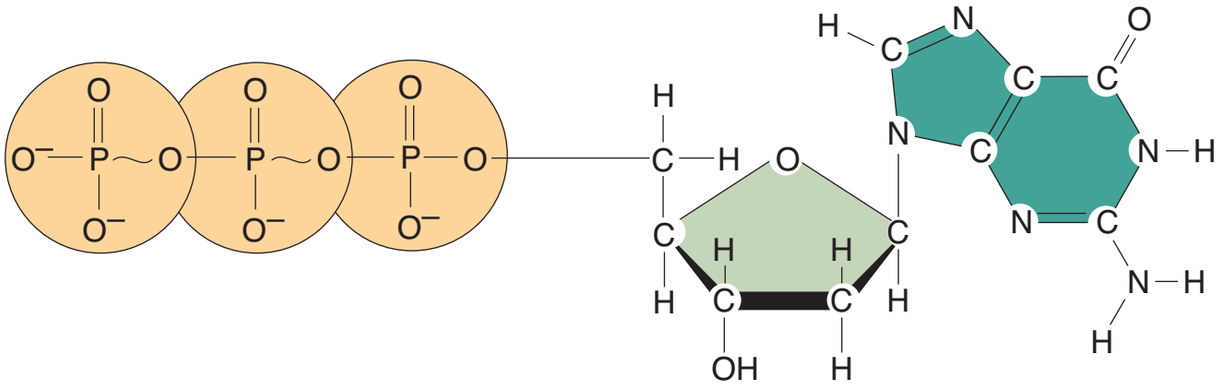


CHAPTER 2

DNA Is the Instruction Book for Life

Nucleic Acid Structure and Organization

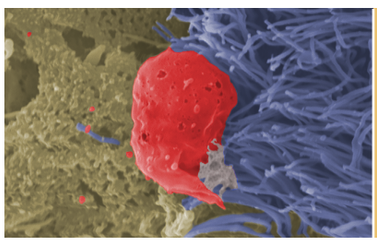
CHAPTER OUTLINE

2.1 The Big Picture

One of the most important trends discussed in Chapter 1 is that, as organisms evolved into increasingly complex beings, the amount of genetic material (deoxyribonucleic acid, or DNA) they contained increased as well. This is not coincidental: the biological codes responsible for developing and sustaining this complexity rely on the molecules encoded by sequences of nucleotides in DNA to serve as the actors that obey the rules in biological codes. The more genetic material a cell has, the more information it can access, and the more variation it can generate. Accessing this information requires cells to follow one of the most ancient biological codes, the genetic code.

DNA is one of the most critical molecules in living organisms because it contains the essential information passed down from generation to generation.

The sequence of nucleotides in DNA reflects its evolutionary history, similar to written documents that can trace our genealogical history to our ancestors. (Indeed, genealogists use DNA sequences to trace family histories for exactly this reason.) Put more simply, sequences of nucleotides in DNA are the functional equivalent of words in a book, and every time a cell divides, it must replicate its entire library of genetic sequence information. Changing even a single letter in one word can change the meaning of the word, as well the sentence and book that contains it. Nothing in nature is perfect: cells make errors when replicating their DNA and pass these errors on to their descendants. These changes constitute the genetic variation required for evolution to occur: variation in the words changes how biological codes are enacted and thereby changes the way organisms function. By studying when uncorrected errors in this sequence appeared over evolutionary history, biologists can reconstruct the time when the information in the library changed and thereby trace the history of all organisms back to LUCA (the last universal common ancestor). This concept forms the foundation of **cell biology principle 2: DNA is the instruction book for life**.



CELL BIOLOGY PRINCIPLE 2

DNA is the instruction book for life.

Photo courtesy of Andrew S. Mount, PhD; Jonathan Stewart; Nichole Hickman; Michael Groce; Okeanos Research Lab, Clemson University.

The implications of this principle are profound. Virtually every cellular function can be traced back to biological codes acting on ribonucleic acids (RNAs) and proteins encoded by DNA. Unpacking how this happens remains one of the primary challenges in cell and molecular biology, and the lessons we learn along the way directly impact nearly all aspects of modern life, including medicine, agriculture, manufacturing, and even warfare.

The goal of this chapter is to explore the structure of nucleotides and nucleic acids, as well as their structural organization inside eukaryotic cells. This will serve as the foundation for addressing how DNA is replicated (Chapter 7) and how cells access DNA information (Chapters 8 and 12). We will frequently reference sections of these chapters to reinforce the concept that a carefully organized "library" is a prerequisite for cells to respond properly to challenges in their environment.

This chapter covers four major ideas:

- First, DNA is a storehouse of "cellular information." Like code biology, the concept of biological information is also somewhat abstract; notice how a relatively simple linear code can produce an enormous number of different RNA and protein products.
- Second, the linear structure of a DNA molecule is relatively simple. It can often be learned largely by memorization.
- Third, DNA organization in a eukaryotic cell is relatively complex. To assist in working through this subject, we have divided the physical organization into five levels, from poorly organized to very highly organized.
- Fourth, cells can chemically modify DNA and the proteins that organize it to help control their structure and function. Understanding this idea sets the stage for explaining how cells read specific sections of a DNA molecule at specific times.

2.2 All of the Information Necessary for Cells to Respond to Their External Environment Is Stored as DNA

Key Concepts

- Cells store the information necessary to build RNA and proteins in the form of DNA, a simple nucleic acid found in all living organisms.
- When a cell divides, each of the resulting daughter cells contains an almost exact copy of the parental cell DNA.
- DNA must transmit its information to other molecules to be useful.
- The functional unit of DNA information is called a gene.
- The DNA information in genes is converted into matching, single strands of RNAs by a process called transcription; cells transcribe many different types of RNA from their DNA.
- The sequence information in messenger RNAs (mRNA) is converted into a sequence of amino acids by a process called translation.
- Mutations in DNA are passed on to RNAs and proteins, resulting in variations that are acted upon by natural selection.

As we stated in Chapter 1, to be alive, cells must have the ability to self-replicate and self-correct. This statement implies that cells can sense their surroundings, so they know when it is safe to replicate, and that they are aware of their functional state and know when they are damaged. To "know" these things, cells must contain information (see Box 2-1).

BOX 2-1 TIP

Here is a helpful tip for every chapter in this book: before reading the chapter in detail, flip through the pages and note how many figures are devoted to each of the major subdivisions. Remember that it takes considerably more effort to generate a figure than it does to write an equivalent volume of text, so if an idea warrants a figure, it likely merits your attention. Looking at the figures before reading reveals a visual roadmap of some of the most important concepts in the chapter. By noting how many figures are used for each subdivision and what they contain, students will have a good idea of what to expect as they work their way through a chapter.

