

Chapter 2

Digital Rules for Proteins

Lise Meitner (1878–1968) was “the physicist first to recognize that experiments reported by two former colleagues in Berlin meant that atoms had been split, never got [the Nobel] prize, even though one of those colleagues, Otto Hahn, did in 1944.” (October 20, 2003, New York Times.) Element 109, meitnerium, is named in her honor.

Digital rules are a hallmark of a mature science. Protein associations are the elementary events in biology when examined at the molecular level. This book attempts to develop digital rules for interactions involving proteins. We begin with a sketch of some of the main ideas that the book will cover. This is not an outline but rather is a narrative that introduces the main goals and challenges to be addressed and gives a glimpse of some of the major advances.

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

2.1 Digital Nature of Molecules

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

2.1.1 *Digital Nature of Atoms*

Atoms can be characterized by the number of electrons, protons, and neutrons of which they are composed. Some atoms of primary interest in protein biochemistry are listed in Table 2.1, and this table expresses digital rules at different scales: subatomic, atomic, and molecular.

There are other subatomic digital rules that we will generally take for granted here, like Pauli’s electron spin exclusion principle and Hund’s rule about electron orbital occupation.

Several rules are encoded in Table 2.1. The first rule is used to reduce the number of columns: the number of protons always equals the number of electrons (the net charge is zero). A second rule is that the typical number of neutrons in the dominant isotope is nearly the same as the number of protons. But the most important rule is the **octet rule**: the number of the electrons in the outer shell plus the number (listed in the “lacking” column) of electrons contributed by atoms covalently bonded to it is always eight (except for hydrogen), up to chlorine, and then, it is eighteen for the larger atoms (for even larger atoms, the magic number is thirty-two, but those atoms do not concern us). This simple rule facilitates the determination of molecular bond formation.

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

Table 2.1 Subset of the periodic table. The column “+/-” denotes the number of protons and electrons in the atom. The column “outer” is the number of electrons in the outer shell. The column “lacking” is the number of electrons needed to complete the outer shell. The column “ion” gives the charge of the atomic ion (copper has two possible forms). The column “mass” gives the atomic mass in Daltons (see Chapter 18 for details), reflecting the naturally occurring isotopic distribution. The column “radius” lists the “mean” van der Waals radius [56], with the exceptions marked by *’s taken from various Web sites.

Another rule encoded in Table 2.1 is that the atomic mass is very close to the number of protons plus the number of neutrons (in the specified unit, the Dalton). The mass of the proton and neutron are approximately the same, and the rest mass of an electron is less than 0.0006 times this amount; see Table 2.2. The units of atomic mass are discussed in Section 18.2.1, but for now, the main point is that the mass of the proton and neutron are about one in the standard unit for atomic measurements, the Dalton. For reference, we list in Table 2.2 these masses in more familiar units.

As we see in Table 2.1, the number of neutrons can vary. We have listed what is known as the dominant isotope, followed (in parentheses) by the other possible stable isotopes (involving the listed numbers of neutrons). Neutrons add mass but not charge. Various isotopes are important in certain contexts; a hydrogen atom with an extra neutron is called deuterium. Atoms occur naturally in different isotopic forms, and the atomic mass reflects this natural variation. Otherwise, the atomic mass would be essentially the sum of the numbers of protons

and neutrons, with a small correction for the electronic mass, as well as another correction that we will discuss shortly. For chlorine, about a quarter of the atoms have 20 neutrons, and thus, the atomic mass is about halfway between integer values. The given atomic masses are themselves only averages, and any particular set of atoms will vary in composition slightly; see the Periodic Table in [383] for more details.

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

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

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

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

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

The digital description of an atom is to be contrasted with the analog description of the Schrödinger equation [385]. This equation describes the electron (and proton) distribution, which is the key determinant of atomic interaction. It is a continuum equation predicting the electronic distribution at all points in space, and there is a separate three-dimensional space at the least for each electron and in the general case for the protons as well. Even if it were simple to solve this equation (which it is not), it would be difficult to determine simple facts from such a representation. We are forced to consider effects on this level in many cases, but operating at the atomic level has clear advantages.

proton	neutron	electron	Dalton
1672.622	1674.927	0.910938	1660.539

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

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

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

There are other simple rules in chemistry that allow prediction of bond formation, such as the electronegativity scale (Section 8.2.1) and the resonance principle (Section 14.1). The electronegativity scale allows the determination of polarity of molecules (Section 8.2). The resonance principle states that observed states of molecular bonds are often a simple convex

combination of two elementary states. For example, a benzene ring can be thought of as being made of alternating single and double bonds, whereas in reality, each bond is closely approximated by a convex combination of these two bonds. The resonance principle may be thought of as a Galerkin approximation to solutions of the Schrödinger equation.

The size data for atoms listed in the “radius” column in Table 2.1 provide another way to distinguish between different atoms in a simple way, although the lengths do not correspond to a tangible boundary. If we could look at atoms at this level, the nucleus would be a tiny dot, and the electrons would be a fuzzy cloud, extending beyond the stated radius with a certain (nonzero) probability (cf. Figure 1 in [56]). Rather, the length corresponds to the size of an “exclusion zone” to give an idea of the size of a region where other atoms would not (typically) be found. A similar notion of length will be discussed in Section 5.6 regarding the size of atomic groups that form proteins. The length variation may not seem extreme, but the corresponding volume variation is over an order of magnitude.

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

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

Table 2.4 Relation between volume r^3 and length r in the range of lengths (in Ångstroms) relevant for atoms.

