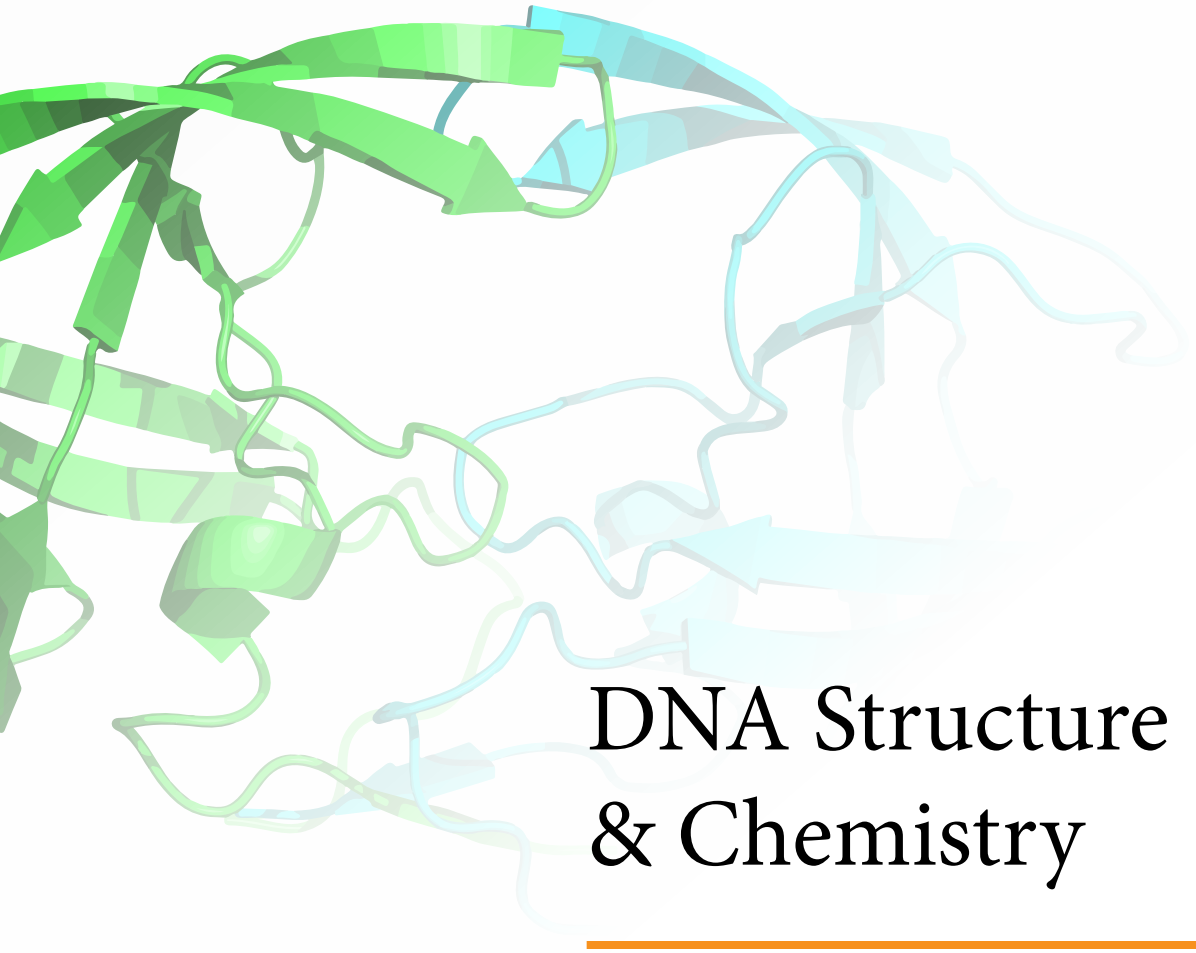


8

DNA Structure & Chemistry

Goal

To understand the structure and chemistry of DNA and the significance of the double helix.

Objectives

After this chapter, you should be able to

- explain the structural basis for the directionality of polynucleotide chains.
- describe how hydrogen bonding and geometry dictate base pairing.
- describe the forces that stabilize the DNA double helix.
- explain the significance of the grooves in the DNA double helix.
- describe how DNA is compacted into chromosomes.

Proteins, as we have seen, are the workhorses of the cell; they exhibit extraordinarily varied structures, which enable them to perform myriad tasks. This is in sharp contrast to **deoxyribonucleic acid** (**DNA**) molecules, which with few exceptions exhibit a single, common structure: the double helix. This makes sense because DNA has just one function, information storage. Just as a compact disc can contain the blueprint for building a house, DNA is a storage depot for the information needed to make the diverse proteins of the cell. The instructions for specifying the primary structure of a protein are stored in a stretch of DNA known as a gene. Almost all of the carbon-containing molecules in living systems are either proteins that are directly encoded by DNA or non-protein molecules that are generated by the actions of enzymes, which like other proteins are themselves specified by DNA. In Chapters 8-12 we will learn how the cell replicates DNA molecules and does so faithfully, how it retrieves genetic information to direct the synthesis of proteins, and how it regulates this flow of genetic information from DNA to protein. But first, in this chapter we look closely at the structure and chemistry of DNA in order to learn how its double-helical architecture allows information to be stored, duplicated, and accessed.

Each DNA strand is an alternating copolymer of phosphates and deoxyribose sugars

As its name implies, the double helix is composed of two polynucleotide chains that are wrapped around each other as helices. We focus first on the

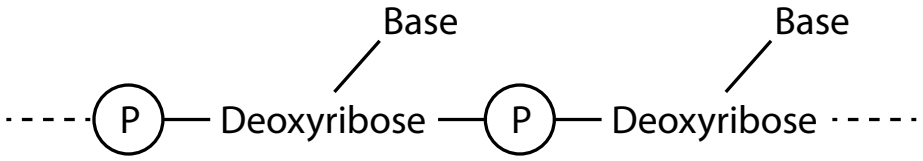
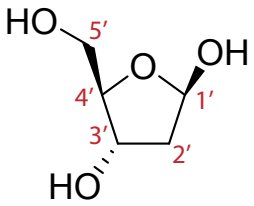


Figure 1 **DNA is an alternating polymer of deoxyribose and phosphate**



2′-deoxyribose

Figure 2 **The carbon atoms in deoxyribose are numbered 1' to 5'**

chemical nature of the individual chains. Each is a copolymer composed *alternately* of phosphate groups and sugar units (Figure 1). The sugar units are pivotal to the chains, as the four **nucleobases** (or more simply **bases**), which we will consider soon, are attached to the sugars. The sugars are pentose sugars, meaning that they contain five carbon atoms. The positions of the carbons in the sugars are labeled with apostrophes, which are referred to as primes (1'-5'), to distinguish them from the numbering of positions in the bases (Figure 2). The bases are attached to the sugars at the 1' ("one-prime") position via a **glycosidic** linkage (simply meaning that the base is bonded to a sugar). Notice that the sugars are five-membered rings in which an oxygen atom links the 1' and 4' carbons. Notice also that the 5' carbon is off the ring, being attached to it via the 4' carbon. Thus, only four of the five carbons contribute to the five-membered ring.

The sugars in DNA are **2'-deoxyribose** sugars in that the carbon at position 2' lacks a hydroxyl group and instead has two hydrogen atoms. RNA, which we will consider in a later chapter, is a similar copolymer, but its sugars are **ribose** sugars, which contain a hydroxyl group at the 2' position in place of one of the two hydrogen atoms.

Importantly, the phosphates in the alternating copolymer are joined to the sugars at the 3' and 5' positions via **phosphate ester** linkages (in which a phosphorous atom double-bonded to oxygen is joined to a carbon-containing group via a second oxygen atom) (Figure 3). Because the phosphates are joined to the sugars through two ester linkages, these are said to be **phosphodiester** bonds.

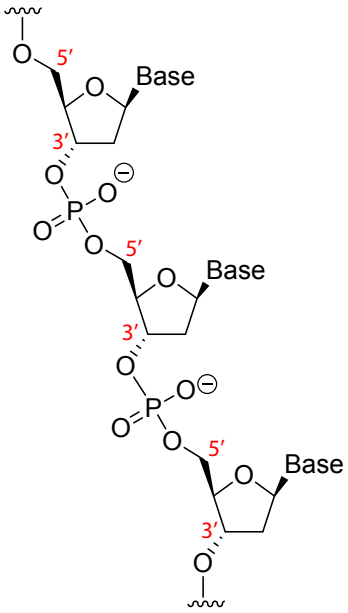


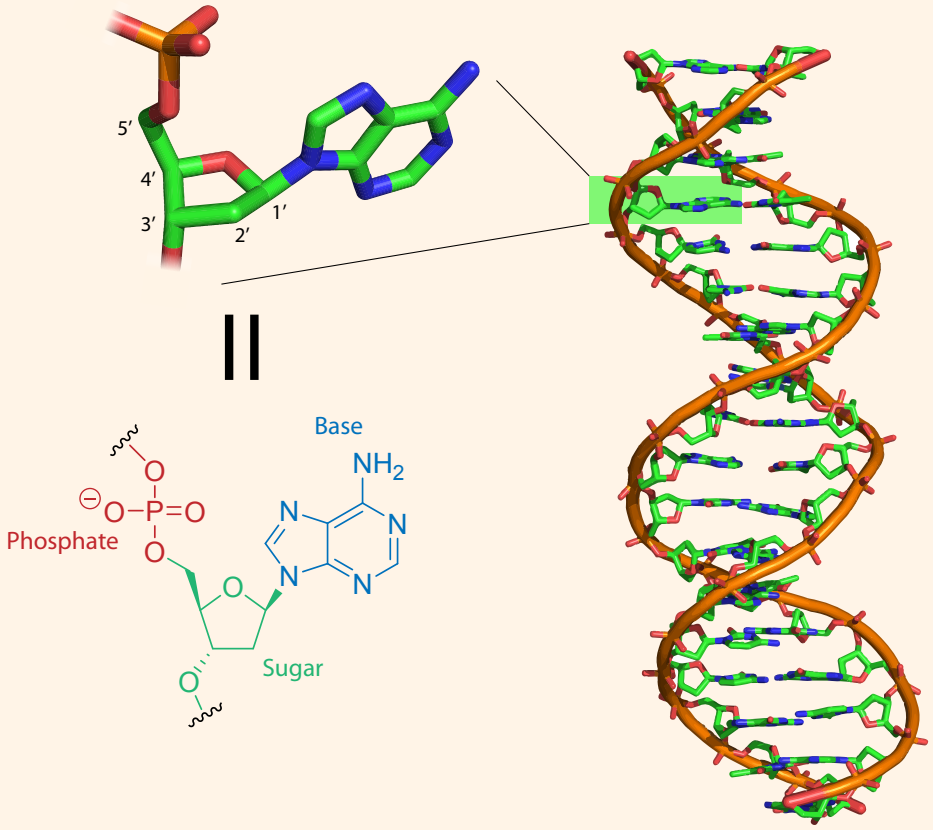
Figure 3 **Phosphodiester linkages connect the 3' carbon of one deoxyribose sugar to the 5' carbon of the adjacent sugar in the DNA polymer**