What is cellular information? The concept of *information* is fairly abstract. While we are all quite familiar with its use in our daily lives, we cannot point to it. Most often, when we refer to information, we mean the way it is represented in physical form: as written text, recorded sound, or images, for example. One of the best physical examples of information storage in our everyday lives is a library (see Box 2-2). Cells don't use any of these forms to store information, but they nonetheless capture and store a tremendous amount of information about their surroundings and their physical state. This information is stored in the form of molecules. Just as we use different physical forms to store different

BOX 2-2 THE LIBRARY OF CONGRESS ANALOGY, PART 1

The information stored in DNA is analogous to the letters, numbers, and other symbols used in books. Units of these books, such as chapters, are the genes in DNA. For an information storage system to be useful, someone has to open and read it; in a library, this is done by people, and in cells, this is done by proteins. A nucleus is similar to the Library of Congress in that entry is restricted and the removal of information is prohibited. So, proteins in the nucleus "copy" the genes in the DNA "books" by creating RNA copies; this is called transcription because the language of a DNA "book" is transcribed to a new form. **FIGURE A** shows how these concepts apply to both cells and our hypothetical library. While the source materials (books or genes) must remain within the structure (library or nucleus), several copies can be made rather quickly and distributed to numerous locations outside the library building. Note that in a cell's "library," many of the books resemble "instruction manuals" more than novels or textbooks. Thus, photocopies of these manuals are useful only if they perform a job. Most RNA molecules control cell behavior directly, analogous to reproducing a work of art and putting it on display. A notable exception is messenger RNA (mRNA), which requires conversion of its information into sequences of amino acids we call proteins. The reading of mRNA photocopies to build polypeptides and proteins is called translation for this reason: words are converted into physical objects (proteins) that do the actual work. The reading of mRNA is done by ribosomes, which can be thought of as factories (several proteins and RNA molecules working together as a team). Note that this strategy protects the books quite well, while also making them useful to the population.

The recent effort to digitize library holdings by scanning and photographing predigital-age works illustrates how humans have put the lessons of cell and molecular biology to use. Instead of storing information on paper, canvas, clay, stone, parchment, vinyl, and magnetic tape, all of this information is being converted into a simple binary computer code that can be readily reproduced and edited. Whereas binary code uses only two elements (zeros and ones) to store information, DNA uses four building blocks, called nucleotides. Digital libraries more closely resemble how cells manage information than conventional libraries, but even they must confront physical storage demands: server farms are highly specialized, organized spaces dedicated to storing and securing the hardware holding the encoded information. DNA has even been proposed as a digital information storage device to replace binary code (Panda et al., 2018).

This analogy also demonstrates how mutations can affect cells. Imagine that someone (or even entire teams of people) tried to reproduce every letter of every book in the library. Chances are good that he or she would make several mistakes (even digital copies can contain errors), and the new books would have different words in them. If these mistakes are not corrected, all subsequent photocopies of these books would have different information, and many of the new words would be unrecognizable; this is why many mutations are harmful to cells. On very rare occasions, the mistakes yield changes that are useful, such that the RNA (and possibly proteins) resulting from these new books helps the cell better adapt to its environment. These cells have a slightly higher probability of surviving long enough to divide and pass on this new information to the next generation of cells. This is how evolution by natural selection works.

FIGURE A The Library of Congress analogy for information storage and transmission in cells.

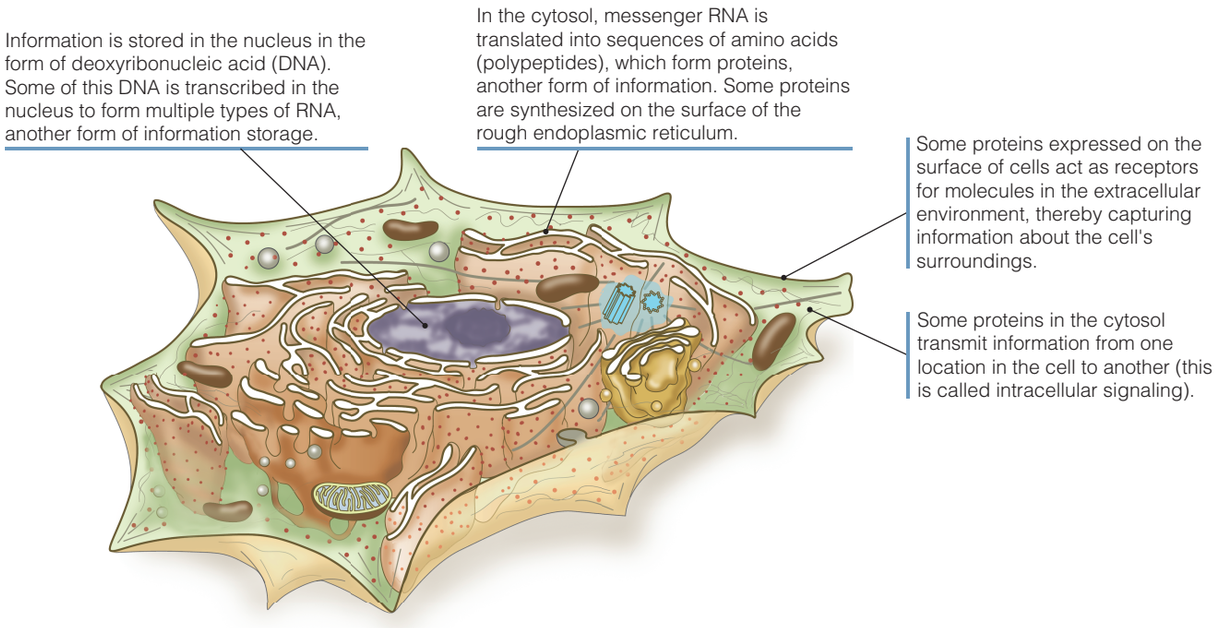


FIGURE 2-1 Some forms of information storage in cells.

kinds of data, cells use different types of molecules to store the different types of information they need to stay alive, as illustrated in **FIGURE 2-1**.

One of the most important molecules used to store cellular information is DNA. Cells use DNA to store information for constructing other complex biological molecules, such as RNAs and proteins. (How cells use DNA to construct these other molecules is discussed in Chapter 8.) DNA stores the most fundamental information necessary for life; every cell, and, therefore, every living organism, uses DNA for this purpose.

A Cell's DNA Is Inherited

DNA stores information in the form of a linear sequence of repeating units, somewhat analogous to how a linear sequence of letters or characters in an alphabet can store information in the form of words, sentences, or paragraphs. DNA uses a very simple language, composed of only four molecules, commonly known as A, C, T, and G. As we will see below, each of these letters is a molecule known as a **deoxyribonucleotide**. These deoxyribonucleotides are attached end-to-end to form very large structures in cells.

When cells replicate, they divide into two parts called daughter cells, each of which inherits a complete and nearly identical copy of the parental cell's DNA. This requires the parental cell to replicate its DNA before cell division. (The mechanisms of DNA replication are discussed in Chapter 7.) In multicellular organisms such as humans, most cells (called **somatic cells**) only pass their DNA on to the cells that replace them during that organism's lifetime; a specialized set of cells, often called **germ cells** (e.g., eggs and sperm), is usually responsible for passing DNA from one organism to its offspring.