2.1.2 Ionic Rules

One rule that bridges between atoms and molecules is related to ions. Certain ions are key to life in humans. There many important ionic effects that are specific to individual ions and are not easily explained by simple rules [291]. However, the charge of many ions can be deduced from a model that fits the data in Table 2.1 quite well for most, but not all, ions. The rule is that the atomic ion charge c is related to the number of electrons ℓ lacking in the outer shell and the total number of electrons O in a complete outer shell by

$$c = \begin{cases} O - \ell & \text{positive ions} \\ -\ell & \text{negative ions.} \end{cases} \quad (2.3)$$

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

2.1.3 Carbon/Hydrogen Rules

Another example of a simple rule at the molecular level relates to the common occurrence of hydrogens bonded to carbons. These occur so commonly in bio-organic chemistry that they are often only implied in graphical representations. In biomolecules, the hydrocarbon skeleton is typically represented without direct labeling of the atoms involved, with combinatorial rules that subsume the valence information implied in the electronic outer shell of hydrogen and carbon. For example, a benzene ring might be written as a simple hexagon, without any labeling of hydrogen and carbons. The chemical formula for benzene is C_6H_6 . Implicit in the graphical representation is that a carbon is located at each vertex of the hexagon and that each carbon is bonded to a hydrogen. The rules apply to a wide variety of molecules, such as β -D-galactose (GAL) and N-acetyl-D-glucosamine (NAG) (Section 4.6.3); the three letter names for these molecules are their names in the Protein Data Bank. The three letter names for benzene in the Protein Data Bank are BNZ.

2.2 Digital Nature of Proteins

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

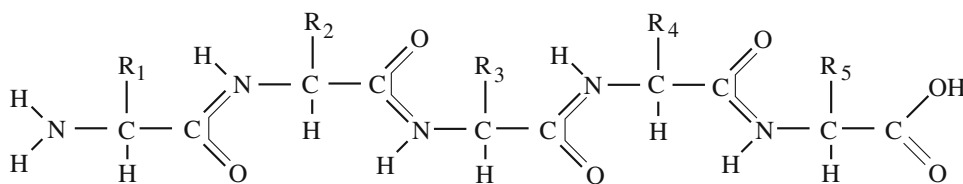


Fig. 2.1 Schematic representation of a peptide sequence with five units (all in the trans conformation). The bonds between C's and N's denote the transition point between one peptide unit and the next. The residues are numbered in increasing order starting at the N-terminal end.

The digital and deterministic nature of protein function is implied by the fact that their structure is encoded by a discrete mechanism, DNA. This linear description is then translated into protein, a small fragment of which is depicted in Figure 2.1. But this linear representation of proteins belies the functional geometry of a protein, which is three dimensional. It is the latter representation that appears at first to be anything but a discrete widget. However, we will see that it is so. There are posttranslational 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 digital ways is the significant role played by hydrophobic effects (Section 2.6) in most protein–ligand interactions. Such interactions account not only for the formation of protein complexes, but also for signaling and enzymatic processes. But the hydrophobic effect is essentially nonspecific. Thus, its role in a discrete system is intriguing.

We will see that it is possible to quantify the effect of hydrophobicity in discrete ways. The concept of *wrapping* (see Chapter 8) yields such a description, and we show that this can affect many important phenomena, including protein association (molecular recognition) (Chapter 7) and the flexibility of the peptide bond (Chapter 14). In these two examples, we will see behaviors that are essentially digital in nature that can be predicted based on quantitative measures. In particular, both of these effects yield what can be viewed as switches, things that can be turned on and off. There are other examples of switches in protein behavior, and even the fundamental hydrogen bond can be viewed in this way to a certain extent.

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

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

2.2.1 Hydrophobicity and Hydrophilicity

Any molecular material is termed hydrophobic if it does not interact favorably with water, typically because it cannot make appropriate hydrogen bonds with water. One major objective of the book is to clarify the effects of hydrophobicity in particular cases, including its role in protein–ligand interactions and other important phenomena. Hydrophobic moieties displace water and, as we will see, decrease the dielectric capacity of the environment. Hydrophobic molecular groups can reduce the dielectric effect locally, which in turn enhances certain nearby bonds. Particular atomic groups in protein sidechains will be identified as being hydrophobic. A major objective of the book is to clarify these concepts.

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

2.2.2 Solvation $\neq$ Salvation

The life of a protein in water is largely a struggle for the survival of its hydrogen bonds. The hydrogen bond (cf. Chapter 6) is the primary determinant of the structure of proteins. But water molecules are readily available to replace the structural hydrogen bonds with hydrogen bonds to themselves; indeed, this is a significant part of how proteins are broken down and

recycled. We certainly cannot live without water [376], but proteins must struggle to live with it [146, 309].

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 (unfolded) 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 [191], 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 [3, 464, 490, 516], 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 sidechains of amino acids look very much like water to a water molecule. But more importantly, the protein backbone hydrogen bonds can be replaced by hydrogen bonds with water, and this can disrupt the protein structure. This can easily lead to the break-up of a protein if water is allowed to attack enough of the protein's hydrogen bonds.

The primary strategy for protecting hydrogen bonds is to bury them in the core of a protein. But this goes only so far, and inevitably, there are hydrogen bonds formed near the surface of a protein. And our understanding of the role of proteins with extensive noncore 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 sidechains (Section 4.1.1). Simply counting carbonaceous groups (e.g., CH_n for $n = 0, 1, 2$ or 3) in the sidechains shows a striking range, from zero (glycine) to nine (tryptophan). Most of the carbonaceous groups are nonpolar and thus hydrophobic. Having the right amino acid sidechains 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 nonconducting tape.

We refer to the under-protected hydrogen bonds, which are not sufficiently wrapped by carbonaceous groups, as **dehydrons** (Section 2.6.2) to simplify terminology. The name derives from the fact that these hydrogen bonds are stabilized and strengthened by being dehydrated.

2.2.3 Energetic Ambivalence

One could imagine a world in which all hydrogen bonds were fully protected. However, this would be a very rigid world. Biology appears to prefer to live at the edge of stability. Thus, it is not surprising that new modes of interactions would become more prevalent in biology than in other areas of physics. For example, it has been recently observed that exposed hydrogen bonds appear to be sites of protein-protein interactions [174]. 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 epid-iorthotric stimulant. Similarly, snowflakes and raindrops tend to form around small specs

of dust. Such forces also have analogies in personal, social, and political interactions where forces based on detrimental circumstances cause a beneficial outcome. A couple who stay together because they do not want to be alone provides such an example. The defect of an under-protected hydrogen bond gives rise to just such an epidiorthotic force. The action of this force is indirect, so it takes some explaining.