Box 1 Nucleotides are tripartite repeating units in polynucleotides

We have so far presented the polynucleotide chain as an alternating copolymer of phosphates and deoxyribose sugars to which bases are attached. An alternative way to think about the chain is as a simple polymer of units consisting of a phosphate, a sugar, and a base. Such tripartite units are referred to as **nucleotides**, hence the name polynucleotide. Nucleotides in DNA have a single phosphate, but free nucleotides can have two or three (and sometimes more) phosphate groups. Nucleotides bearing three phosphates at the 5' position of deoxyribose will become important in subsequent chapters when we consider the biosynthesis of polynucleotide chains. A bipartite structure consisting of a sugar and a base but no phosphates is referred to as a **nucleoside**.

Figure 4 Polynucleotides are also polymers of nucleotides

Shown is a single nucleotide within a DNA double helix and its three-dimensional stick representation as well as a corresponding standard line drawing. Atoms in the three-dimensional structure are colored as follows: carbon, green; oxygen, red; nitrogen, blue; phosphorus, orange.



Polynucleotide chains have a 5'-to-3' directionality and align in an anti-parallel orientation in the double helix

A fundamental feature of the polynucleotide chain is that its ends are dissimilar. Thus, the 3' hydroxyl is displayed at one terminus, the 3' end, and the 5' hydroxyl at the other terminus, the 5' end. This means that polynucleotide chains have an intrinsic directionality. This is analogous to the directionality of polypeptide chains, which, as we have seen, have an N-terminus and a C-terminus.

Since the double helix consists of two polynucleotide strands, what is the orientation of the two helical strands relative to each other? The answer is that the two strands are oriented such that the 5'-to-3' directionality of one strand aligns with the 3'-to-5' directionality of the other strand. That the directionality of the two strands is anti-parallel is an invariant rule that governs the interaction of polynucleotide chains (RNA as well as DNA) with each other.

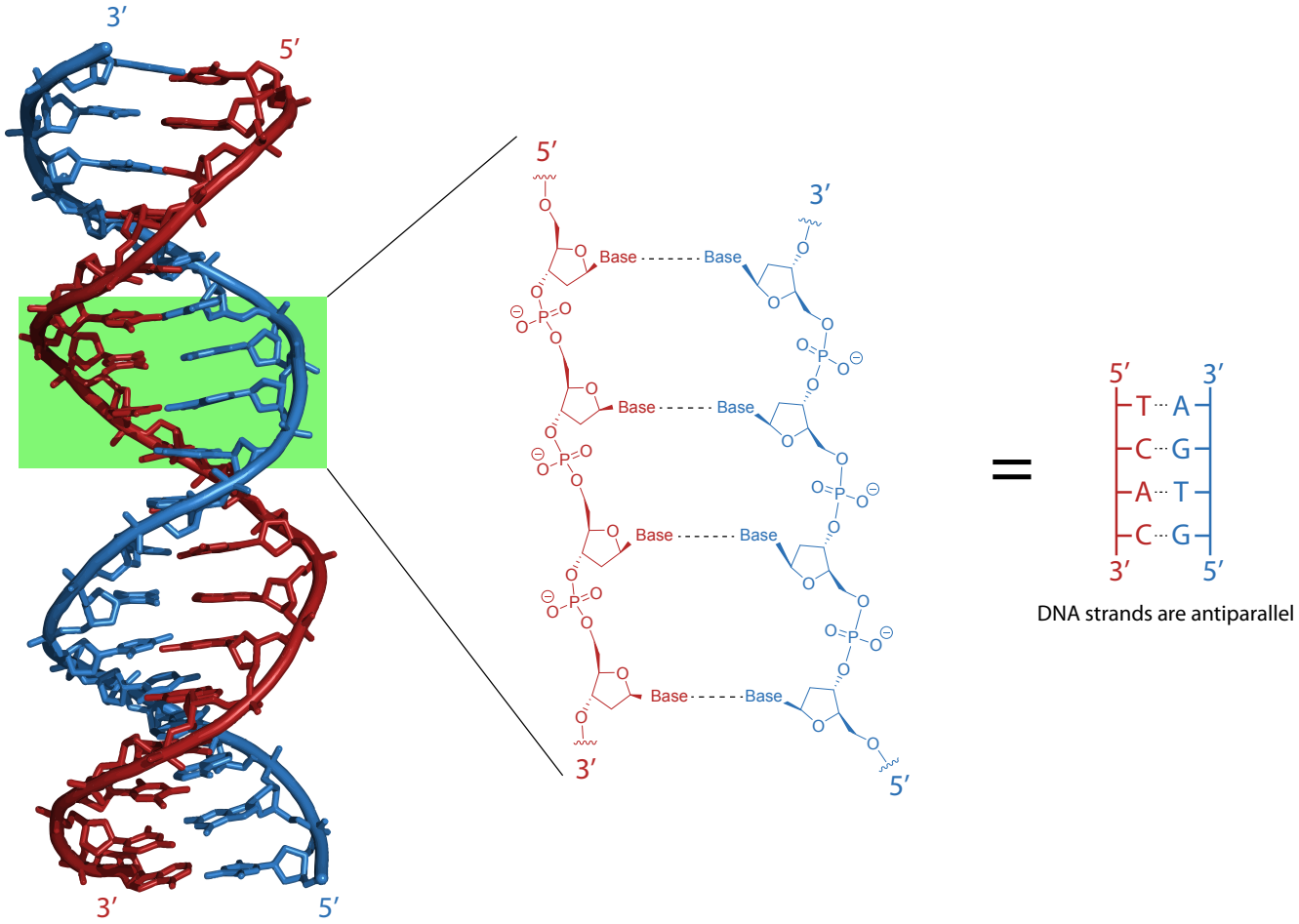


Figure 5 **Polynucleotide chains are antiparallel in the double helix**

A:T and C:G are abbreviations for the bases and their pairing (adenine paired with thymine and cytosine paired with guanine), as we come to next.

As we will consider in detail in later chapters, the directionality of polynucleotides and of polypeptides is the basis for three foundational rules that govern the duplication and retrieval of genetic information. These are that: (1) the synthesis of polynucleotide chains always proceeds in a 5'-to-3' direction, (2) the synthesis of polypeptide chains always proceeds in an N-to-C-terminal direction, and (3) the information for amino acid sequences from the N-terminal amino acid to the C-terminal amino acid is specified sequentially in a 5'-to-3' direction in polynucleotide chains.

The two strands of the double helix interact with each other via two pairs of complementary bases

Each deoxyribose in the polynucleotide chain is attached to one of four bases that mediate interactions between the two strands of the double helix. Bases are flat rings consisting of carbon and nitrogen atoms; they are said to be **heterocyclic** because they are composed of rings containing other atoms than just carbon. There are two kinds of bases, pyrimidines and purines. Pyrimidines are single, six-membered heterocyclic rings, whereas purines have a bicyclic structure consisting of five- and six-membered heterocyclic rings. The positions of carbon and nitrogen atoms in the pyrimidine and purine rings are numbered as shown in Figure 6. Pyrimidines are joined

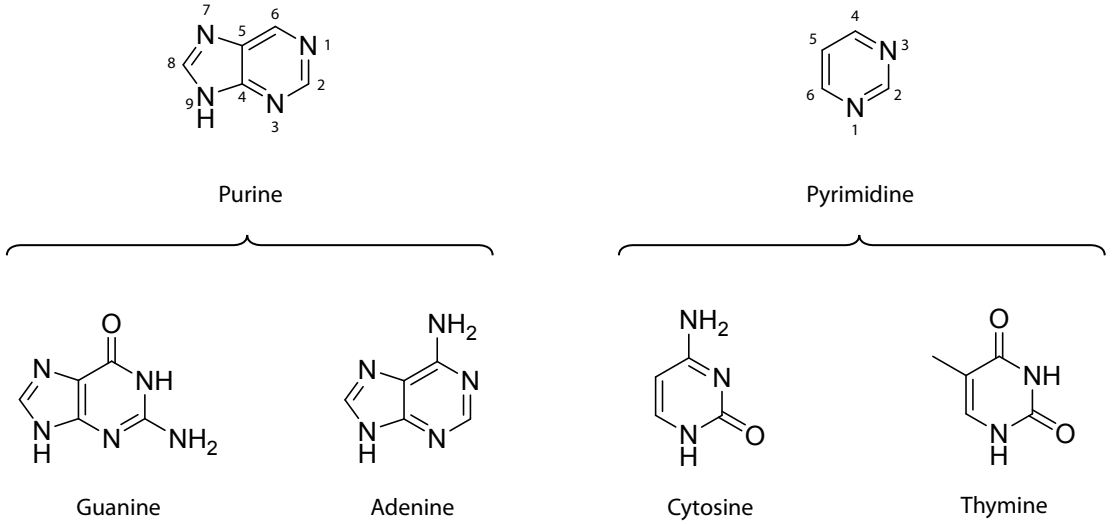


Figure 6 **The bases are purines and pyrimidines**

The four bases in DNA belong to two families whose numbering systems are shown.

via a glycosidic linkage to the 1' position of deoxyribose via the nitrogen at position 1, whereas purines are attached via the nitrogen at position 9.