Mutations in DNA Are Passed from Generation to Generation

When cells replicate their DNA, they frequently make mistakes, as seen in **FIGURE 2-2**. Some of these mistakes result in changes to the deoxyribonucleotide

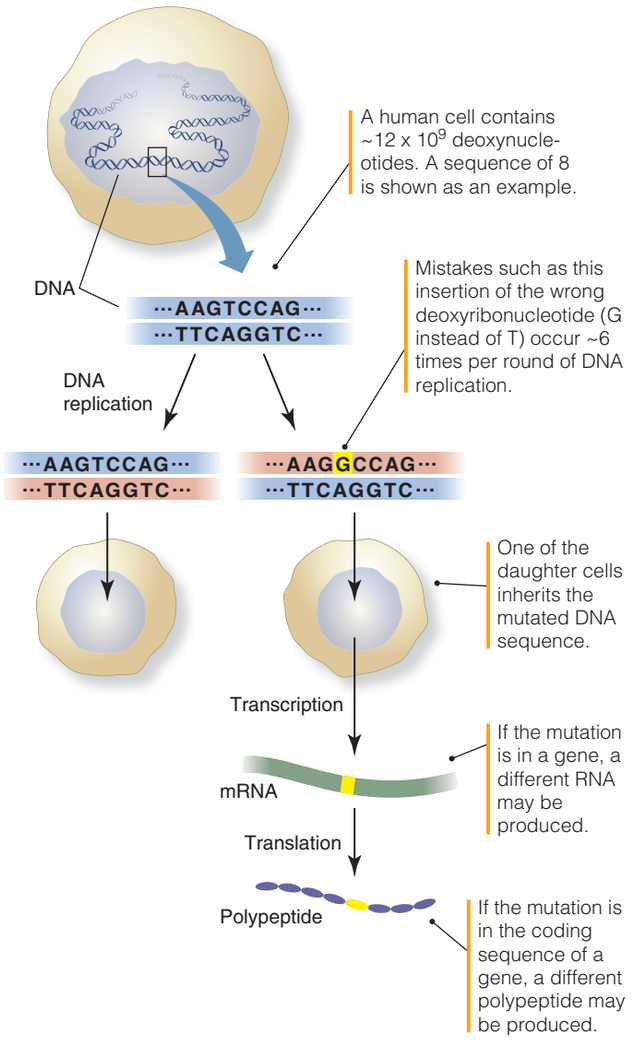


FIGURE 2-2 Mistakes in DNA replication may cause mutations.

sequence. This is understandable, considering how many DNA replication events take place during an organism's lifetime. For example, let's take a look at the human body. Each human somatic cell (i.e., not eggs or sperm) contains approximately 12 billion (12×10^9) deoxyribonucleotides in its DNA. The number of cells in an average human body at any given time is estimated to be about 3×10^{13}, which replicate to form approximately 10^{16} cells during a person's lifetime. Generating 10^{16} cells from a single fertilized egg, therefore, requires $(12 \times 10^9) \times 10^{16} = 12 \times 10^{25}$ nucleotides to be replicated *in the correct order.* Nothing in nature can achieve the level of accuracy necessary to replicate all of these nucleotides perfectly. In fact, DNA replication errors are actually quite common in humans, especially in rapidly dividing cells, such as blood cells or cells lining the digestive tract: the rate of inserting the wrong deoxyribonucleotide in a DNA sequence (often called a **point mutation**) is approximately 1 per every 2×10^{10} deoxyribonucleotides, or about six errors each time a cell replicates.

Other changes in the original DNA template sequence can also occur: extra deoxyribonucleotides can be inserted, some can be left out; even large pieces of DNA can be accidentally deleted, added, and/or moved to another

location in the DNA sequence. The net effect of these changes is that every cell, and every organism, is at least slightly different from its ancestors, siblings, and other relatives. This heterogeneity in DNA sequences contributes a great deal to the variation found in populations of organisms that undergo evolution by natural selection.

DNA Must Be Read to Be Useful

Like other forms of coded information, DNA is not useful in isolation. For example, the information in a book is meaningful only if it is read and put to use. The same principle applies to DNA in a cell: only the portions of DNA that are "read" are meaningful. Cells don't literally read a DNA sequence, of course. Instead, they use proteins that bind to specific deoxyribonucleotide sequences in the DNA, and this binding changes the behavior of the proteins, as illustrated in **FIGURE 2-3**. The information is then converted into a useful form (an RNA molecule). Some of these proteins are responsible for *transcribing* deoxyribonucleotide sequences in DNA into RNA sequences; some of these RNA sequences are then *translated* into amino acid sequences in proteins, a topic we will cover in much greater detail in Chapter 8. We will see numerous other examples of how DNA information is used throughout the book.

For many years, scientists thought that a large percentage of the DNA in most eukaryotic cells was useless because they could find no evidence that proteins would bind to these regions. More recently, they discovered that these regions contain characteristic patterns of repeating DNA sequences. We now know that many of these sequences either bind to proteins directly or control the shape of neighboring DNA sequences that bind proteins. (These noncoding regions are analogous to the "behind-the-scenes" roles that individuals play in a theater production: without them, the performance could never take place.) Much of this DNA contains deoxyribonucleotides that are chemically altered (e.g., by methylation: see Section 2.4). These chemical modifications can have a profound impact on which portions of DNA are read by a cell. These modifications are so important that a new field of biology, called **epigenetics**,

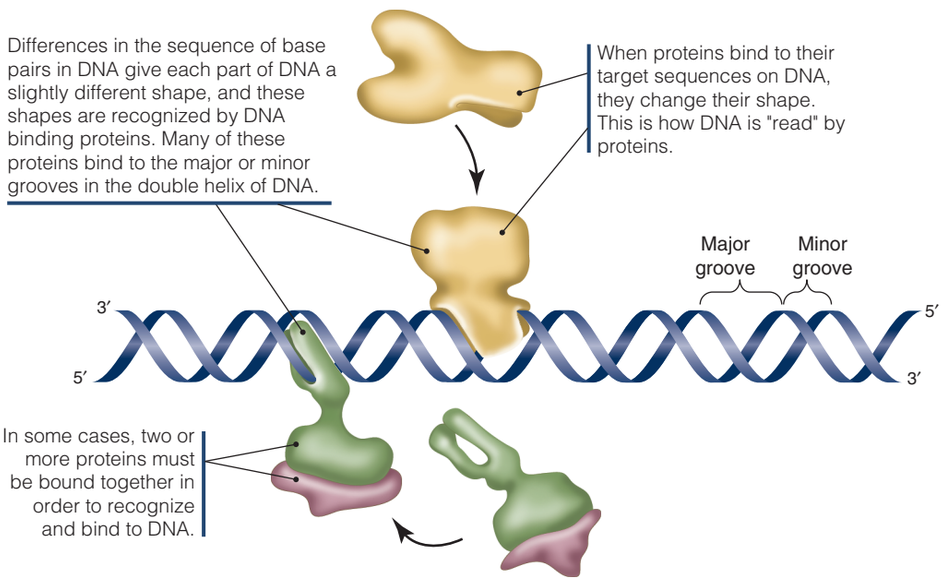


FIGURE 2-3 DNA information is "read" by proteins.

has emerged for studying how these modified sequences impact the phenotypes and behaviors of organisms. We will present more specific examples of epigenetic modifications later in this chapter (see Figure 2-23) and in Chapter 12.