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

If the removal of water from an under-protected hydrogen bond is energetically favorable, then this means there is a force associated with attracting something that would exclude water. Indeed, one can measure such a force, and it agrees with what would be predicted by calculating the change energy due to the change in dielectric (Section 9.1). You can think of this force as being somewhat like the way that adhesive tape works. Part of the force results from the removal of air between the tape and the surface, leaving atmospheric pressure holding it on. However, the analogy goes only 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 epidiorthotic force associated with water removal from an under-protected hydrogen bond provides a mechanism to bind proteins together. This is a particular type of hydrophobic effect, because wrapping the bond with hydrophobic groups provides protection from water. It is intriguing that it arises from a defect which provides an opportunity to interact.

2.3 Pchemomics

The term “omics” refers to the use of biological databases 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 [287]. Indeed, the original study of the structure of the peptide bond (see section 8-4 of [382]) 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 Bank (**PDB**) provides three-dimensional structures that yield continuing opportunities for proteomics discoveries. Using the perspective of physical chemistry in data mining in the PDB, some simple laws about protein families were determined by studying patterns of under-wrapped hydrogen bonds [162]. We examine just one such result in Section 2.3.4; many other results in physical chemistry can be likewise explored.

A simple view of the PDB gives a representation suitable only 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 involves a type of physical chemistry lens.

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 [241] 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 [179] 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 [314]. This means that for very subtle issues, one must still be careful about even all-atom simulations. The mysteries of water continue to confront us. But its role in biology will always be central.

2.3.1 A New Tool?

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

Typical data mining 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 [482], the following quote is featured:

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

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

2.3.2 Data Mining Definition

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

Data mining is sorting through data to identify patterns and establish relationships.

Data mining parameters include:

- Association — looking for patterns where one event is connected to another event
- Sequence or path analysis — looking for patterns where one event leads to another later event
- Classification — looking for new patterns (may result in a change in the way the data is organized but that's ok)
- Clustering — finding and visually documenting groups of facts not previously known

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

2.3.3 Data Mining Lens

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

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



In the early research on proteins [482], discovering such formulæ was a major step. But a much bigger step came with the realization that proteins are composed of sequences of amino acids. This allowed proteins to be described by alphabetic sequences, and the descriptions come in different forms: DNA, RNA, amino acid sequences. One can think of these from a linguistic perspective, and indeed, this has been a productive approach [254].

The function of DNA is largely to store sequence information, but proteins operate as three-dimensional widgets. Not all proteins have a stable three-dimensional representation, but most biologically relevant proteins function via three-dimensional structures. Indeed, random proteins would be expected not to form stable three-dimensional structures [120]. The PDB is a curated database of such structures that provides a starting point to study protein function from a physical chemistry perspective.

But structure alone does not explain how proteins function. Physical chemistry can both simplify our picture of a protein and also allow function to be more easily interpreted. In particular, we will emphasize the role of the modulation of the dielectric environment by hydrophobic effects. We describe a simple way this can be done to illustrate the effect on

individual electronic entities, such as bonds. But there is need for better lenses to look at such complex effects.

One way to describe what we are doing in modern terms is “big data analytics.” The definition of what is “big” is subjective, but the exercises in the book require handling thousands of PDB files. Following such small steps in a research direction can lead to much larger datasets, but we do not discuss here the issues related to high-performance computing necessary to do computing at that scale [444]. The “analytics” aspect requires the implementation of many mathematical ideas. Most of these require at most calculus, but the need to understand the three-dimensional geometry of proteins forces a review of concepts that are accessible in the precalculus curriculum. In using this book as a text, such mathematical concepts can be emphasized to a greater or lesser extent, as desired.

2.3.4 Hydrogen Bonds are Orientation-Dependent

The hydrogen bond provides a good starting example of the use of “pchem” data mining to reveal its properties. Figure 6 of [287] shows clearly both the radial dependence and the angular dependence of the hydrogen bond. Similar evidence is found in later papers; Figure 3 in [483] suggests that hydrogen bonds are stronger when they are both shorter and better aligned. However, the precise relationships between angle and distance can depend on the context, being different in different types of protein structure [32]. Figure 8 of [535] 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 [355]. All of these involved data mining.

An alternative method for modeling hydrogen bonds is to study their energetics via quantum mechanical calculations and to interpolate the resulting energy surfaces [364, 460]. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations has also been reported [283]. Despite the inherent interest in hydrogen bonds, a general model of them has not yet been developed. In particular, hydrogen bonds do not appear as primary bonds in molecular dynamics simulations. Due to the primary importance of the hydrogen bond in protein structure, we will review what is known and not known in Chapter 6.

2.3.5 What is an Answer?

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

A probability distribution provides a way to give answers that combine the types of answers you get with high-school algebra and those you get with calculus. An answer that is a number is a Dirac δ -function, whereas a function corresponds to a measure that is absolutely continuous.

This added level of sophistication is especially helpful in a subject where it seems almost anything can happen with some degree of probability.

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

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

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

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

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

2.4 Multiscale Models

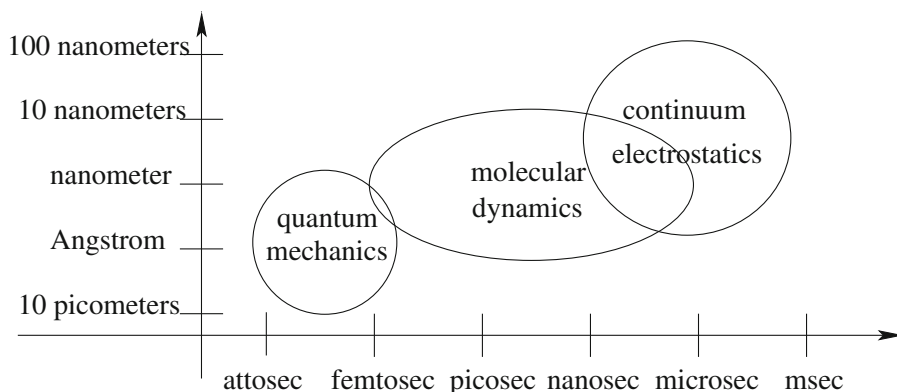


Fig. 2.2 Cartoon of spatial and temporal scales of biomolecular models. See the text for more details.