The pyrimidines are **cytosine** (**C**) and **thymine** (**T**), and the purines are **adenine** (**A**) and **guanine** (**G**) (Figure 6). Base pairs consist of a pyrimidine and a purine such that cytosine pairs uniquely with guanine (**C:G**) and thymine specifically with adenine (**T:A**) (Figure 7). (As we will see in Chapter 10, RNA contains the pyrimidine uracil in place of thymine; like thymine, uracil pairs with adenine.) Compared to the 20 disparate side chains of amino acids, the four bases are relatively similar-looking. In contrast to side chains, however, they all principally do one thing: pair with a single complementary base.

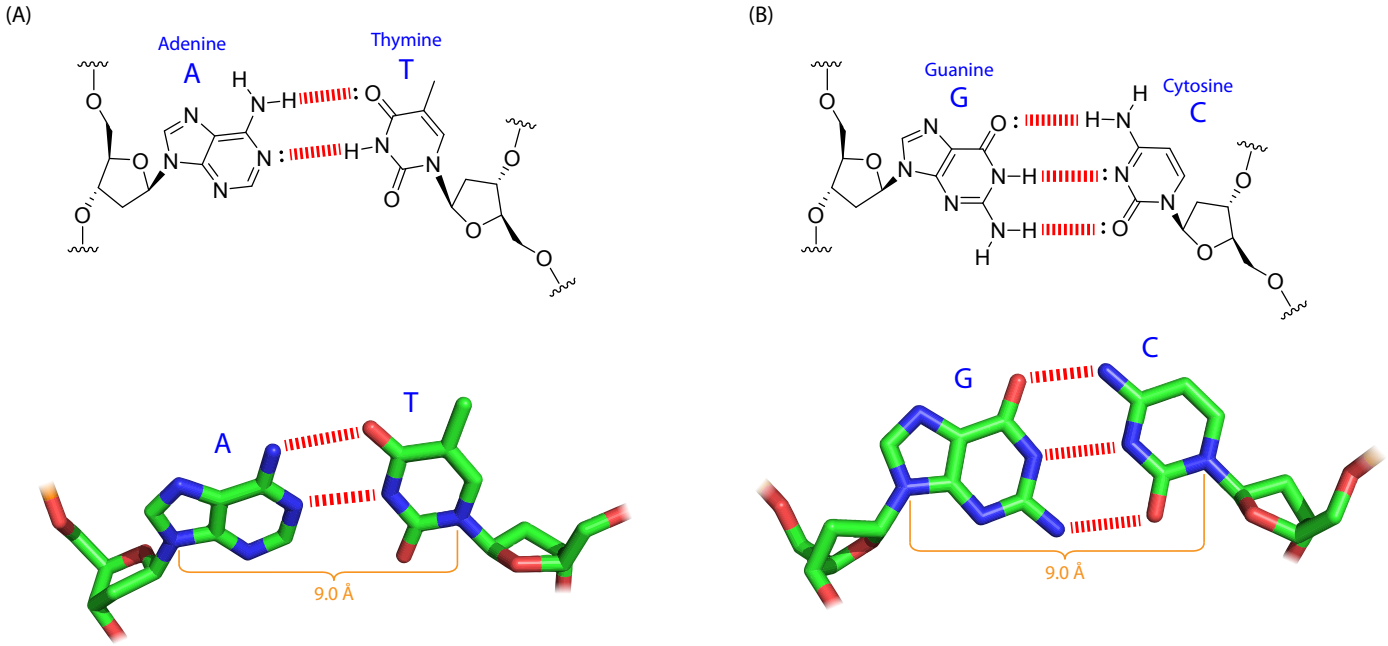


Figure 7 **Each base specifically pairs with one other base**

Adenine and thymine form two complementary hydrogen bonds to form the A:T base pair (A), whereas guanine and cytosine form three complementary hydrogen bonds to form the G:C base pair (B). Both the A:T and G:C base pairs have the same width, as shown in orange.

What is the basis for the *specific* pairing of adenine with thymine and cytosine with guanine? The specificity of pairing is dictated by hydrogen bonding and geometry. Hydrogen bond donors and acceptors in adenine and thymine and in guanine and cytosine align in the double helix to allow two and three hydrogen bonds to form, respectively, between the bases (Figure 7). Notice that both exocyclic groups (functional groups outside the rings, such as the exocyclic amino group attached at position 6 in adenine) and ring atoms (such as the ring nitrogen at position 3 in thymine) participate in base pairing. Notice also that the bonds between the bases are straight, in accordance with the rule that strong hydrogen bonding requires that the bond angle between the donor and the acceptor be close to 180°. The contribution of geometry to complementarity stems from the fact that G:C and A:T base pairs have the same dimensions across their long axes. This matters, as we will see presently, in that it allows the double helix to maintain a uniform diameter (20 Å) while accommodating two kinds of base pairs each in two possible orientations.

The double helix embeds irregularity within regularity

What is remarkable about the double helix as an information carrier is that it embeds irregularity in a regular structure. DNA is irregular because of the irregular order of the base pairs. This is what makes DNA an information carrier. Notice that A:T and G:C base pairs are accommodated in either orientation in the double helix; any of the four bases can be present on either strand at any one position. Base pairs provide four units of information (G, C, T or A) in a linear sequence.

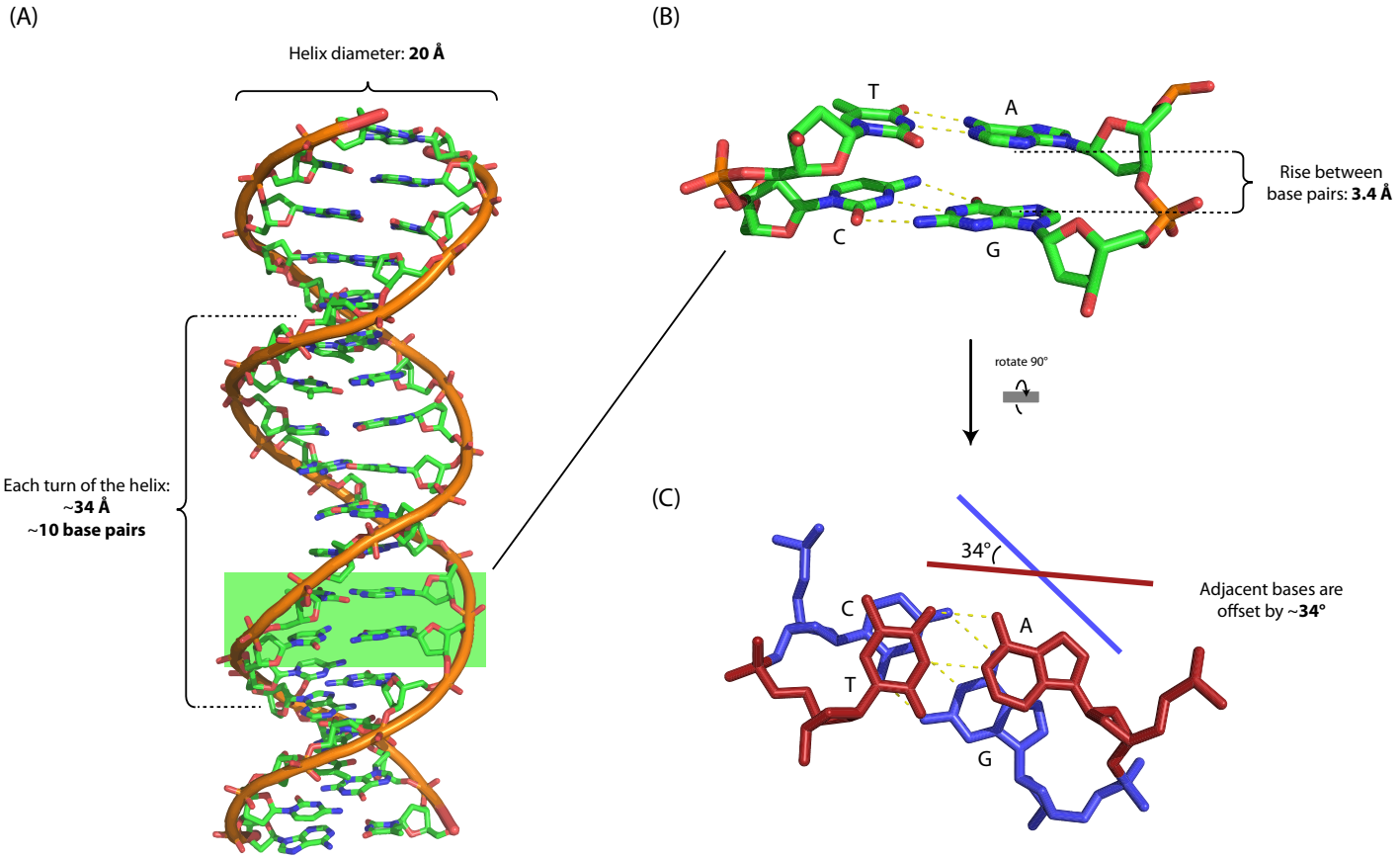


Figure 8 **The DNA double helix has a uniform structure along its length**

But the overall architecture of the double helix is regular (Figure 8). That is, and as mentioned above, A:T and G:C base pairs are equally accommodated without distorting the diameter (20 Å) of the double helix. Each base pair is displaced from its neighbor by a rotational angle of ~34°, resulting in ~10 base pairs per turn of the helix (360°). The thickness (rise) of the bases is ~3.4 Å; hence each turn of the helix is ~34 Å in length. Even though the sequence of base pairs varies, the helical structure is largely consistent along the length of the DNA. This consistency is the opposite of proteins, in which irregularity in the order of amino acids results in extraordinary diversity in protein structure; it underscores the fundamental difference between the roles these two kinds of macromolecules play in living systems.

An additional noteworthy feature of the double helix is its handedness. We learned in Chapter 6 that the α-helix of proteins is right-handed. Likewise, the two helices of the double helix are each right-handed. You can convince yourself of this using the physicists' right-hand rule. In your mind's eye, imagine your right hand wrapped around the DNA molecule in Figure 8 with your thumb pointing along the long axis of one of the strands. Your fingers will follow the twist of the helical backbone in the direction that your thumb is pointing. Now try this with your left hand; it doesn't work! Therefore, DNA is a right-handed double helix.

