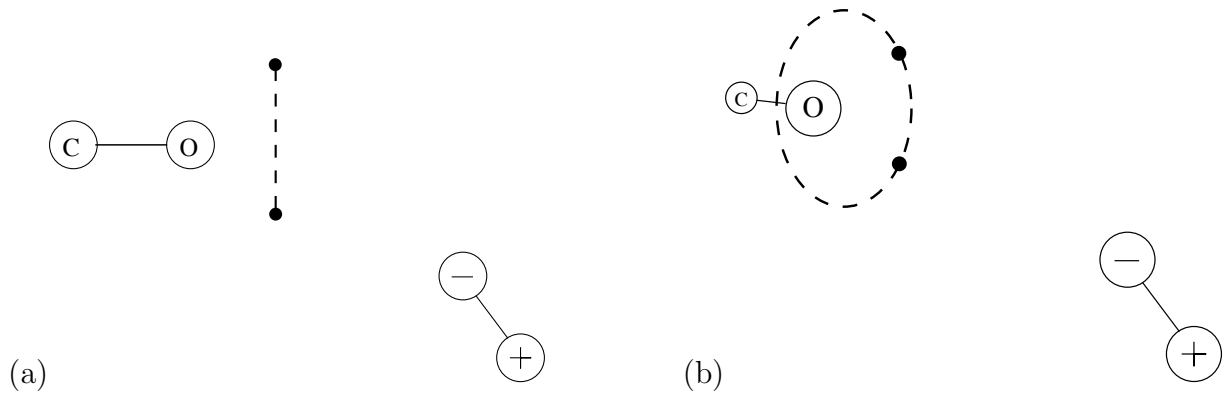


Figure 9.16: Configuration for the placement of hydrogen at the end of the sidechain of serine or threonine in response to a nearby dipole. The dashed line indicates the circle of possible hydrogen placements. (q) The plane of the circle is orthogonal to the plane of the page. (b) The plane of the circle is skew to the plane of the page.

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

We can also use the expression (9.16) to find the optimum θ. In coordinates determined so that the hydrogen and the dipole lie in a plane, the interaction field (9.16) 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

$$\nabla V(x(\phi), y(\phi), z(\phi)) \cdot (x'(\phi), y'(\phi), z'(\phi)) = 0. \tag{9.65}$$

9.6 Exercises

Exercise 9.1 *Show that the interaction energy (9.4) tends to the asymptotic form*

$$\frac{-2\cos\phi}{R^3}. \tag{9.66}$$

Exercise 9.2 *Show that the interaction energy (9.7) tends to the asymptotic form*

$$\frac{-\frac{1}{2} - \frac{3}{2}\cos\phi}{R^3}. \tag{9.67}$$

Exercise 9.3 *Verify that the second term in the energy expression in (9.11) is indeed the same as (9.5). Also verify that the fourth term in the energy expression in (9.11) is correct.*

Exercise 9.4 *Plot the expression in (9.12) 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 9.5 *Plot the expression in (9.22) and verify that it is symmetric around $\phi = \pi/2$ and has a simple minimum there.*

Exercise 9.6 *Carry out the calculations leading to the expression in (9.22) in the case that the charge group has a positive charge as well as the negative charges, as shown in Figure 9.10. 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 9.7 *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 9.8 *Prove that the asymptotic expression (9.14) is valid for fixed θ and large r. (Hint: show that*

$$V(r,\theta) = \frac{a}{r}\left(\frac{1-\sqrt{1+2r^{-1}\cos\theta+r^{-2}}}{\sqrt{1+2r^{-1}\cos\theta+r^{-2}}}\right) \tag{9.68}$$

and expand the expression in the numerator. Is this asymptotic approximation uniformly valid for all θ?)

Exercise 9.9 *Determine the percentage error in the approximation (9.14) when $\theta = \pi/4$ and $r = 3$.*

Exercise 9.10 *Show that a charge system with only the charges as indicated in (9.55) forms a dipole provided $a = b$ and examine its asymptotic behavior.*

Chapter 10

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.

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.

10.1 Biochemical units

There are natural units associated with biochemical phenomena. 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 Avogradro'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.

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 we can think of a joule as one kilogram-(meter/second)^2. The standard (SI) unit for energy is the joule, but it differs from the older calorie only by a numerical factor. Energy has units mass times velocity squared, as we know from Einstein's famous relation. In particular, a joule is 6.7006×10^9 amu c^2.

A natural unit of time for biochemistry is the femtosecond (10^{-15} second). 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 time scale resolves molecular motion, but does not over resolve it by much: 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. Note that our typical 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 femtosecond, about 2.9979×10^{-7} meters, or 300 nanometers. This may seem odd. You might have expected a spatial unit on the order of an Ångstrom (roughly the radius of the smallest atoms), but this is three-thousand Ångstroms. 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-femtosecond (lfs).

10.1.1 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 $2.25 \times 10^{24} q_e$.