But why don't we just write down a mathematical model and use it to simulate protein dynamics? This is a reasonable question, and we attempt here to show why such an approach at the moment would not be productive. The difficulty is the particular multiscale aspect of the problem: the temporal scales are huge but the spatial scales overlap, as depicted in Figure 2.2. Of course, existing models are useful in limited contexts. However, we will explain limitations in two such models that must be addressed in order to use them on more challenging simulations.

Models for many systems have components which operate at different scales [236]. Scale separation often simplifies the interactions among the different scales. The differences often occur in both spatial and temporal scales. Scale separation often simplifies the study of complex systems by allowing each scale to be studied independently, with only weak interactions among the different scales. However, when there is a lack of scale separation, interactions among the scales become more difficult to model.

There are three models of importance in protein biochemistry. The different spatial and temporal scales for these models are depicted in Figure 2.2. The smallest and fastest scale is that of quantum chemistry. 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 timescale of molecular dynamics is much longer than the quantum scale. But the length scale is comparable with the quantum scale. For example, the Ångström can be used effectively to describe both without involving very large or very small numbers.

Finally, the electric properties of proteins are mediated by the dielectric behavior of water in a way that is suitable for a continuum model [110, 447]. 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 timescale for the continuum dielectric model, but we have depicted in Figure 2.2 the timescale for so-called Brownian dynamics models which are based on a continuum dielectric model [329].

The lack of physical scale separation, linked with the extreme time separation, in biological systems is the root of some of the key challenges in modeling them. Note that the temporal scales in Figure 2.2 cover fifteen orders of magnitude, whereas the spatial scales cover only three or four orders of magnitude. Many biological effects take place over a timescale 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 [379], and these models may require further examination at the quantum level.

Multiscale 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 [145]. The electric field is determined by effects at the largest scale and causes a change in behaviors at the smallest scale, forcing a restructuring 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. Learning theory [41] is being used to improve standard models used for molecular systems. Recent work [26, 27] suggests that high-dimensional problems like the Schrödinger equation admit accurate low-rank approximations. Density functional theory (DFT) models purport to reduce the complexity of quantum chemistry calculations dramatically, but they require an approximation to the exchange-correlation potential, and at the moment, this remains a heuristic issue, albeit one that is actively studied [92, 256, 434].

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 representation of the many-body forces in biomolecules, so the predictions are not 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 [12]. The first issue is actively being addressed by improving the interaction potentials used in molecular dynamics simulations [112, 253].

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 [447, 531, 532]. Similarly, attempts are being made to represent the affect of ions more accurately [318].

We can anticipate improvements to the models used for biochemical simulation, and we hope that these will contribute to improved computational techniques in molecular biology. In addition to improvements in models at the various scales, we also anticipate advances in linking models at different scales. While this is an extremely difficult problem, significant advances are being made [121, 128].

2.5 Simple Models

The approach we take in this book is to look for simple models that capture the essential features of protein biophysics. These are sometimes referred to as toy models, or cartoon models, but whatever name is used, the intent is to capture the physical mechanisms with computationally tractable models. A key difficulty of the models discussed in Section 2.4 is that they involve significant cancellations. For example, the molecular dynamics model involves numerous positive and negative charges, yet the net charge of most proteins is small, if not zero. Thus, any inaccuracies in the estimation of charge size or location get amplified by a type of round-off error, since the number of such terms is large. By contrast, we study models where the inherent error may be much larger, such as the dehydron, but there is no cancellation due to competing effects.

2.6 Hydrophobic Interactions

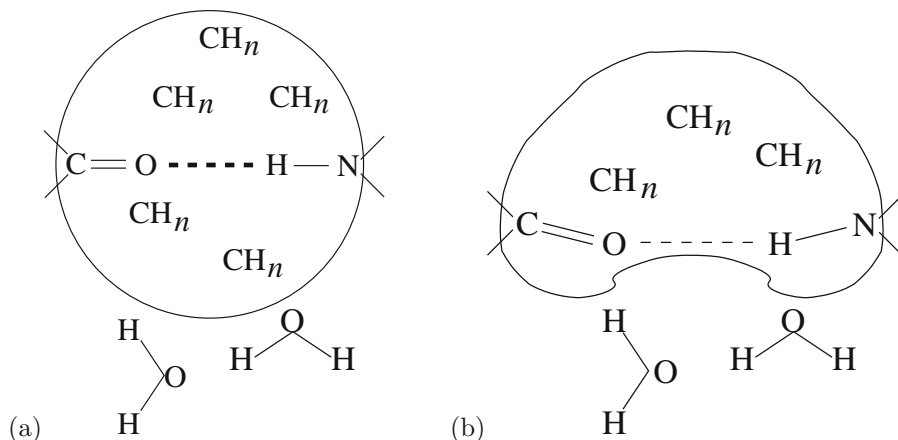


Fig. 2.3 Caricature of wrapping of backbone hydrogen bonds by carbonaceous groups CH_n . (a) Well-wrapped backbone hydrogen bond. (b) Under-wrapped backbone hydrogen bond.

Hydrophobic interactions are sometimes said to be more important than even the hydrogen bond [263]. Although not completely understood, the **hydrophobic force** [48] derives from the **hydrophobic effect** [481]. This effect is one of the central topics of our study. However, the hydrophobic effect has many manifestations in protein behavior.