Box 2 History of determining the structure of the double helix

How was the structure of the double helix determined? Jim Watson and Francis Crick famously solved the structure by model building. In particular, physical models of the bases led to the discovery of base pairing and the contributions of hydrogen bonding and geometry. But the discovery of the double-helical structure of DNA benefited importantly from the work of chemist Rosalind Franklin, who carried out X-ray diffraction studies of oriented fibers of DNA. Her iconic X-ray photograph of the diffraction pattern, "Photo 51", was famously shared with Watson and Crick without her knowledge. Embedded in "Photo 51," perhaps the most famous image of an experimental result in all of science, were clues to the principal features of the double helix: the "Maltese cross" revealing that DNA is helical, the missing "layer line" revealing that the helix is double-stranded, and the intense reflection at 3.4 Å revealing the rise (thickness) of base pairs (Figure 9).

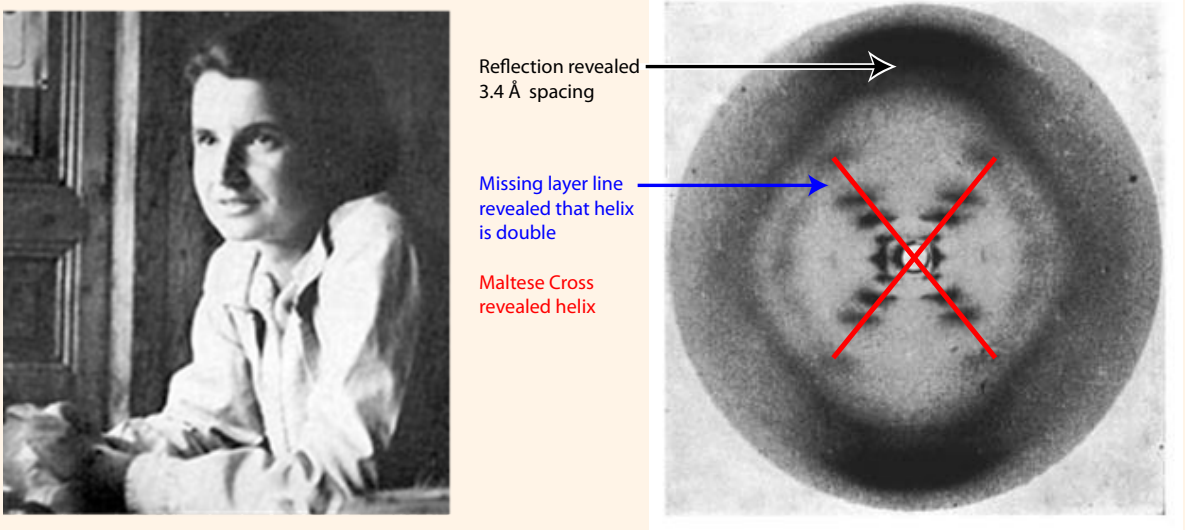


Figure 9 **Rosalind Franklin's X-ray crystallography experiments were crucial to the discovery of the DNA double helix**

The structure of DNA suggested a mechanism for the duplication of the genetic material

Why was the structure of DNA of such epic significance when it was revealed by Watson and Crick in 1953? The answer is that it solved the historic riddle of how the genetic material is duplicated. One of the fathers of the field of molecular biology, the Nobel Prize-winning physicist Max Delbrück, believed that the riddle could not be solved without invoking new principles of chemistry or physics. But the structure of the double helix suggested a simple mechanism that could be understood in terms of existing concepts. The fact that the individual strands of double-stranded DNA form a string of base pairs meant, as we have seen, that the two strands in a double helix are complementary. In other words, the exact sequence of bases in either strand can be deduced from the sequence of the opposite strand simply by pairing A with T, C with G, G with C, and T with A. That DNA is a self-complementary double helix meant that each strand could serve as a template for the synthesis of its complement, thereby solving one of the great intellectual mysteries in biology (Figure 10). That the structure provided a simple solution to the mystery was noted but not spelled out by Watson and Crick in their landmark 1953 publication in one of the most famous understatements in the history of science: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." The next chapter will present a classic experiment that proved that this copying mechanism is correct and discuss the chemical mechanisms by which template-directed synthesis of polynucleotide chains takes place.

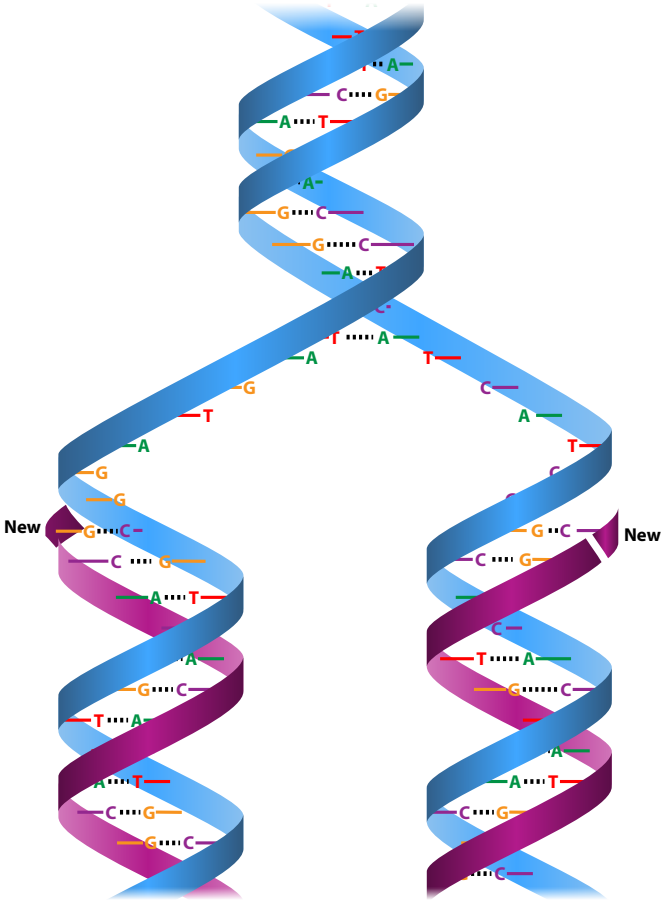


Figure 10 **The self-complementarity of the DNA double helix suggested a simple mechanism for the duplication of the genetic material in which each strand serves as a template for the synthesis of a new, complementary strand**

Box 3 The negative charges of the phosphate groups in DNA protect the polynucleotide chain from hydrolysis

The conjugate acids of the phosphate groups in the DNA backbone have a pK_a of about 2. So at physiological pH they are deprotonated and carry negative charges (Figure 11). These negative charges usually form ionic interactions with positively charged ions such as magnesium ions, and in cells they are often also associated with positively charged proteins. The formation of ionic interactions between positively charged amino acids, such as arginine and lysine, and negatively charged phosphate groups is one of the primary mechanisms by which living systems interact with and manipulate DNA.

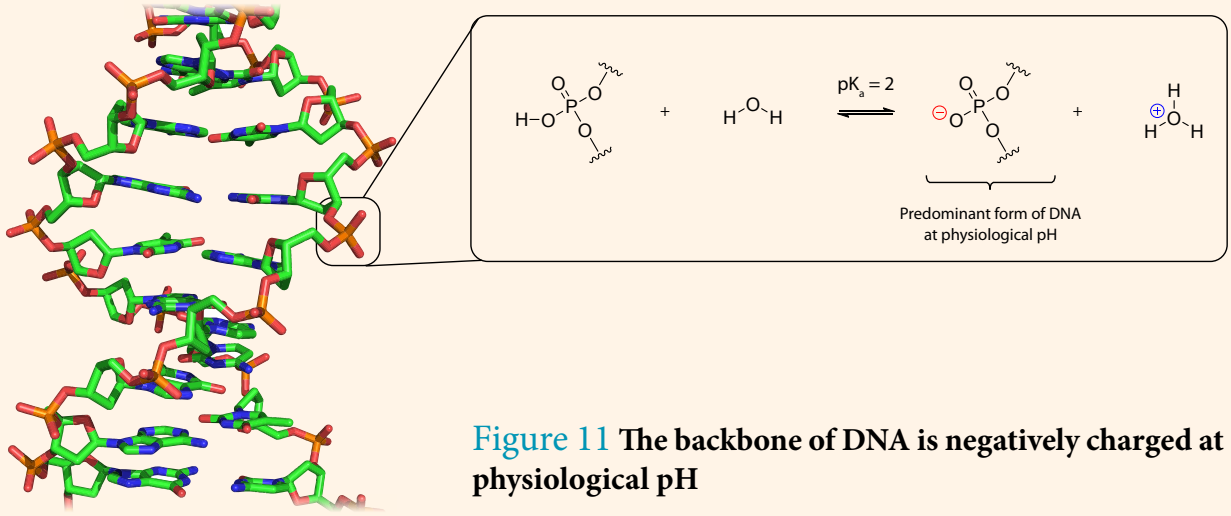


Figure 11 **The backbone of DNA is negatively charged at physiological pH**

An important consequence of the negative charges on DNA is that they contribute to the stability of the DNA strand. DNA's crucial role in information storage requires that DNA be resistant to the common ways by which molecules spontaneously degrade. One of the most common ways that the molecules of life degrade in water is through a process called hydrolysis, as we saw for proteins in Chapter 4. In the case of DNA, hydrolysis results in breakage of the DNA backbone. The first step in DNA hydrolysis involves bringing the non-bonded electron pairs on the oxygen atom of water into close proximity with one of the phosphorus atoms in the DNA backbone. In this process oxygen acts as the nucleophile, and phosphorus acts as the electrophile. As we have previously seen, electrons repel each other, and when the non-bonded electron pairs of an oxygen atom approach the phosphorus atom in a phosphate group, they are strongly repelled by the negative charges of the oxygen atoms that surround the phosphorus atom in phosphate (Figure 12). In other words, the abundance of electron density on the phosphate oxygen atoms effectively shields the phosphorus atom from nucleophilic attack by water, making phosphorus a poor electrophile. Indeed, when DNA is chemically modified so as to remove these negative charges, it becomes much more prone to hydrolysis.