DNA Information Is Packaged into Units Called Genes

The most familiar form of DNA information is a segment of deoxyribonucleotides known as a **gene** (**FIGURE 2-4**). Despite its common usage in many fields of biology, there is no universally accepted definition for this term. In this book, we will define a gene as the linear sequence of deoxyribonucleotides necessary for converting a portion of that sequence into a complementary sequence of ribonucleotides inside a cell. (In this context, two nucleic acid sequences are called complementary when they can form stable hydrogen bonds between their base pairs.) Less formally, we can say that a gene is a portion of DNA that can be converted into RNA, plus some additional sequences that are absolutely necessary for this conversion to take place. A gene is *always* a single linear sequence of deoxyribonucleotides on a single piece of DNA; a gene cannot be fragmented into different portions of DNA scattered throughout different DNA molecules. The average length of a human gene is 10,000–15,000 nucleotides (often abbreviated as **base pairs**, or bp), though there is considerable variation in size.

Genes are the best-known units of biological inheritance. When Gregor Mendel (1822–1884) first discovered the principles of genetic inheritance in the mid-nineteenth century, he was studying how individual genes of pea plants in his garden were passed from generation to generation. A few decades later, Heinrich Wilhelm Gottfried Waldeyer-Hartz (1836–1921) discovered linear strands in the nucleus that changed color when he added stain to the cells, and he called these structures chromosomes (derived from the Greek word meaning "colored bodies"). Another 20 years passed before Thomas Hunt Morgan (1866–1945) determined that genes are arranged on chromosomes.

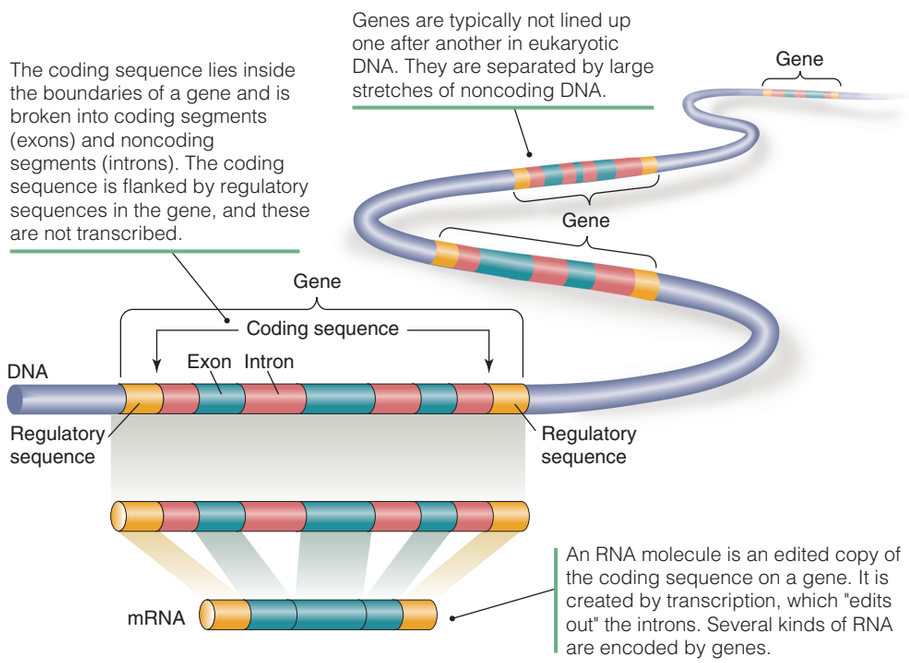


FIGURE 2-4 The smallest functional unit of DNA is a gene. A sample of eukaryotic DNA encoding a messenger RNA is illustrated.

A single chromosome may contain thousands of genes lined up one after another, each with its distinct packet of information (see Figure 2-4). The modern definition of a **chromosome** is a genetic element containing genes essential to cell function. **Genetics**, the field of study devoted to uncovering the mechanisms governing the inheritance and expression of genes, has contributed greatly to our understanding of cell behavior. We will discuss the mechanisms controlling gene expression in greater detail in Chapter 12.

Genes Are Transcribed into RNAs

All genes share one important trait: some portion of their deoxyribonucleotide sequence can be converted by a cell into a complementary sequence of ribonucleotides (also known as RNA). In many genes, the portion that is replicated as RNA includes a region called the **coding sequence**. The coding sequence of many eukaryotic genes is often broken up into segments, called **exons**, separated by noncoding sequences called **introns**. The process of synthesizing an RNA molecule is called transcription, illustrated in Figure 2-4. Currently, biologists recognize seven major classes of genes, according to the function of the transcribed RNAs they generate (see Box 2-3):

BOX 2-3 THE LIBRARY OF CONGRESS ANALOGY, PART 2

The library analogy works best for mRNAs, as mRNAs are the photocopies cells read to make proteins. However, the analogy doesn't work too well for the rest of the RNAs, and this is because they are perfectly useful to cells *without* having to be translated into proteins. Most photocopies are pretty useless by themselves: they can be made into paper airplanes but not real planes. So think of these other RNAs as very special photocopies that function perfectly well without having to undergo translation.

- **Ribosomal RNA (or rRNA)** molecules are an essential component of the large and small subunits of ribosomes, the molecular complexes that make proteins. In humans, the small subunit rRNA is about 1,900 nucleotides long (sometimes referred to as simply bases), and the large subunit rRNA is about 5,000 nucleotides long. Biologists have found over 1,200 introns at over 150 unique sites in the small and large subunit ribosomal RNA genes across all known species of organisms. These intron sequences must be clipped out before the mature rRNAs can function properly.
- **Messenger RNA (or mRNA)** molecules, unlike rRNAs, do not play an *active* role in any cellular activity. Instead, they serve as templates for the assembly of the ribosomes that will build a specific new polypeptide. These are called messenger RNAs (mRNAs) because they are intermediates in the genetic code, carrying the encoded message from DNA to proteins. Put another way, mRNAs are short-lived, intermediate copies of DNA information that are translated into proteins. Human mRNAs average 2,500 nucleotides in length and contain an average of 7.8 introns.
- **Transfer RNA (or tRNA)** molecules are bridge molecules that link amino acids to a specific three-nucleotide sequence on mRNA. They specifically deliver the correct amino acids to ribosomes, where they are added to the polypeptides being synthesized. (This is analogous to mail carriers and couriers, who ensure packages are delivered to the correct address.) Compared to the rRNAs and mRNAs they interact with, tRNAs are comparatively tiny (typically about 73–93 nucleotides long) and contain a small number of introns that can be traced back to the archaeal ancestor

of eukaryotes (see the subsection *Eukaryotes Arose Through Teamwork by Prokaryotes* in Chapter 1).

- **MicroRNA (miRNA)** molecules are noncoding RNAs with an average length of 22 nucleotides. The genes encoding miRNAs do not contain introns, but some are contained within the introns of genes encoding mRNAs. They perform many different functions in cells, but most inhibit translation by binding to a noncoding region of mRNA molecules and inhibiting the ribosome. Their discovery in 1993 revolutionized cell and molecular biology by introducing a fourth critical player (in addition to rRNA, mRNA, and tRNA) in the biological code that converts DNA into proteins. In addition to inhibiting translation of their own genes, cells can also use miRNAs to defend against infection by viruses. These molecules are covered in much greater detail in Chapter 11.
- **Short interfering RNA (siRNA)** molecules are so-named because they are small (21–23 base pairs in length) and interfere with translation by inducing the destruction of targeted mRNAs. Biologists think these molecules evolved in plants and animals to silence genes of the viruses and other pathogens that infect them. Mammals, including humans, do not express siRNAs, but they inherited the ability to use foreign siRNAs from their evolutionary ancestors. This discovery in 2001 opened up a new and potentially powerful way of combating genetic diseases, such as cancer, by inserting synthetic siRNA molecules that target and inhibit cancer-causing genes in tumors.
- **Small nuclear RNAs (snRNAs)** are slightly larger than miRNAs (100–200 nucleotides long) and function primarily in the nucleus, where they control transcription of mRNA genes and export of RNA and proteins from the nucleus to the cytosol. These are discussed in greater detail in Chapter 11.
- The seventh and most diverse class of RNA molecules plays a multitude of regulatory roles, including splicing of introns from other RNAs and editing of gene sequences. X-inactive-specific transcript (Xist) RNA is a noncoding RNA that plays an essential role in inactivating X chromosomes, as we will discuss in Section 2.3. Another notable member of this class is called clustered regularly interspaced short palindromic repeats (CRISPR) RNA. We discuss CRISPR in much greater detail in Chapter 12.