The hydrophobic effect [48, 481] was proposed in the 1950s [482] as a major contributor to protein structure. However, it is only recently that the detailed nature of hydrophobic forces has been understood. Here, we will characterize a type of intramolecular hydrogen bond, named dehydron, that is strengthened and stabilized as water is removed from its surroundings. Indeed, the dehydron can be viewed as promoting a particular type of hydrophobic effect.

In two recent papers, further understanding of hydrophobic forces have been provided [117, 118]. It was seen that the role of hydrophobic modulation of solvent dielectric is critical to the hydrophobic force [117].

There is a simple view of how hydrophobic forces work. There are certain molecules that are hydrophobic (cf. Section 2.6.2 and Chapter 8), meaning that they repel water. Regions of proteins that have many such molecules, e.g., a protein with a large number of hydrophobic residues on a part of its surface, would tend to prefer association with another such surface to reduce the frustration of having two water–hydrophobe interfaces. It is this simple effect that makes cooking oil form a single blob in water even after it has been dispersed by vigorous stirring.

More precisely, the argument is that the elimination of two hydrophobic surfaces with a water interface is energetically favorable. One could also argue by considering volume changes (cf. Section 5.6) since hydrophobic sidechains 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 [117, 118]. These arguments are compelling, but they suggest a nonspecific interaction. Indeed, hydrophobic attraction leads to nonspecific binding [176].

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

2.6.1 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 [146].

Another type of solvent effect that occurs 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 (see Section 5.2 and Chapter 6) formed by backbone amide or carbonyl groups, either with other backbone or sidechain groups, or with water. In some situations, water removal can cause a switch in the resonance state to a flexible mode [145].

Another example of a change of electrical properties resulting from differences in the water environment involves a more gross change. Proteins which penetrate a cell membrane go from a fully solvated environment to one that is largely solvent-free (inside the membrane). This can be related to a large-scale change in the secondary structure of the protein conformation that has implications for drug delivery [176].

More generally, solvent mediation can alter any electrostatic force via dielectric effects (Chapter 15). Changes in dielectric properties of the environment can have a substantial impact on any electrical property. Much of our study will be related to the dielectric effect and its modulation by hydrophobic groups. But rather than try to address this by introducing a precise model (see Chapter 15), we prefer to introduce the concept by example. We thus begin by looking at one particular example of hydrophobic modulation of the dielectric behavior of water around hydrogen bonds.

2.6.2 Dehydrons

In [174], 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 preformed, but under-protected, hydrogen bonds are effectively adhesive, since water removal from their vicinity contributes to their strength and stability, and thus, they attract partners that make them more viable (see Section 2.6.5 and Chapter 9).

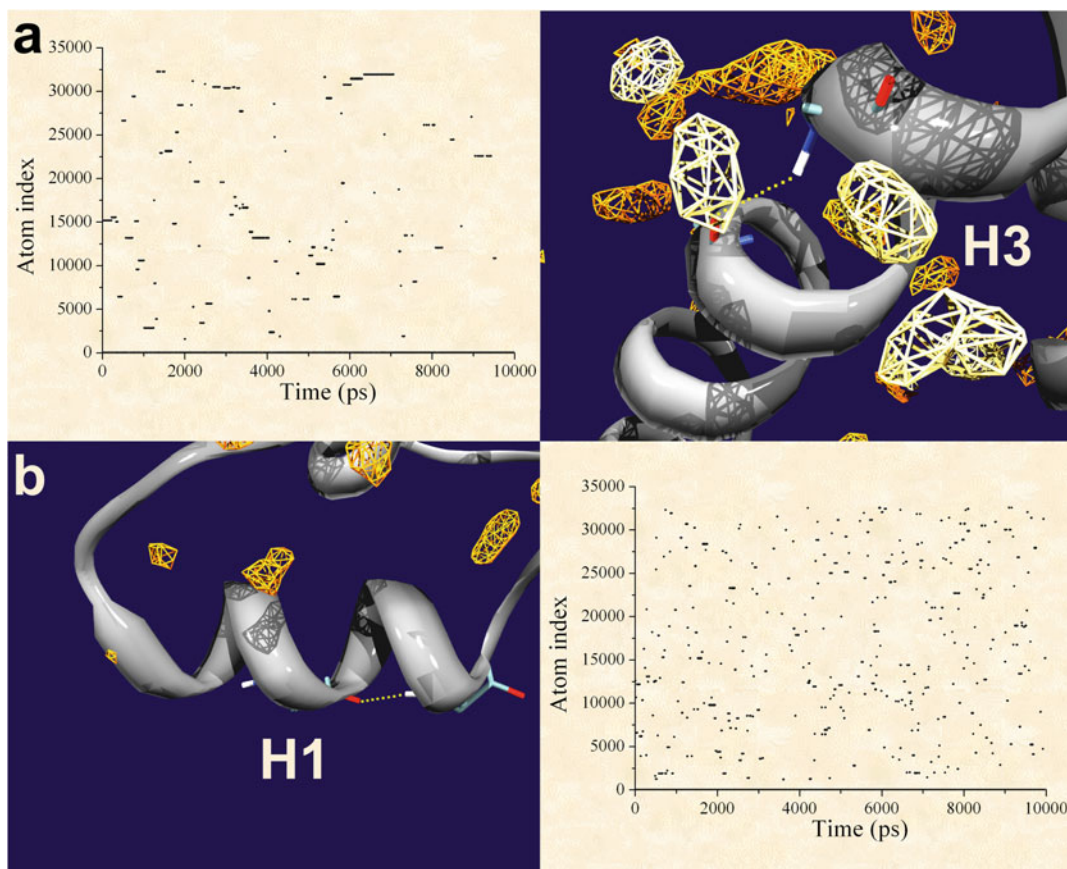


Fig. 2.4 Dynamics of water near hydrogen bonds, adapted from [107, Figure 5]. Plotted on the axes are the residence times of individual waters (numbered as indicated on the vertical axis) as a function of the simulation time (horizontal axis). The orange and yellow polyhedra indicate regions with high residence times for waters. (a) Hydrogen bond in helix H3 is well-wrapped. (b) Hydrogen bond in helix H1 is under-wrapped. The relevant hydrogen bonds are indicated by dotted lines.