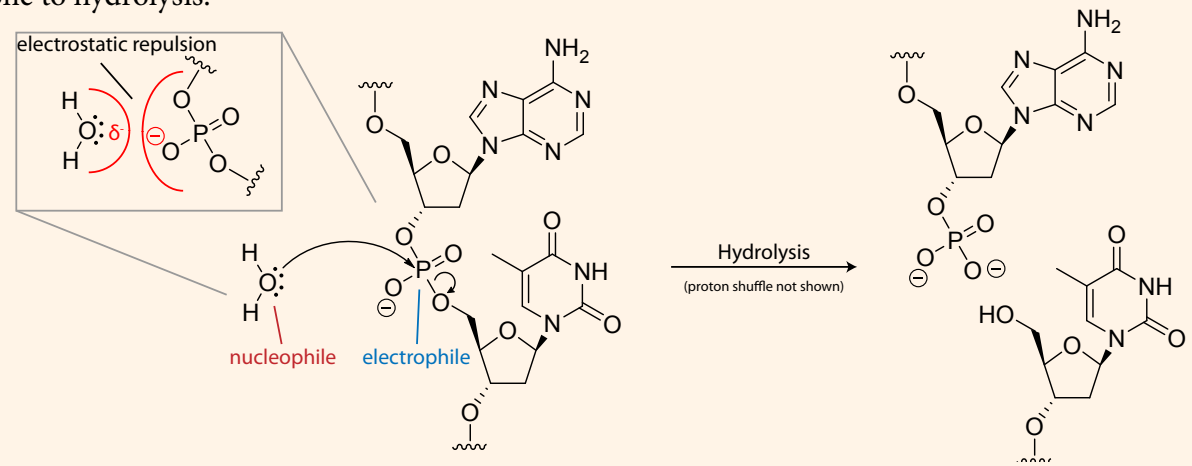


Figure 12 **DNA's negative charge impedes its hydrolysis**

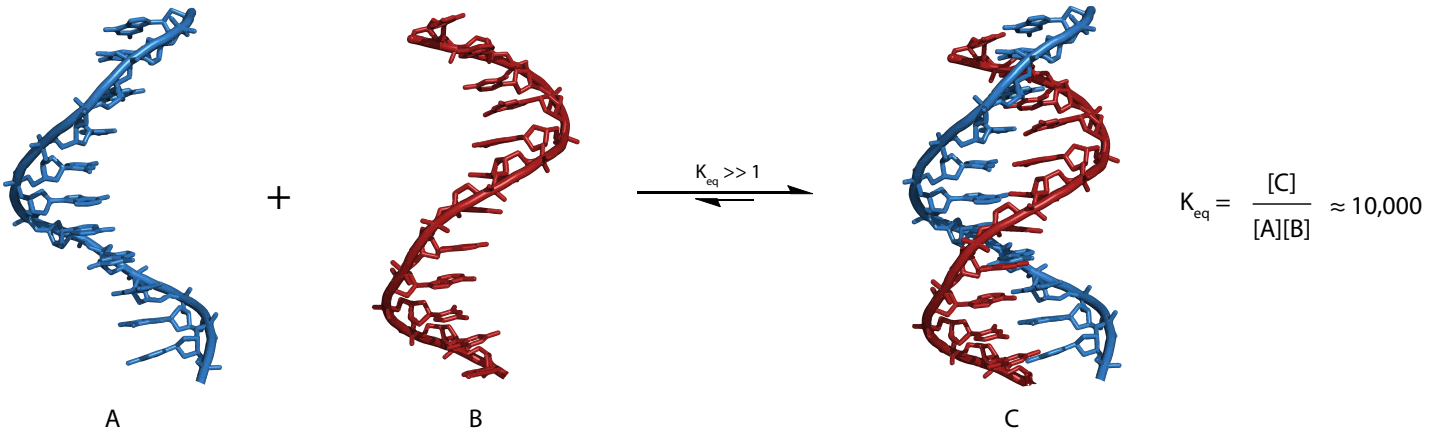


Figure 13 **Annealing of DNA strands is thermodynamically favorable**

Double-stranded DNA is thermodynamically favored over single-stranded DNA

Double-stranded DNA consists of two polynucleotide chains that are wrapped around each other but are not covalently linked. Is this double-stranded structure stable? Is its formation favored over the state of two separate single strands of DNA? Suppose we caused the two complementary strands of a double helix to come apart (for example, by heating them in solution) and then lowered the temperature under physiological conditions of pH and ionic strength. Would the two strands re-**anneal**? In other words, would this be a thermodynamically favorable reaction? Consider that at neutral pH, the phosphate groups in the DNA backbone are negatively charged, which would cause repulsion between the strands (although under physiological conditions charge repulsion is mitigated by cations, such as sodium and magnesium). Also, going from a state of two separate strands to a double helix results in a decrease in entropy ($\Delta S_{DNA} < 0$); single-stranded DNA is more flexible than double-stranded DNA and going from two individual strands to a double helix results in a decrease in disorder. Nonetheless, formation of double-stranded DNA is highly favored over single-stranded DNA, and the equilibrium constant for the formation of a double helix from two unpaired, complementary strands of DNA is approximately 10,000 (Figure 13).

What then accounts for the relative stability of double-stranded DNA? As we now explain, multiple factors contribute to the favorability of the annealing reaction, but the stacking of bases in the double helix is paramount.

The double helix is principally stabilized by base stacking

The primary factor that stabilizes the double helix is **base stacking**, which is the layering of bases on top of each other in the double-stranded structure. Base stacking is thermodynamically favorable because of the hydrophobic effect and because of van der Waals interactions between stacked bases. At first glance the bases might not appear to be hydrophobic, as they contain several polar bonds. However, all of the partial charges are located on the edges of the bases, whereas the top and bottom surfaces of these

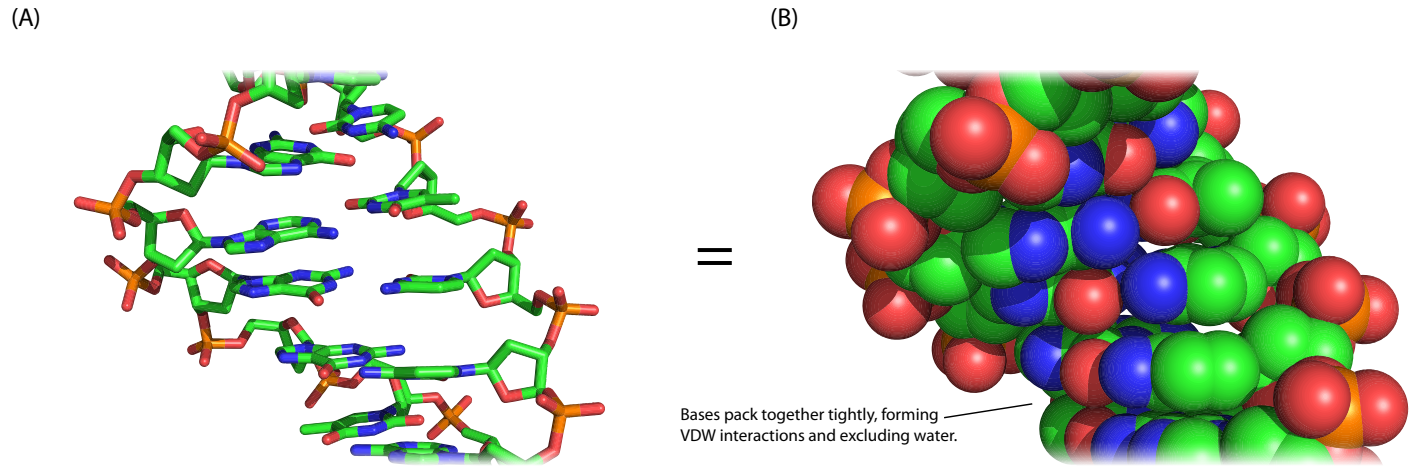


Figure 14 Base stacking results in van der Waals interactions and the exclusion of ordered water

The stick representation of a DNA double helix shown in (A) can be misleading, as it seems to suggest that there are large spaces between adjacent base pairs. The space-filling model shown in (B) shows the amount of space occupied by each atom in the structure. This representation clearly shows that there is no free space between stacked bases. VDW, van der Waals.

flat molecules are relatively nonpolar. When DNA forms a double helix, the hydrophobic tops and bottoms of the bases are stacked next to one another and hidden from water (Figure 14). Because the flat surfaces of the bases are hydrophobic, water forms ordered structures around them when DNA is single-stranded, but when DNA strands anneal, the ordered water molecules are released, resulting in an increase in entropy ($\Delta S_{water} > 0$). This is the same principle that we encountered in Chapter 6 regarding protein folding, in which the favorability of protein folding was in part due to ordered water molecules that were released when hydrophobic side chains were sequestered in the protein's interior. Base stacking also contributes importantly to the enthalpic favorability of the double helix because it allows van der Waals interactions to form between instantaneous dipoles in the hydrophobic surfaces of the bases.

Hydrogen bonds between paired bases contribute modestly to the stability of the double helix but are critical for the specificity of base pairing

Hydrogen bonds between bases contribute to the stability of DNA but not as much as might be expected. The reason for this is that DNA is bathed in a high concentration of water, and water molecules are excellent hydrogen bond donors and acceptors. As a consequence, the hydrogen bonds that form between A and T or between C and G during annealing replace hydrogen bonds that were present between the edges of unpaired bases and water when the DNA was single-stranded (Figure 15A). The hydrogen bonds between the bases and water when DNA is single-stranded may not be as enthalpically favorable as those between paired bases in double-stranded DNA (in which the bonds are in the optimal straight alignment of dipoles). Nonetheless, because annealing replaces one set of hydrogen bonds with another, the overall contribution of hydrogen bonding to the stability of the double helix is not the major driving force in the formation