Messenger RNAs Are Translated into Proteins

The combined product of rRNAs, tRNAs, mRNAs, and miRNAs working together is a newly synthesized chain of amino acids called a polypeptide, as shown in **FIGURE 2-5**. miRNAs help determine when a specific mRNA is available for translation. During translation, three-nucleotide-long **codons** in the coding sequence of mRNA are matched with **anticodons** in tRNA by ribosomes (including rRNA) to determine the sequence of amino acids in the resulting polypeptide. Even though 64 different codons are possible (four different nucleotides can fill each position, therefore 4^3 or $4 \times 4 \times 4 = 64$), each codon does not code for a unique amino acid. Instead, 3 codons (UAG, UAA, UGA) are designated as "stop" codons, which halt translation, and in the remaining 61 codons, redundancies ensure that an mRNA specifies only 20 different amino acids. The average size of a human gene coding sequence is about 500–600 codons, yielding a polypeptide of 500–600 amino acids in length. Therefore, an average-sized human gene can theoretically encode at least 20^{500} different polypeptide sequences (20 different amino acids per each of the 500 codons; see **TABLE 2-1**). In reality, the actual number of polypeptides produced by any single organism is far less than this because most of these polypeptides would not be useful to cells. Polypeptides (either as

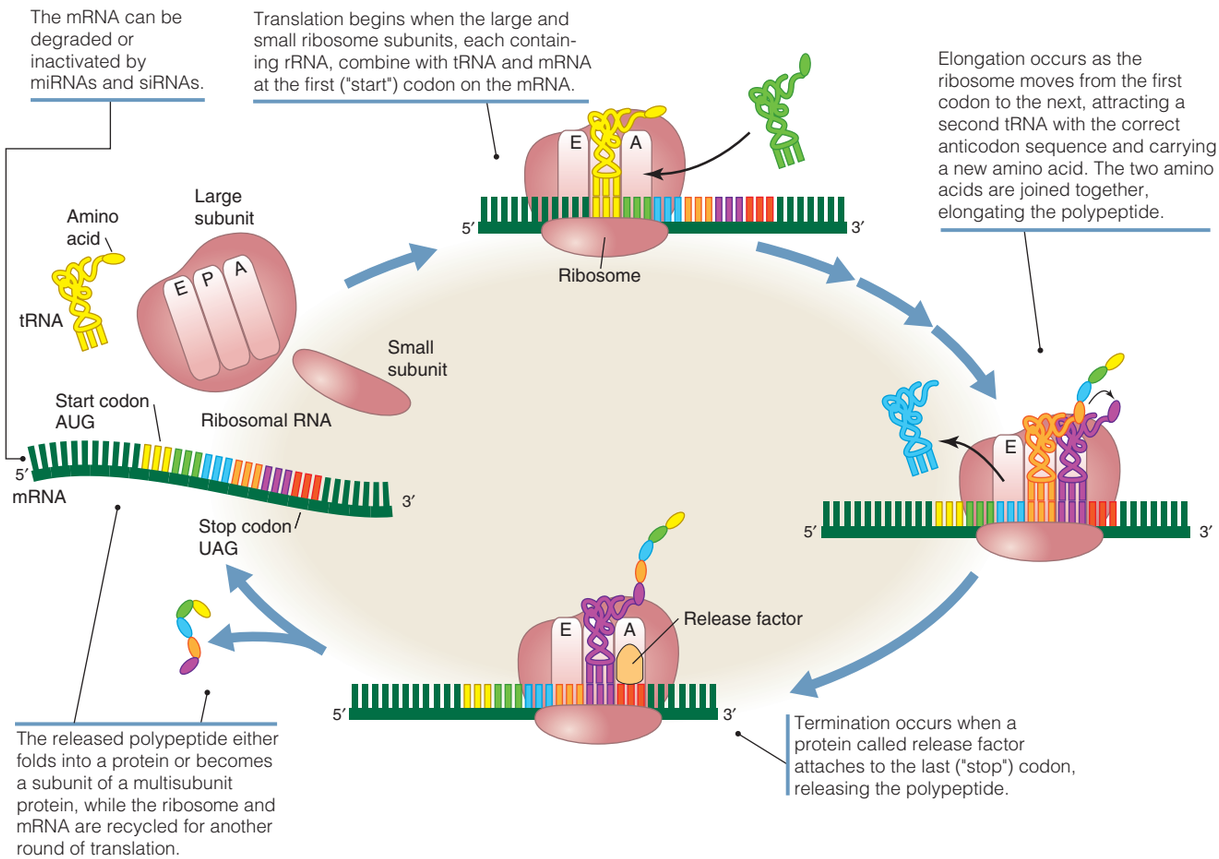


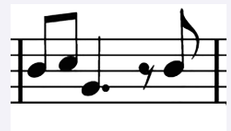
FIGURE 2-5 An overview of translation in eukaryotes.

individuals or as assembled groups) form proteins, some of the most common molecular structures in cells, and proteins have some strict requirements for how their polypeptides must behave. We call these requirements the three traits of proteins, and we will discuss them in detail in Chapter 3.

The information stored in DNA can undergo two transformations to become useful to cells. The first transformation, **transcription**, simply changes the DNA informational sequence into a complementary RNA informational sequence. This has a profound benefit for cells: relative to the typically very large DNA molecules, RNA molecules are quite small and can be transported to different regions in a cell relatively easily. Also, multiple RNA molecules can be generated from the same DNA sequence, so that a cell can configure its RNA profile by simply changing the number of RNA molecules it copies from each of its genes. This is one easy way for cells to become specialized.

The second transformation occurs when mRNA information is used to build polypeptides. The process of converting the ribonucleotide sequence of mRNA into an amino acid sequence in a polypeptide is called **translation**. We will discuss the molecular events responsible for translation in much greater detail in Chapter 8. Note that mRNA is the only known type of RNA that is translated. This, too, offers great benefits to cells. Each DNA and RNA molecule is composed of only four different subunits (nucleotides), so the number of variations available is relatively small. By comparison, proteins are composed of twenty different subunits (amino acids), which permit them to form much greater numbers of different molecules. Translation allows the information in DNA to be expressed in a myriad different proteins.