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 [48, 481] and its role in modulating the dielectric effect (Chapter 15) through water exclusion. At the simplest level, wrapping occurs when sufficient nonpolar groups (CH_n , $n = 0, 1, 2, 3$) are clustered in the vicinity of intramolecular hydrogen bonds, protecting them by excluding surrounding water [172]. The concept of wrapping of a hydrogen bond is depicted informally in Figure 2.3. A well-wrapped hydrogen bond, Figure 2.3(a), is surrounded by CH_n groups on all sides, and water is kept away from the hydrogen bond formed between the C–O group of one peptide and the N–H group of another peptide (Section 4.1). An under-wrapped hydrogen bond, Figure 2.3(b), allows a closer approach by water to the hydrogen bond, and this tends to disrupt the bond, allowing the distance between the groups to increase and the bond to weaken.

It is possible to identify dehydrons as **under-wrapped hydrogen bonds (UWHB)** by simply counting the number of hydrophobic sidechains in the vicinity of a hydrogen bond. This approach is reviewed in Section 8.3. More accurately, a count of all (nonpolar) carbonaceous groups gives a more refined estimate (Section 8.4). However, it is possible to go further and

quantify a force associated with dehydrons which provides a more refined measure of the effect geometry [174] of the wrappers (Section 8.5).

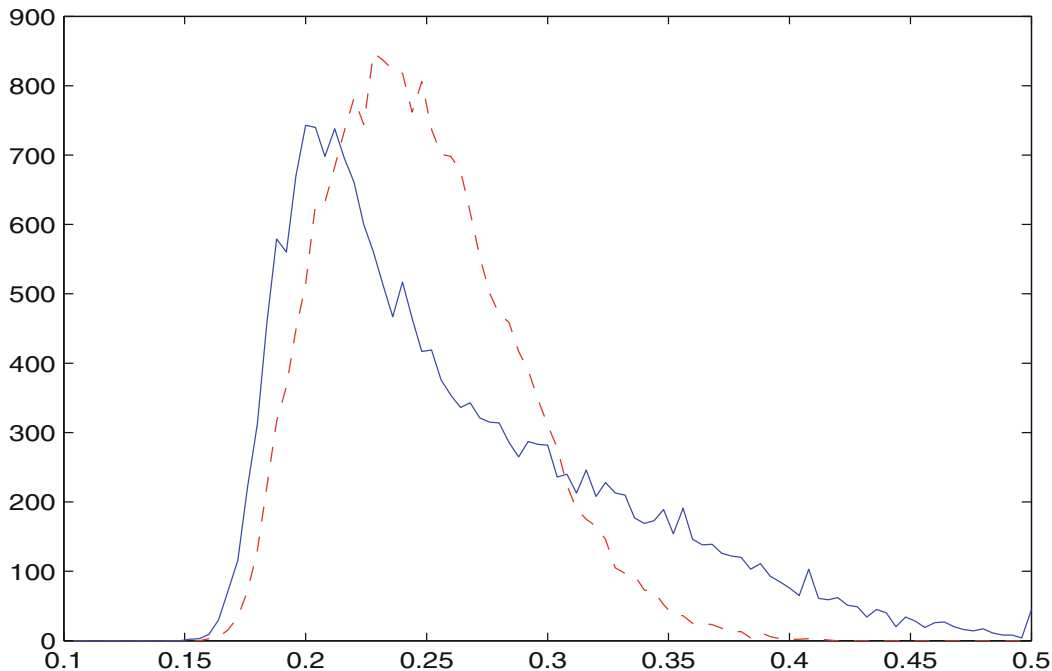


Fig. 2.5 Distribution of bond lengths for two hydrogen bonds formed in a structure of the sheep prion [107]. The horizontal axis is measured in nanometers, whereas the vertical axis represents numbers of occurrences taken from a simulation with 20,000 data points with bin widths of 0.1 Ångstrom. The distribution for the well-wrapped hydrogen bond (H3) has a smaller mean value but a longer (exponential) tail, whereas the distribution for the under-wrapped hydrogen bond (H1) has a larger mean but Gaussian tail.

At the structural level, a significant correlation can be established between dehydrons and sites for protein complexation (Chapter 8). The HIV-1 capsid protein P24 complexed with antibody FAB25.3 provides a dramatic example, as shown in [174, Figure 2] and in a cartoon in Figure 11.1(a).

2.6.3 Dynamics of Dehydrons

The extent of wrapping changes the nature of hydrogen bond [107] and the structure of nearby water [165]. Hydrogen bonds that are not protected from water do not persist [107]. Figure 5 of [107] shows the striking difference of water residence times for well-wrapped and under-wrapped hydrogen bonds. Private communication with the authors of [107] has confirmed that there is a marked difference as well in the fluctuations of the hydrogen bonds themselves. Under-wrapped hydrogen bond lengths are larger (on average) than well-wrapped hydrogen bonds. More strikingly, the distributions of bond lengths as shown in Figure 2.5 are quite different, confirming our prediction based on Figure 2.3 that the coupling of the hydrogen bond characteristics with the water environment would be different.

The H-bond R208–E212 depicted in [107, Figure 5(A)] is well-wrapped, whereas V189–T193 depicted in [107, Figure 5(B)] is a dehydron (see [167, Figure 3a, 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 rebonding with different water molecules [107].