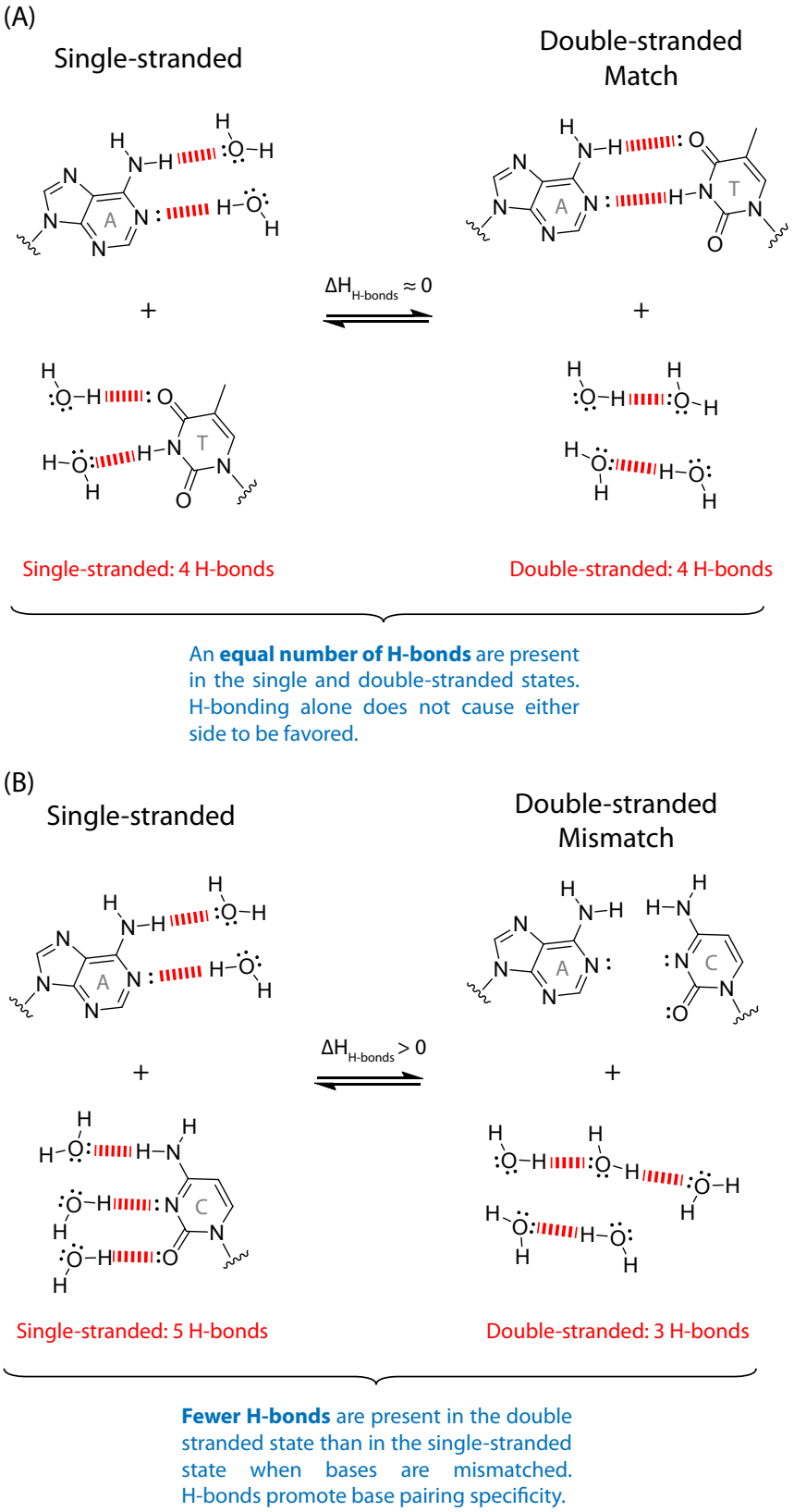


Figure 15 Hydrogen bonds contribute modestly to the stability of the double helix but are responsible for the specificity of base pairing

(A) An equal number of hydrogen bonds are formed in the single-stranded state as in the double-stranded state. (B) In the case of a mismatched A:C base pair, fewer hydrogen bonds are formed in the double-stranded state.

of the double helix. You will recall that a similar concept applies to protein folding, in which the energetic contribution of hydrogen bonds between polar side chains in the tertiary structure of a protein needs to be weighed against the favorability of hydrogen bond formation with water molecules when the protein is unfolded. Thus, for both proteins and DNA, hydrogen bond formation is less important to stability than is the hydrophobic effect, in which water molecules are excluded from hydrophobic surfaces.

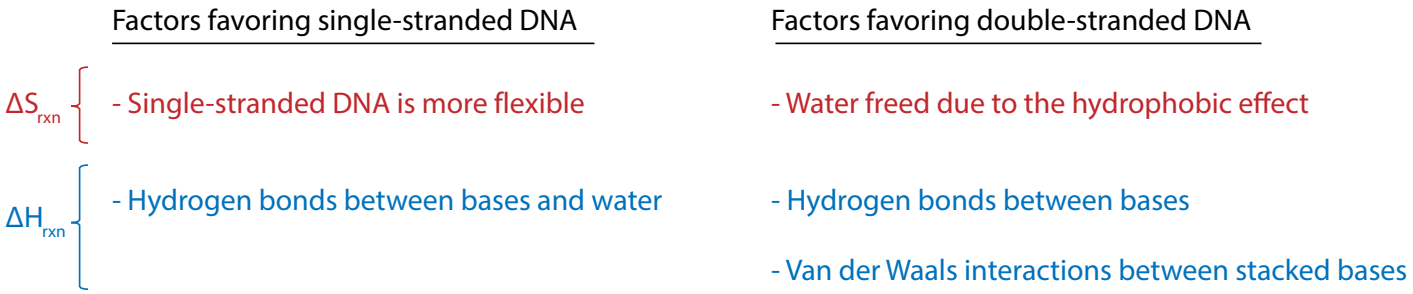


Figure 16 **Multiple factors affect ΔS_{rxn} and ΔH_{rxn} for DNA annealing**

In light of these considerations, we can ask whether hydrogen bonds are truly important in base pairing and, if so, why and when? The answer is that they are critically important for the *specificity* of base pairing. Imagine, for example, a mismatched base pair between A and C. Since A is a purine and C is a pyrimidine, the two bases are complementary in terms of size, so you might expect such a base pair to be geometrically accommodated by the double helix. But the failure of the bases to form hydrogen bonds with one another has a major energetic cost. When two strands harboring an A:C mismatch anneal with each other, the hydrogen bonds that had existed between the bases and water when the DNA was single-stranded cannot be replaced by hydrogen bonds between the bases when the DNA is double-stranded (Figure 15B). Moreover, the double helix does not allow enough room for water molecules to fit between the mismatched bases. Thus, there is a net loss of hydrogen bonds during annealing when A is paired with C. If, on the other hand, and as we have seen, when A is paired with its complement T, hydrogen bond formation between A and T entirely compensates for the loss of hydrogen bonding with water molecules. Briefly put, the overall process of forming a mismatched base pair is enthalpically unfavorable because the hydrogen bonds between the bases and water that were broken to allow for pairing are not replaced with equivalent hydrogen bonds once the mismatched base pair is formed. Thus, and whereas hydrogen bonds contribute modestly to the stability of the double helix, they are critically important for favoring complementarity in the annealing of strands and in ensuring specificity during template-directed DNA synthesis (Figure 10), as we will see in Chapter 9.

In sum, multiple entropic and enthalpic factors contribute to the stability of DNA, with the hydrophobic effect and van der Waals interactions between bases being the most important, as summarized in Figure 16.

Cooperativity contributes kinetically to double helix formation

So far we have considered energetic contributions to DNA annealing. One additional factor that acts kinetically also contributes to the formation of the double helix. This is **cooperativity**, which arises from the fact that once two stretches of complementary bases have paired with each other, neighboring bases are likely to find their complements more rapidly because of their close proximity (Figure 17). That is, if two strands of DNA have begun to anneal,

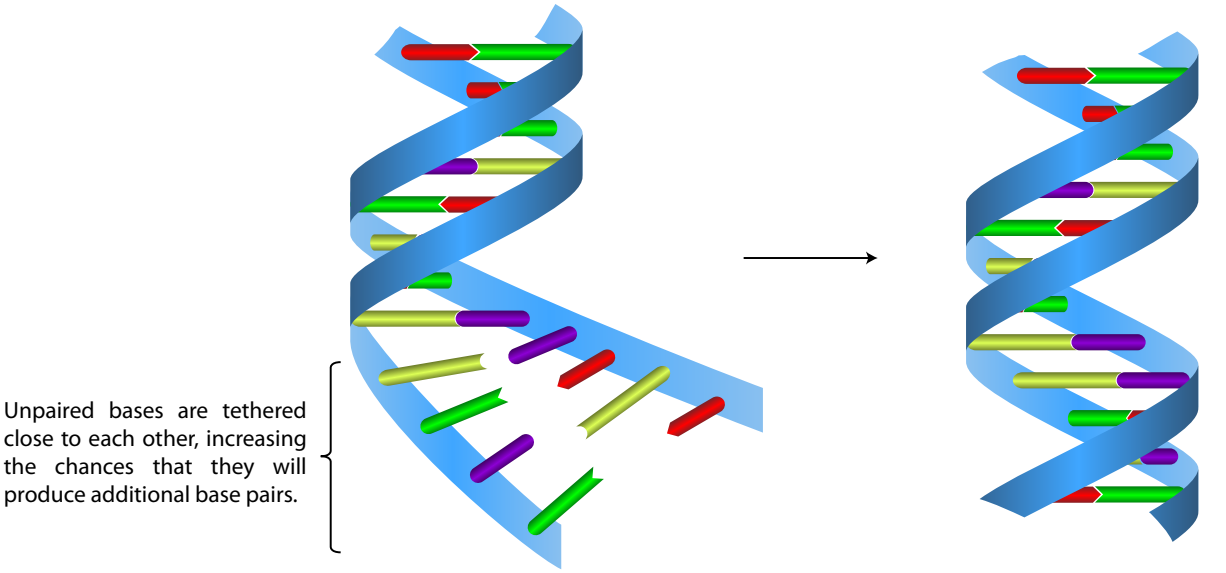


Figure 17 **DNA annealing is a cooperative process**

Double helix formation is cooperative, meaning that the formation of some base pairs increases the probability that additional base pairs will also form. This occurs because the unpaired bases are physically tethered close to one another, increasing their local concentration and increasing the chance that they will find one another in solution.

the adjacent bases on each strand, which are physically tethered to those that have already paired, are brought into close proximity to each other. In effect, there is a **high local concentration** of the adjacent unpaired bases. This high local concentration increases the likelihood that the adjacent complementary bases will pair. Thus, DNA annealing can be thought of as a zippering process that facilities further annealing. Increasing the local concentration is a recurring theme in the chemistry of life and is central to the modes of action of many enzymes.