TABLE 2-1 Comparison of the Information Storage Capability of the Genetic Code, ASCII Computer Code, and Musical Notation

	Genetic Code	ASCII Code	Musical Notation*
	GATCCCTGCC TGGGCTTTGC CTCTGCAGCC CCCCGCCCCA CAGGTTCACA CCTCGGGTCT TCTCCACCGC TGCCACACGC CAGAGCCTGT AGCGGGGCCT CAGAGTCTGG GAGGTGGGAC TCCTGCACCT CAGCCATCAT CAGACCCATG GGGCCACCCA GGGAACCTTG GCAGGGACCA TTACCAGTGA CCTGCCGAGG CCCCGGACTC TGCCAGCCAG CTGTGCCGGC CACCCTGCCC CGGACAGTGC CGGTTCATGT GGGAACTAGG GGACGATGTG GTTCTTCGCA TCTGATGATG AAGGCCCTGG GCCACTTGGC ACGGGCGGGC GCTCCCGAGA TGGATATGAG GAGCCCCCTC TGCCCAACTC CCAGAAAGGC CGAGGCTCTG CAGCGGGAGG AAGTCCTGCG ATGTCCTGGG GGGCAGCAGC GCAGGGCACA GGGACAGCCC CCCTCCACAG CTCTTCCTGG CCAGCCCTCC CCACTATCTG CCAGGAGGTT GCTTCTTCCA GGAGGCTTTT CCCGACCAGC CCAGGGGTCC AGGGTCTGGG GCTCCCAGCT GCTGTGAGTG CTGCACATTC TCTTGAGGAC AGCCCCCTCC CTCCCCCACC CACTTCTGGT GCCCACTGTG GCCACAGCAA GCACTGGGGC CTGCACTCAG GGACCTCGGG GCCTCCTGGG GAGCTGCTGA CCCTAGGCAG AGAGATTGCA CATCCCTAAG AGTCTACAGA CACCCCAGTG TTTGCCAGTG TTTGCCCGTG TTCACCAGTG TTTGCCAGTG TTTGCCAGTA TTTGCTCGCC AGTGTTCGCC ACTTGTCCCT CTGGCTGCAA GAGTGACTGG GTTTGGGCGG GAAGTTGCAG GTCCCTCCAG GACAGTTGGC CGATGACGTG GAGACAGACC CACCCCCCAA TCCTGGCTCC CTGCAGGACG CGGGGCCCCC CGAGATCCTG GCGGTGCTCA GCACGACGGG CACCTCCGTG TTCACCAGTC CAATGGGCAC GGAGCGTGGC TTTATTTGCA TGTCTGGATT CCTAACGACT TCAGCCTCTG CACCTCCTGG GTTTTCCCTG CTGCAAATTG CCATTTGGCG TCGTCCCCAA TTTCCGGCCA AGGCCGCGTC GTCGTGCTGC TGTGTAATTT GATGTGTGGA GTTCTAGATA CCAAGTGTCT GTCGGTTTTA GACATCGCAA ACGTCCTTCC CAGTGTGGCC CGTCCATTCG CTTCTGTGCA GCAAAATCTT TAATTATTTG ATGGCATCAA AATGTGTGTC CAGTTTTACC TTCTAGTTTA TACTTTCGAA CATTTGTTTG AGAAATCTTT CTCCCACCTG TGGCTGATAG TGACGTCTTC TAACTTCCCA TTTACTATGT TACATTCAGA CCCATCATCT TCAGGAAGAC GCTTGTGTGC GAGACGGGTA TGAGGCCCCC ACACCCCGCC TCAGGACCAC TGTCCATGGT TCCACCCCTG ACCCCGGACT CCGCTCCCCA GACCTCCTAA	ASCII Code - Character to Binary 0 0011 0000 1 0011 0001 2 0011 0010 3 0011 0011 4 0011 0100 5 0011 0101 6 0011 0110 7 0011 0110 8 0011 1000 9 0011 1001 A 0100 0001 B 0100 0010 C 0100 0011 D 0100 0100 E 0100 0101 F 0100 0110 G 0100 0111 H 0100 1000 I 0100 1001 J 0100 1010 K 0100 1011 L 0100 1100 M 0100 1101 N 0100 1110 O 0100 1111 P 0101 0000 Q 0101 0001 R 0101 0010 S 0101 0011 T 0101 0100 U 0101 0101 V 0101 0110 W 0101 0111 X 0101 1000 Y 0101 1001 Z 0101 1010 a 0110 0001 b 0110 0010 c 0110 0011 d 0110 0100 e 0110 0101 f 0110 0110 g 0110 0110 h 0110 1000 i 0110 1001 j 0110 1010 k 0110 1011 l 0110 1100 m 0110 1101 n 0110 1110 o 0110 1111 p 0111 0000 q 0111 0001 r 0111 0010 s 0111 0011 t 0111 0100 u 0111 0101 v 0111 0110 w 0111 0111 x 0111 1000 y 0111 1001 z 0111 1010 : 0011 1010 ; 0011 1011 ? 0011 1111 · 0010 1110 ‘ 0010 1111 ! 0010 0001 ‘ 0010 1100 “ 0010 0010 (0010 1000) 0010 1001 space 0010 0000	
Name of basic message	Polypeptide	sentence	song
Name of smallest functional unit	Codon	byte	measure
Length of functional unit	3 nucleotides	8 bits	4 notes
Number of different values in each position of functional unit	4 (A, C, T, G)	2 (0, 1)	36 (12 notes × 3 octaves)
Number of different functional units possible	Theory: $4^3 = 64$ Actual: 20 (due to redundancy)	$2^8 = 256$	$36^4 = 1{,}679{,}616$
Estimated number of units in an average “message”	500 codons in average polypeptide	140 bytes in average sentence	80 measures per average four-minute song
Number of different average-size messages possible	Theory: 64^{500} polypeptides Actual: 20^{500}	256^{140} sentences	$1{,}679{,}616^{80}$ songs

*Traditional musical notation uses sequences of twelve different notes (C, C#, D, etc.) to designate different pitches in an octave, and most modern music uses a range of about three octaves, or thirty-six notes (we will ignore music-specific rules such as tempo, key, and scales, which influence the choice of notes in a given musical piece). The standard unit of musical notation is a *measure*, and for the sake of simplicity, we assume each measure contains only four notes.

Mutations in DNA Give Rise to Variation in Proteins, Which Are Acted on by Natural Selection

Mutations such as those discussed previously have the potential to alter the structure and function of RNAs and/or proteins. If a mutation occurs in the coding sequence of a gene, this change in the DNA informational sequence is reflected by a corresponding change in the RNA informational sequence that arises from it through transcription. In some cases, point mutations have little or no effect on the structure and function of the RNA, but in other cases, the effects of mutations can be profound, resulting in the formation of a dramatically different informational RNA or even no RNA at all.

Because three types of RNAs play a role in synthesizing proteins, mutations in their sequences have the potential to alter the sequence of amino acids created from an mRNA. This is especially true when a sequence in the coding region of an mRNA is altered because this region of mRNA determines the order of tRNAs it binds to and thus of amino acids in a new polypeptide. If the coding region of an mRNA is changed, the order of tRNAs that bind to it will be changed accordingly, and the resulting polypeptide will have a different sequence of amino acids, as shown in **FIGURE 2-6**.