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

$$\begin{aligned}
\epsilon_0 =& 8.8542 \times 10^{-12} \mathrm{C}^2\mathrm{N}^{-1}\mathrm{m}^{-2} \\
=& 3.450 \times 10^{26} q_e^2 \mathrm{N}^{-1}\mathrm{m}^{-2} = 3.450 \times 10^{26} q_e^2 \mathrm{J}^{-1}\mathrm{m}^{-1} \\
=& 1.444 \times 10^{27} q_e^2 \mathrm{cal}^{-1}\mathrm{m}^{-1} = 1.444 \times 10^{30} q_e^2 \mathrm{kcal}^{-1}\mathrm{m}^{-1} \\
=& 2.40 \times 10^{6} q_e^2 (\mathrm{kcal/mole})^{-1}\mathrm{m}^{-1} = 2.40 \times q_e^2 (\mathrm{kcal/mole})^{-1}\mu\mathrm{m}^{-1} \\
=& 0.72 q_e^2 (\mathrm{kcal/mole})^{-1}\mathrm{lfs}^{-1} .
\end{aligned} \tag{10.1}$$

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 [46] used $\epsilon_0 = 1$ as a unit, 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. If this is the chosen spatial unit, then $\epsilon_0 = 1$ in these units. For reference, very large viruses [251] are between one and two tenths of a micron in diameter.

10.1.2 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

$$\begin{aligned} k_B =& 1.380 \times 10^{-23} \text{J/K} \\ =& \frac{1.380 \times 6.022}{4.1868} \text{cal/mole-K} \\ =& 1.984 \, \text{cal/mole-K} \end{aligned} \tag{10.2}$$

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 atomic mass units, then $k_B \approx 0.831$.

Planck's constant = 6.626068 10^{-34} m^2 kg / s = 39.90165 Ångstroms^2 amu/picosecond. The constant $\hbar = h/2\pi$ is then $\hbar = 6.35055$ Ångstroms^2 amu/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.

10.2 Quantum chemistry units

The Schrödinger equation has three terms which must have the same units in order to be dimensionally correct. If we divide (15.1) 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 (15.1) we have implicitly assumed that the permittivity of free space $\epsilon_0 = 1$. We can do this, as noted above, but we need to choose the right spatial unit to make it all work out. Unfortunately, that spatial unit is quite large for quantum chemistry, since it is four orders of magnitude larger than the typcial scale of interest.

A more typical choice of spatial unit at the quantum scale [219] would be to use the Bohr radius $a_0 = 0.529189$Å. If we also adopt the Hartree[1] $\mathcal{E} = 4.356 \times 10^{-18}$ joules $= 1.040 \times 10^{-21}$ kcal, or about 626.5 kcal/mole (cf. Table 3.1), and we adopt the mass of the electron m_e as the unit of mass, then $\hbar^2/m_e a_0^2 = \mathcal{E}$. Moreover, we also have the coefficient of the potential in (15.1) equal to $\mathcal{E}$; that is, $e^2/(4\pi\epsilon_0 a_0) = \mathcal{E}$.

The time-derivative term in (15.1) 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, and on the scale of a few femtoseconds (the typical time step of molecular dynamics simulations), the time-derivative term in (15.1) can plausibly be ignored, or rather time-averaged.

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

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.

10.3 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. 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 amu, since it is roughly the mass of the smallest atom. With this as the mass unit, the masses in the Schrödinger equation are of order one. 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 $\hbar/c = 0.212 \times 10^{-15}$ amu-meters. If we want $\hbar = 1$ and $c = 1$, 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 $\hbar/c^2 = 0.7066 \times 10^{-24}$ amu-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 0.212 femtometers, time to be measured in multiples of 0.7066 yoctoseconds, and mass in atomic mass units, then $c = \hbar = 1$. As noted above, a joule in these units is 6.7006×10^9. So $k_B = 9.2468 \times 10^{-14} K^{-1}$.

10.4 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.

10.5 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. Estimates vary around one base pair per million years. 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, there has not been enough time for this type of mutation to cause a complete change to a typical chromosome. So from the point of view of a dynamical system, we are neither just at the beginning of a full cylce, nor have we seen more than one cycle. Rather we are just well into one complete cycle. Thus we would not expect to see any limiting behavior yet.

10.6 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 [92, 83].

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 $\alpha \approx 1.2 \text{Å}^3$.

10.7 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 10^{23}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 the typical O-O distance in water (about 2.75 Å).

10.8 Fluid viscosity and diffusion

Fluids display an aggregate behavior known as **viscosity**. Fluid dynamicists [] call the viscosity μ and physicists [] call it η. The units of the coefficient of viscosity (often called **dynamic vis-**

cosity) 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, over a factor of five-hundred smaller.

10.8.1 Kinematic viscosity

Another scaling factor is significant in fluid flow, namely the fluid density. The ratio of viscosity (or dynamic viscosity) and density is called **kinematic viscosity**, usually labelled ν. This has units length-squared per time, since density has units of mass per length-cubed. Thus kinematic viscosity has the same units as a spatial diffusion constant. The stoke is one centimeter-squared per second. The kinematic viscosity of water is about one millimeter-squared per second, or one centistoke, whereas the kinematic viscosity of air is rougly two times *larger*. That is, air is more viscous than water! The viscosity of fluids varies significantly with temperature, but we have provided values at roughly the same temperature (293 K) for comparison.

Viscous drag is the effective force of viscosity in opposing motion. It provides a retarding force in the direction oppositve to the motion. The drag coefficient has the units of force divided by velocity, or mass per time unit.

10.8.2 Diffusion

10.9 Exercises

Exercise 10.1 *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.*

[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.