The information content of the double helix is accessible through the edges of base pairs in the major and minor grooves

Critical to the working of the cell are DNA-binding proteins that must locate and bind to specific sequences of bases in the DNA. For example, regulatory proteins (transcription factors) bind to specific sites in DNA to turn genes ON and OFF, a topic to which we will return in Chapter 12. If the two strands of DNA are wrapped around each other in a double helix with the paired bases on the inside, then how do DNA-binding proteins locate their target sequences? Do they scan the genome by prying the two strands apart to read the bases from the inside? This seems unlikely as it would be energetically costly and slow. Instead, there is a simple solution, and it involves the fact that double-stranded DNA has two grooves that are accessible to proteins from the outside of the helix.

These grooves, which are unequal in width, are called the **major groove** and the **minor groove** (Figure 18). The reason that one groove, the major groove, is wider than the other is that the glycosidic bonds that connect the bases to the sugars project from the bases at a 120° angle from each other, causing the bases to form two unequal-sized grooves when stacked on top of each other (Figure 19). Each groove displays the edges of stacked base pairs. Because the major groove is wider, the edges of base pairs are

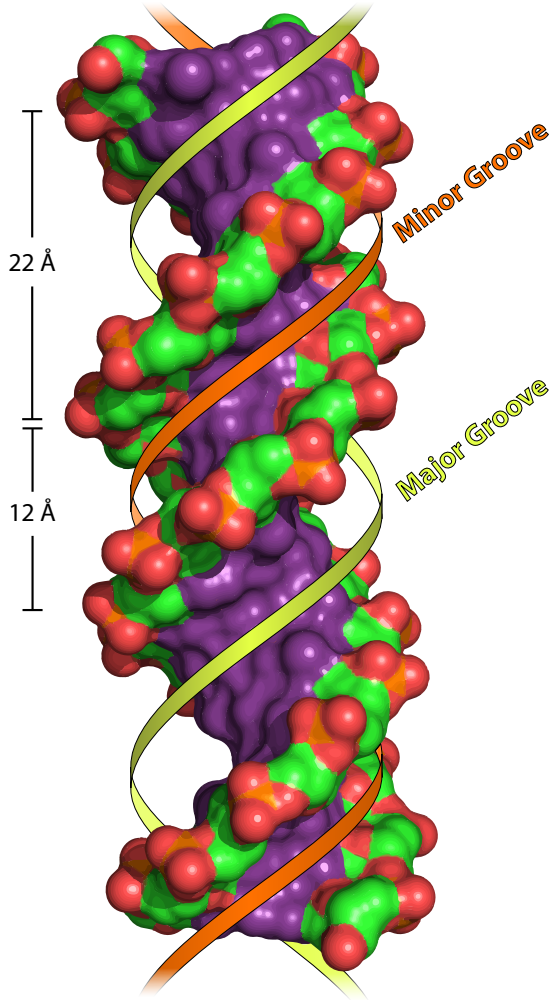


Figure 18 DNA has a major groove and a minor groove

Shown here is a surface representation of DNA in which all atoms in the bases are colored purple. Sugar-phosphate backbone atoms are colored as follows: carbon, green; oxygen, red; phosphorus, orange. The grooves of DNA form two continuous surfaces that wrap around the helix. Each groove is made up of the exposed edges of the base pairs, represented by the purple surfaces in the figure. The major groove (outlined by the yellow ribbon) is wider and more accessible than the minor groove (outlined by the orange ribbon).

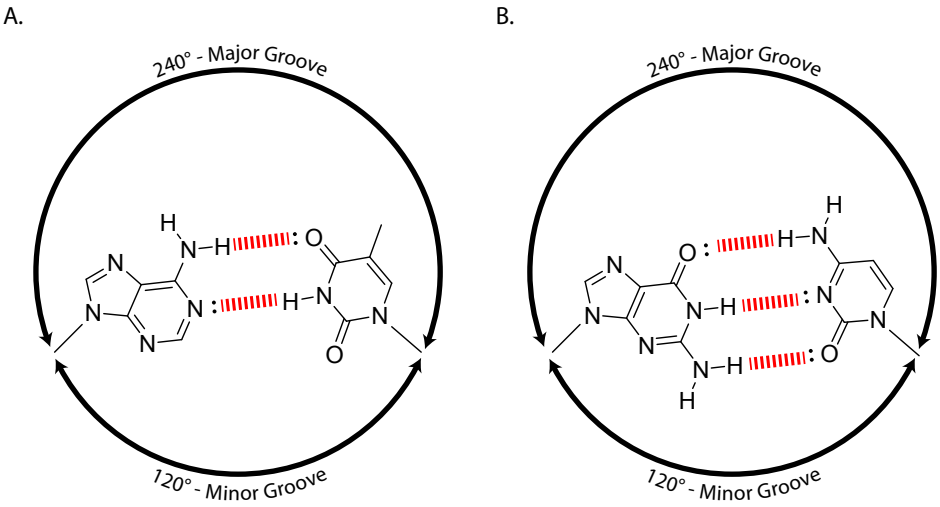


Figure 19 The major and minor grooves result from the angle at which bases connect to the sugars

Base pairs protrude from the sugars of the back bone at an angle of 120° from each other, creating the major and minor grooves. (A) and (B) show A:T and G:C base pairs, respectively.

generally more accessible to proteins in the major groove than in the minor groove, although some proteins contact bases in both grooves.

The significance of the grooves is that the edges of the base pairs are rich in chemical information, especially the major groove. The chemical groups exposed in the major groove are a signature of a specific base pair; they uniquely define the identity of each base pair (Figure 20). For example, a pattern of "H-D-A-A" uniquely identifies a C:G base pair, where "H" is a nonpolar hydrogen atom, "D" is a hydrogen bond donor, and "A" is a

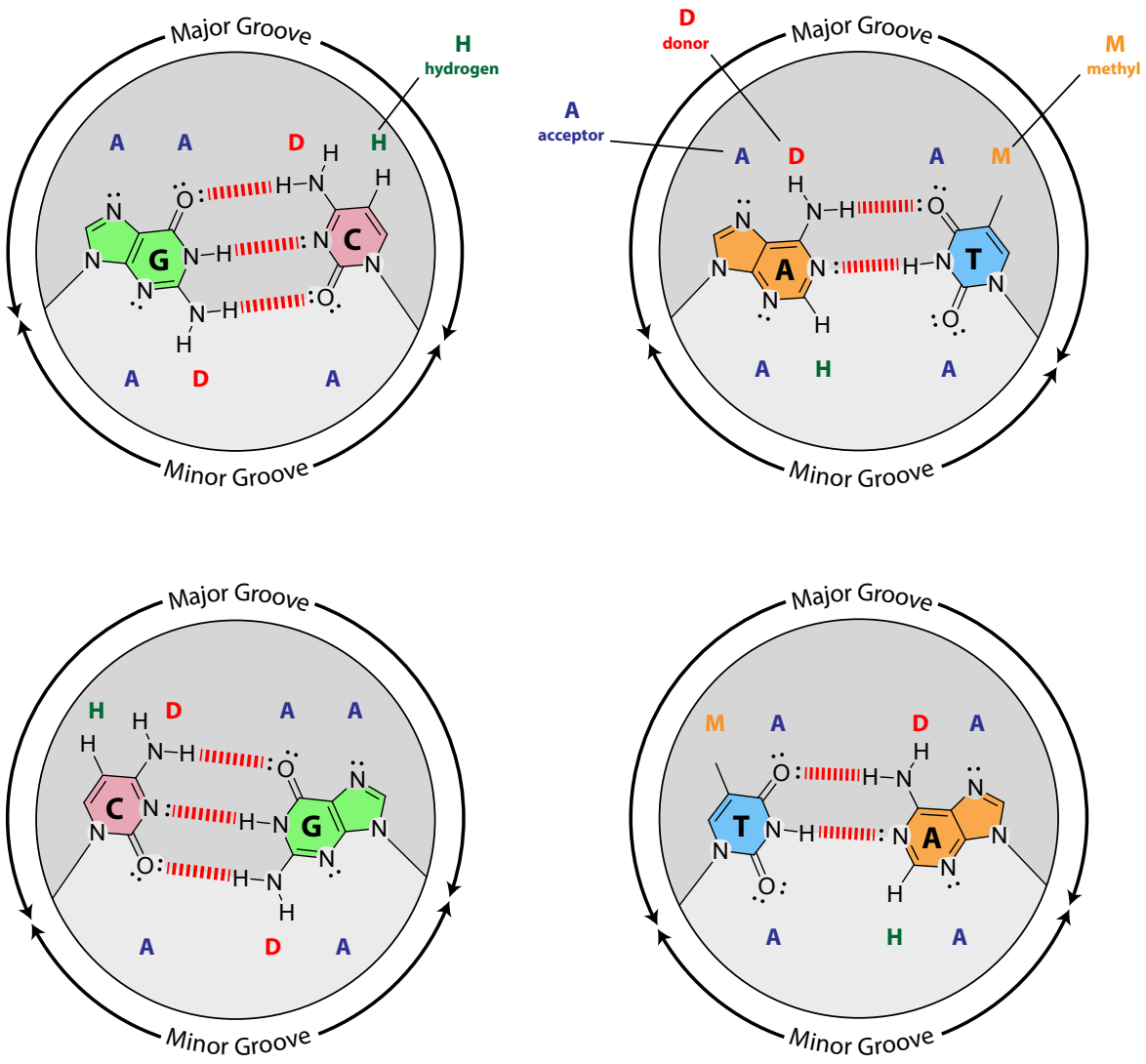


Figure 20 The major and minor grooves display chemical information due to the exposed edges of the base pairs

Each base pair combination is diagrammed to show the pattern of hydrogen bond donors and acceptors that is exposed in the major and minor groove. Atoms found in the major groove are on a dark gray background, while atoms found in the minor groove are on a light gray background.

hydrogen bond acceptor. The reverse pattern, "A-A-D-H," identifies a G:C base pair. Similarly, the patterns "A-D-A-M" and "M-A-D-A" uniquely identify A:T and T:A base pairs, respectively, where "M" is a methyl group ($-CH_3$). Because it is rich in information, and because it is larger and shallower than the minor groove, most DNA-binding proteins bind in the major groove. If the bases in DNA constitute the primary genetic code that specifies amino acids in proteins, then the edges of base pairs represent a second code that enables proteins to read sequence information.