A classic example is the point mutation in the hemoglobin gene that causes **sickle-cell disease**. In this case, *a single deoxyribonucleotide change* in the DNA coding sequence of the hemoglobin gene (the 17th nucleotide is changed from A to T) causes a change in the mRNA and a *single amino acid change* in the hemoglobin protein (the sixth amino acid is changed from glutamic acid to valine). This tiny change causes red blood cells to adopt a characteristic "sickle" shape when the concentration of oxygen in the blood drops because the hemoglobin protein in the red blood cells changes from its normal tetrameric configuration to long polymer strands that distort the membrane of red blood cells. Resulting sickle-shaped cells can get stuck in capillaries and/or hemolyzed (ripped open), thereby interfering with proper circulation and causing a great deal of pain (**FIGURE 2-7**).

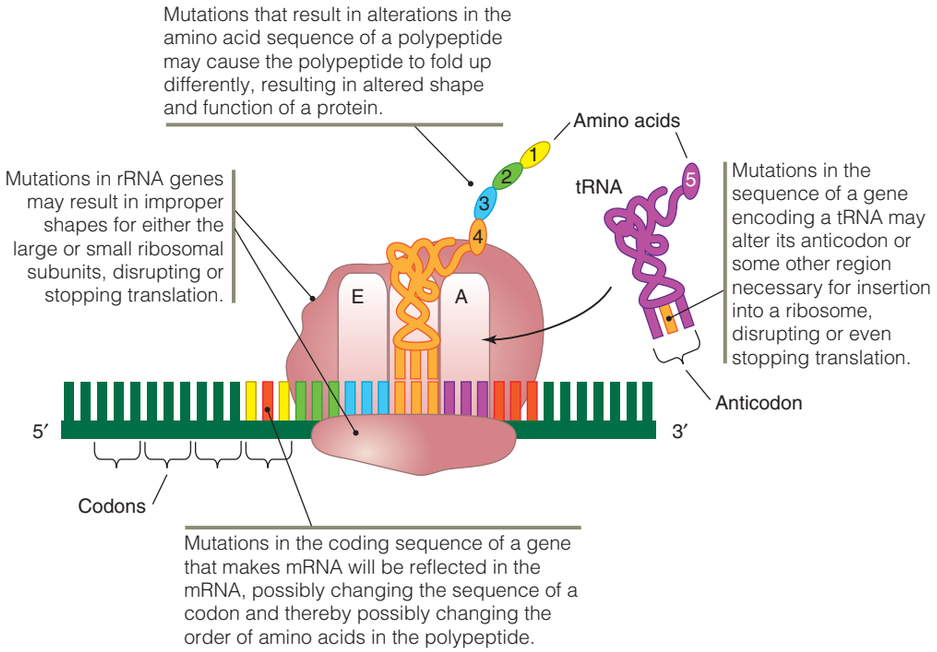


FIGURE 2-6 Mutations can alter amino acid sequence and protein function.

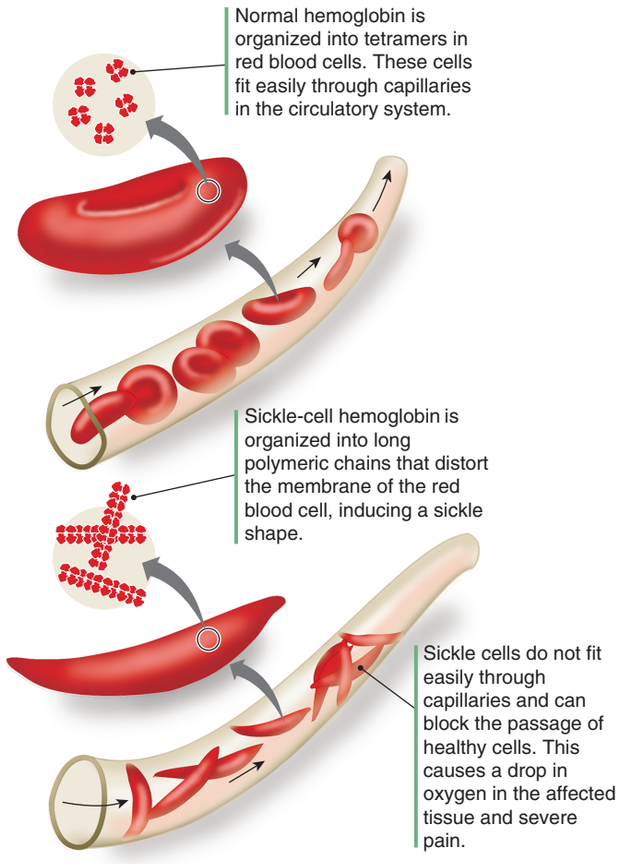


FIGURE 2-7 A single-point mutation causes sickle-cell disease.

Mutations in tRNA or rRNA mitochondrial genes can also have significant effects in cells. (Recall from Chapter 1 that mitochondria and chloroplasts are descendants of prokaryotic organisms and, therefore, perform transcription and translation of their own genes.) Several mutations have been found in the gene encoding the tRNA that carries the amino acid leucine to ribosomes in mitochondria, causing a wide range of problems (such as diabetes, degeneration of muscle fibers and nerves, and stroke-like episodes). Likewise, a single point mutation—the substitution of G for A at position 1,555—in the coding sequence of the gene for the rRNA found in the small ribosomal subunit (often called the 12S subunit) is associated with a form of deafness. The molecular mechanisms linking these mutations to their associated physiological problems are still not clear.

There are as many different possible combinations of mutations in an organism as there are nucleotides in its genotype. Some can be harmful, or even fatal, including those that cause healthy cells to become cancer cells, but most mutations do not cause serious problems for living cells and organisms. The reason for this is fairly simple: typically, alterations in a DNA sequence either (1) have little to no effect on the structure and function of the RNAs and/or proteins produced by a cell or (2) have such a drastic impact on the cell that it (and possibly the entire multicellular organism in which it lives) quickly dies. As a result, most cells in a multicellular organism, or in a population of single-celled organisms, are *subtly* different from the previous generation of cells that divided to form them, as shown in **FIGURE 2-8**. Most scientists believe that a slow, steady rate of mutation persists for several rounds of cell division until enough mutations accumulate to generate a noticeably different cell type,

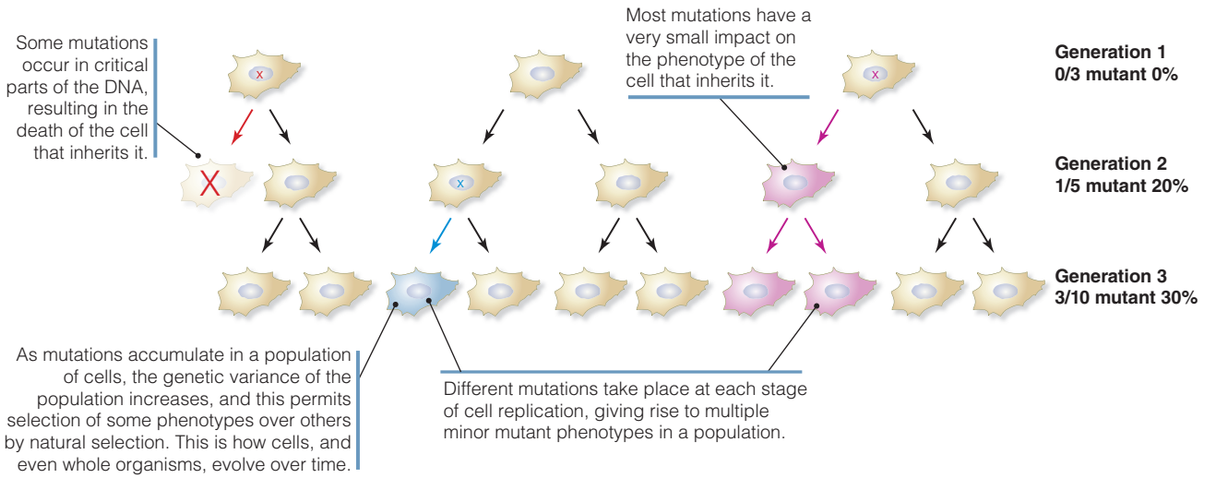


FIGURE 2-8 Mutations accumulate slowly in a population of cells.