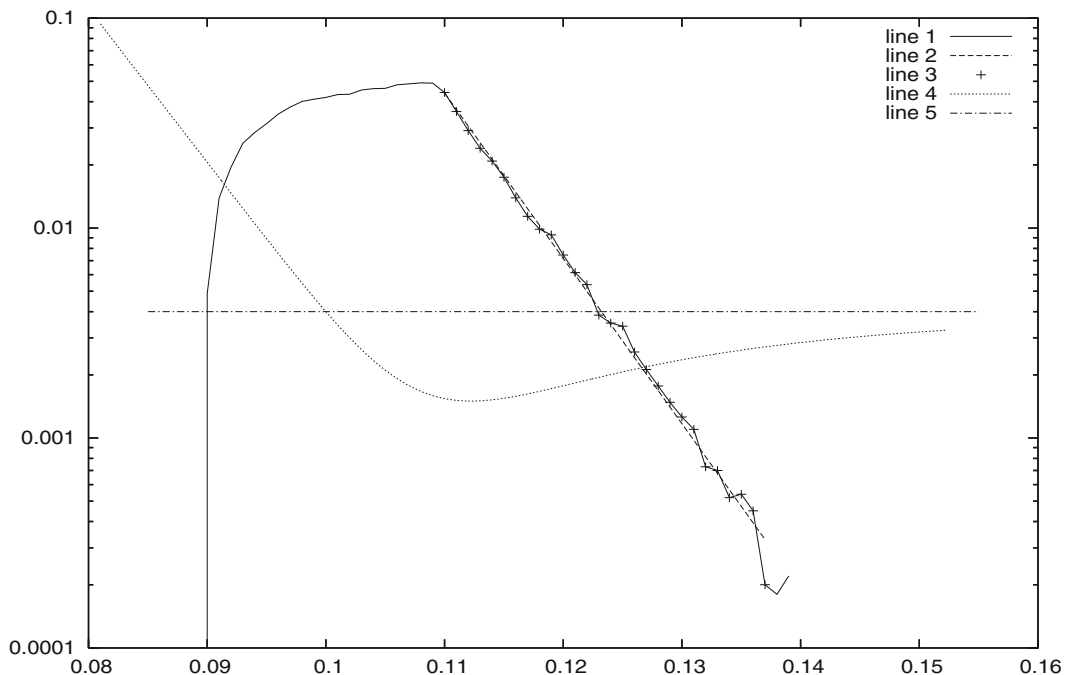


Fig. 2.6 Simulation of a random walk with a restoring force. Shown is the distribution of values x_i defined in (2.5) for 10^5 time steps i , starting with $x_1 = 0.1$, scaled by a factor of 10^{-3} . Also shown is a graph of $\phi + 0.03$ where ϕ is the potential (2.5). The dot-dashed horizontal line provides a reference axis to facilitate seeing where ϕ is positive and negative. The +’s indicate the part of the distribution exhibiting an exponential decay; the dashed line is a least-squares fit to the logarithm of these distribution values. The distribution has been scaled by a factor of 10^{-3} so that it fits on the same plot with ϕ .

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. Second, the fluctuation of the amide and carbonyls comprising the hydrogen bond contributes to a fluctuating electrostatic environment. The bond can switch from the state depicted in Figure 2.3(b) when water is near, to one more like that depicted in Figure 2.3(a) if water molecules move temporarily away. More precisely, the interaction of the bond strength and the local water environment becomes a strongly coupled system for an under-wrapped hydrogen bond, leading to increased fluctuations. For a well-wrapped hydrogen bond, the bond strength and water environment are less strongly coupled.

The distance distribution for under-wrapped hydrogen bonds can be interpreted as reflecting a strong coupling with the thermal fluctuations of the solvent. Thus, we see a

Boltzmann-type distribution for the under-wrapped hydrogen bond distances in Figure 2.5. It is natural to expect the mean distances in this case to be larger than the mean distances for the well-wrapped case, but the tails of the distribution are at first more confusing. The distribution in the under-wrapped case exhibits a Gaussian-like tail (that is, an exponential of the distance squared), whereas the well-wrapped case decays more slowly, like a simple exponential. (See Figure 5.13 for a comparison of these distributions.) Thus, the well-wrapped hydrogen bond is sustaining much larger deviations, even though the typical deviation is much smaller than in the under-wrapped case. To explain how this might occur, we turn to a simulation with a simple model.

2.6.4 Simulated Dynamics

The data in Figure 2.5 can be interpreted via a simulation which is depicted in Figure 2.6. This figure records the distribution of positions for a random walk subject to a restoring force defined by

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

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

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

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

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

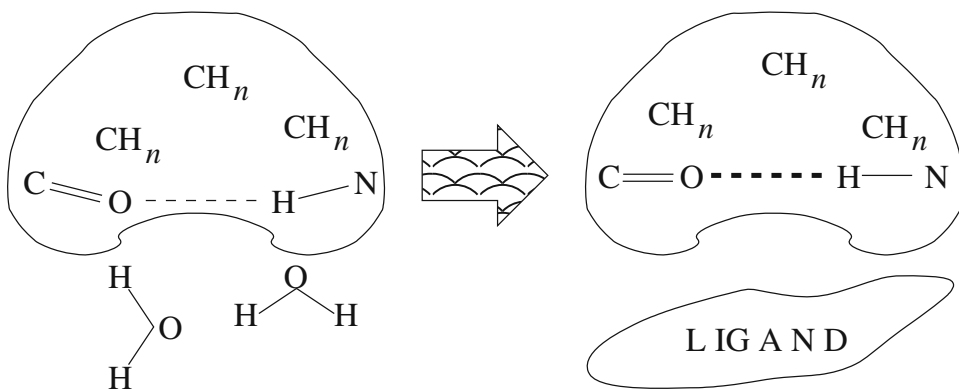


Fig. 2.7 Cartoon showing dehydration due to ligand binding and the resulting strengthening of an under-wrapped hydrogen bond.

2.6.5 Stickiness of Dehydrons

Desolvation of an under-wrapped hydrogen bond can occur when a ligand binds nearby, as depicted in Figure 2.7. The removal of water lowers the dielectric and correspondingly strengthens the hydrogen bond. The resulting change in energy due to the binding effectively means that there is a mechanical force of attraction for a dehydron. This implies that dehydrons are sticky. These concepts are explained in more detail in Chapter 9, and the experimental support for the force of attraction for dehydrons is reviewed as well.