The minor groove, by contrast, contains less chemical information but nonetheless contributes to base pair identification. Whereas C:G and G:C can be distinguished in the major groove, a C:G base pair cannot be distinguished from a G:C base pair in the minor groove, as both display the same "A-D-A" pattern. Similarly, A:T and T:A base pairs cannot be distinguished in the minor groove, once again because the pattern of chemical groups ("A-H-A") is the same in both orientations. In addition,

Box 4 Proteins recognize DNA sequences using chemical features in the major and minor grooves

Figure 21 provides specific examples of how chemical information is read from the grooves. In Figure 21A an arginine side chain is shown forming a pair of hydrogen bonds to a guanine in the major groove, and in Figure 21B a glutamine side chain is shown forming two hydrogen bonds to an adenine in the major groove. Notice that the pattern of hydrogen bond donors and acceptors on adenine in the major groove is different from the pattern found on guanine. Consequently, arginine could not make these interactions with adenine, and glutamine could not make these interactions with guanine. If the DNA sequence were different, and this adenine and guanine were swapped for one another, these hydrogen bonds would not be made, and the interaction between the protein and the DNA would be weakened. It is likely that such a change to the DNA sequence would prevent binding of the protein altogether. As shown in this example, specific DNA sequences are recognized by using a series of amino acid side chains that contact the edges of several bases in the major groove.

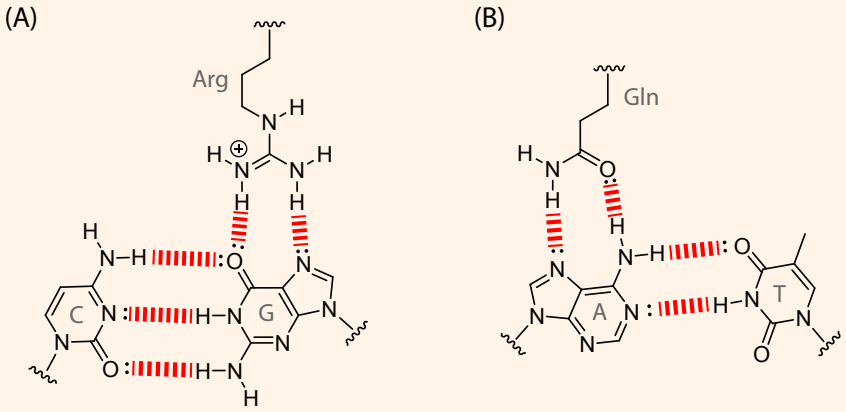


Figure 21 Protein side chains interact with the edges of base pairs in the grooves of the double helix

Shown are the side chains of arginine and glutamine projecting into the major groove, where they have formed hydrogen bonds with the edges of a G:C (A) and an A:T base pair (B).

the minor groove is narrower and deeper than the major groove, which limits its accessibility to amino acid side chains. Although it provides less information than the major groove, some proteins do carry out sequence recognition in the minor groove, a prominent example of which we will encounter in Chapter 10.

DNA is compacted in the nucleus by packaging into nucleosomes

Even though the double helix is narrow in diameter (20 Å) and even though each base pair is only 3.4 Å in thickness, the extraordinarily large numbers of base pairs in genomes, especially those of higher organisms, result in DNA molecules that are remarkably long. Consider all the DNA in the nucleus of a human cell, which is diploid and contains two copies of 23 chromosomes. The entire content of DNA contained in these chromosomes is 6 billion (6×10^9) base pairs. What is the length of 6×10^9 base pairs of DNA? A simple calculation shows that 6×10^9 multiplied by 3.4 Å (3.4×10^{-10} m) per base pair equals about 2 meters. This presents a major challenge for cells, as they have to accommodate all of this DNA in a nucleus that is only about 5 μm (5×10^{-6} m) in diameter.

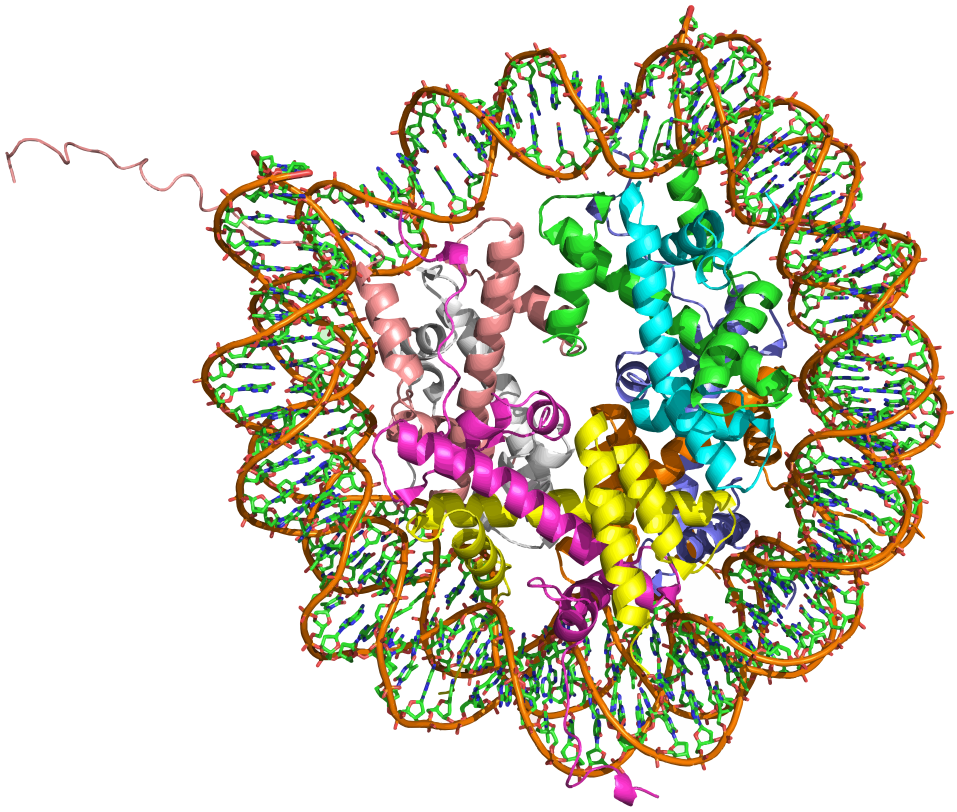


Figure 22 DNA in cells is packaged into nucleosomes

Shown is the X-ray crystal structure of a nucleosome, consisting of DNA wrapped around an octamer of histone proteins, each of which is labeled with a different color.

How do the cells of higher organisms accomplish this feat? They do so by compacting the DNA, and fundamental to this compaction is wrapping the DNA around specialized proteins called **histones** to create structures called **nucleosomes**. Each nucleosome is a heteromeric complex of pairs of four kinds of histone proteins (histones H2A, H2B, H3 and H4) around which the DNA is coiled almost two times (Figure 22). The wrapping of DNA around histones is energetically favorable due to ionic interactions between the negatively charged phosphates of the DNA backbone and positively charged side chains of amino acids (e.g., lysine and arginine) displayed on the surface of the histones. Because of their fundamental role in packaging the vast amounts of DNA in higher cells, histones are among the most abundant proteins in eukaryotic cells. We will return to the topic of nucleosomes and how DNA is packaged in the nucleus when we consider how genes are turned ON and OFF in eukaryotic cells (Chapter 12).

Summary

DNA consists of two polynucleotide chains wrapped around each other as helices. Each chain is an alternating copolymer of phosphates and the pentose sugar 2'-deoxyribose. The carbon at the 2' position of 2'-deoxyribose lacks a hydroxyl group. (RNA, in contrast, contains ribose sugars, which have a hydroxyl group at the 2' position.) The sugar units in the polynucleotide chain are linked to each other by phosphate groups that are joined by ester linkages at the 5' and 3' positions. As a consequence, polynucleotide chains are directional in that they have a 5' terminus and a 3' terminus.

The third component of DNA, the four nucleobases or bases, are joined to the sugar at the 1' position. Thymine and cytosine are six-membered

heterocyclic rings (pyrimidines), whereas adenine and guanine consist of fused five- and six-membered heterocyclic rings (purines). Each repeating unit of a phosphate, a sugar, and a base is referred to as a nucleotide (hence the name polynucleotide).

The two polynucleotide strands of DNA are wrapped around each other in an anti-parallel, 5'-to-3' orientation with a periodicity of ~10 base pairs per turn of the helix. Complementary bases pair with each other between the strands, with A pairing with T and G pairing with C. The specificity of pairing is governed by hydrogen bonds (two for A:T and three for G:C) and geometry, such that both base pairs are accommodated in the double helix with the same diameter (20 Å). Thus, DNA has a regular structure that accommodates irregularity in the form of the sequence of base pairs.

Formation of a double helix from two complementary polynucleotide chains is energetically favorable despite the repulsive contribution of the negatively charged phosphates and the decrease in entropy. Contributing to the stability of the double helix are the hydrophobic effect from the exclusion of water molecules from the hydrophobic surfaces of the bases and van der Waals interactions between the stacked bases. Hydrogen bonding between the bases contributes modestly to the stability of the helix. Although the formation of double-stranded DNA replaces hydrogen bonds that the bases in single polynucleotide chains form with water molecules, the inter-strand hydrogen bonds in the duplex are geometrically optimized. Hydrogen bonding plays a critical role in the specificity of base pairing. Mismatched bases are unable to replace hydrogen bonds from the displaced water molecules, hence incurring an energetic cost.

Base pairs are sequestered inside the double helix. Nonetheless, the sequence of base pairs in the DNA is accessible from the edges of base pairs, which are displayed in the major and minor grooves of the helix, enabling DNA-binding proteins to recognize specific sequences. The difference in width between the two grooves arises from the 120° angle that joins the bases to the sugars. The major groove is wide, shallow, and rich in chemical information, whereas the minor groove is deeper, narrower, and less rich in chemical information. All four base pairs (C:G, G:C, A:T, and T:A) can be distinguished from the chemical groups in the major groove, whereas C:G cannot be distinguished from G:C nor A:T from T:A in the minor groove.

Finally, DNA in the nuclei of higher cells is compacted into nucleosomes by wrapping around a complex of eight histone proteins.