Since dehydrons are sticky, they must be binding sites. Thus, we see in Chapter 7 that they are likely to be found at protein–ligand interfaces. In Chapter 11, we explore several types of protein–ligand interfaces in which dehydrons play a decisive role. Finally, we review in Chapter 13 the role of dehydrons in drug design.

2.7 Dehydron Switch

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

As wrapping is varied, the quality of a hydrogen bond can vary in a highly nonlinear way, as depicted in Figure 2.8. When a hydrogen bond is extremely underwrapped, the addition of one wrapper could have only a small effect, especially if it appears in a region already wrapped. At the other extreme, additional wrappers will have diminishing effects on a fully protected hydrogen bond. In the middle, additional wrappers may have the most effect, leading to a sensitive switch.

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

Unwrapping a dehydron can effectively switch off the hydrogen bond, eliminating a constraint on the protein. This has the potential to increase the entropy of the overall system

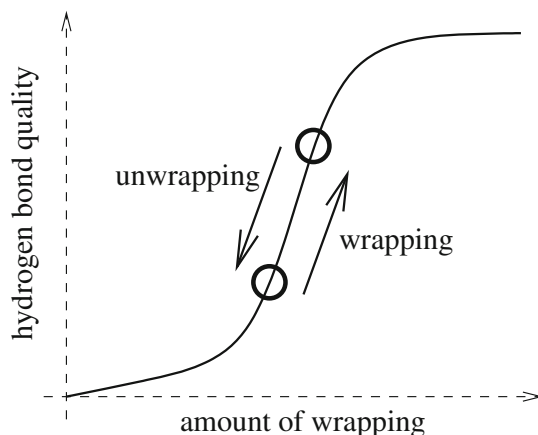


Fig. 2.8 Cartoon depicting the capability of a dehydron to switch on and off when a ligand binds. The binding event may increase wrapping, in which case the hydrogen bond becomes stronger and more stable, or it may decrease wrapping, in which case the hydrogen bond becomes weaker and less stable.

because removing the constraint can increase the degrees of freedom of motion. Thus, both effects have the potential to increase free energy upon association.

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

We will discuss experimental evidence in Section 8.1 that a single nonpolar carbonaceous group can have a measurable effect on the strength and stability of a hydrogen bond. However, these results concern formation of basic protein structure (helices) in an isolated solvent environment. In more complex protein environments, the effect of small amounts of wrapping may have less effect due to environmental constraints. These effects can lead to the more dramatic switch behavior depicted in Figure 2.8.

2.8 Exercises

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

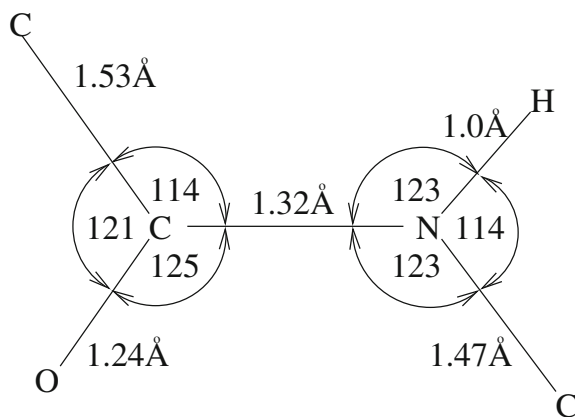


Fig. 2.9 Data reproduced from the figure at the top of page 282 in [382]. Distances are in Ångstroms (indicated by Å) and angles are in degrees. The upper-left and lower-right carbons (C) correspond to sequential C_α carbons in a protein.

Exercise 2.2. Pauling used a simple molecule as a model for the peptide backbone structure in what is called the trans form depicted in Figure 4.2(a). Pauling's data, in the figure at the top of page 282 in [382], is replicated in Figure 2.9. Use this information to compute the distance between C_α carbons according to this model. Explain why you can be sure that these atoms are in a plane (hint: look at the angles). Also use Figure 2.9 to compute the distance between C_α carbons in the cis form as depicted in Figure 4.2(b).

Exercise 2.3. Download a PDB file for a protein and compute the distance distribution between sequential C_α carbons. Instead of plotting the usual distribution, plot the distance on the vertical axis as a function of the order of appearance of C_α carbons in the PDB file. To allow for very small and very large distances, use a log scale for the vertical axis. What causes the outliers? What is the mean of the distribution if you throw out the outliers? Look at PDB file 3SIL and explain why some distances are so small (nearly zero). Look at PDB file 1E08 and explain why one value is significantly smaller than the others. Compare all of this with the distance data in Exercise 2.2.

Exercise 2.4. 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 2.5. Download a PDB file for a protein and compute the N–O distance distribution between the N’s and O’s that are nearest in sequence (cf. Figure 2.1). That is, for each residue i , there is an O_i and an N_i . Plot the distributions for the distances $|O_i - N_i|$ and $|O_i - N_{i+1}|$.

Exercise 2.6. Download a PDB file for a protein and compute the N–O distance distribution between all pairs of carbonyl and amide groups in the peptide bonds (cf. Figure 2.1). Is there part of the distribution that corresponds to pairs forming local interactions (e.g., a hydrogen bond)? Explain your reasoning. Pick a cutoff distance R and plot the distribution of sequence separation k for all N–O pairs whose distance is less than R . Experiment with R between 3 and 7 Å to see how the distribution changes.

Exercise 2.7. Download a PDB file for a protein and compute the N–O distance distribution between each carbonyl and its nearest amide group with which is not covalently bonded in the peptide bonds (cf. Figure 2.1). That is, for each C–O group, find the nearest N and record the N–O distance. Is there part of the distribution that corresponds to pairs forming local interactions (e.g., a hydrogen bond)? Explain your reasoning. Also record the sequence distance k between the N’s and O’s and plot a two-dimensional distributions of k versus physical distance d . (Hint: exclude the N’s and O’s that are near neighbors in the peptide bond backbone.)

A Mathematical Approach to Protein Biophysics

Scott, L.R.; Fernández, A.

2017, XI, 290 p. 110 illus., 27 illus. in color., Hardcover

ISBN: 978-3-319-66031-8