BOX 2-4 FAQ: IS CANCER INHERITED OR NOT?

The answer is *both*. Most cancers occur as a result of mutations that accumulate in a somatic cell, such as a skin or lung cell; these cancers are not inherited because skin and lung cells are not passed from parent to offspring. But if a germ cell (e.g., egg or sperm cell) has a mutation that increases the likelihood of developing cancer, this trait, like any other genetic trait, can be passed onto offspring.

and possibly even a new type of organism if these mutations are passed on from one generation to the next.

In the short term, this can have dire consequences for an organism: most cancers arise from cells that have acquired multiple mutations from their ancestral cells over the course of an individual's lifetime (see Box 2-4). However, the same principles can have a positive impact over the long term. Each generation of cells and organisms is subtly different from that of its ancestors, and it is this variation that permits populations of organisms to adapt to changing environmental conditions. As discussed in Chapter 1 (see Box 1-3), evolution by natural selection only functions if a population of organisms contains some inheritable variation. The persistent mutation rate resulting from errors in DNA replication can, therefore, be viewed as an important tool to help ensure the survival of a species. In effect, every member of a population of organisms, including humans, can be viewed as an experiment in natural selection. We all speak the same genetic language, but we are all mutants, in one way or another.

Concept Check 1

Table 2-1 shows a comparison between the genetic code and two other codes commonly used in our everyday lives. Personal computers typically store numbers and symbols as magnetized particles on a hard drive, organized into bits and bytes, while sheet music stores musical performance instruction as symbols representing notes. What are the similarities and differences between these forms of information storage? In what ways is DNA a better or worse system than the others for storing information?

2.3 DNA Is Carefully Packaged into Five Levels of Organization

Key Concepts

- The fundamental structural unit of DNA is a deoxyribonucleotide; combinations of the four possible deoxyribonucleotides are arranged sequentially to form a linear strand of DNA.
- The simplest form of stable DNA in a cell is called a DNA double helix, formed by two strands of DNA-oriented antiparallel (i.e., oriented in opposite directions) to one another and held together by hydrogen bonds between the atoms in the base portion of the deoxyribonucleotides.
- Three different forms of double-helix DNA have been observed, suggesting that the shape of the double helix may vary in different regions of a DNA molecule.
- The length of DNA in most cells is so great that complex packaging strategies are necessary to make the DNA fit inside the cells.
- Cells construct a protein/RNA scaffold that protects and supports DNA; in prokaryotes, the resulting DNA-scaffold structure is called nucleoid, and in eukaryotes, it is called chromatin. These structures are folded and twisted to further condense the DNA inside cells.
- DNA organization is classified into five levels. The first level is simply the binding of two linear DNA strands to form a double-stranded double helix.
- The second level of packaging is a beads-on-a-string configuration, wherein the DNA double helix is wrapped around a "spool" made up of histone proteins; this shortens the length of a DNA molecule by 7-fold. Chemical modification of the histone proteins is an important mechanism for controlling the expression of genes in eukaryotes.
- The third level of packaging is the formation of 30- to 40-nm-thick fibers, made by twisting the beads-on-a-string structure into a coil. This shortens the length of a DNA strand approximately 42-fold.
- The fourth level of packaging is called looped domains, formed by periodic attachment of 30- to-40-nm fibers to a protein/RNA complex that shortens DNA approximately 750-fold.
- The fifth level of packaging is an organization of the Level 4 DNA/protein/RNA complex into a highly stable structure called a scaffold or matrix. This structure can undergo extensive compaction, also called condensation, when regions of DNA are not being read. This shortens DNA by up to 20,000-fold relative to Level 1.
- DNA compaction beyond the fourth level silences gene expression in DNA; this hypercondensed DNA is only found in eukaryotes and is called heterochromatin.

Because DNA is heritable (i.e., passed on from a cell's ancestors), it reflects the tremendous amount of information that has been gathered throughout billions of years of evolution by natural selection. Even the simplest cells have hundreds of thousands of nucleotides in their DNA. One of the smallest known genomes, that of the microbe *Nanoarchaeum equitans,* contains nearly 500,000 nucleotides. Remember that to be useful as a template, these nucleotides need to be accessible on demand. This presents a special challenge for cells: condensing DNA into a manageable size, while still permitting access to each nucleotide.

The solution to this challenge is complex. To better understand it, we will address the problem one level at a time. In this section of the chapter, we will focus on five levels of physical structure of the DNA molecule, increasing in complexity from a simple double helix to a highly compacted form called heterochromatin. In Section 2.4, we will discuss the chemical modifications cells use to control the behavior of proteins that organize DNA.

DNA Is a Linear Polymer of Deoxyribonucleotides

Before we examine the packing of DNA, let's have a closer look at DNA and how it is built (Box 2-5). We will start by examining the structure of a deoxyribonucleotide, the simplest building block in DNA, and then we will move up in complexity until we reach a complete double-stranded DNA molecule (see Box 2-6). Refer to **TABLE 2-2** to keep track of the names of the different structures as we increase in complexity.

BOX 2-5 THE LIBRARY OF CONGRESS ANALOGY, PART 3

In this section, we will learn how to create letters (deoxyribonucleotides) and link them together to form a language (DNA). We will then focus on how the letters are organized into words, sentences, books, bookshelves, bookcases, and so on, as we explore how DNA is packaged inside a cell.

BOX 2-6 TIP: STRUCTURE BEFORE JARGON

Just as we found in Chapter 1, as we go through the structure of DNA, it is likely that many new words will come up. We strongly suggest focusing initially on the fundamentals of DNA structure, and then, once the concepts are understood, spend the necessary time memorizing the names.

TABLE 2-2 The Bases, Nucleosides, and Nucleotides of RNA and DNA

	RNA	DNA		
Bases	**Nucleoside**	**Nucleotide**	**Deoxynucleoside**	**Deoxynucleotide**
Purines				
Adenine (A)	Adenosine	Adenosine monophosphate (AMP)	Deoxyadenosine	Deoxyadenosine monophosphate (dAMP)
Guanine (G)	Guanosine	Guanosine monophosphate (GMP)	Deoxyguanosine	Deoxyguanosine monophosphate (dGMP)
Pyrimidines				
Cytosine (C)	Cytidine	Cytidine monophosphate (CMP)	Deoxycytidine	Deoxycytidine monophosphate (dCMP)
Uracil (U)	Uridine	Uridine monophosphate (UMP)	—	—
Thymine (T)	—	—	Deoxythymidine	Deoxythymidine monophosphate (dTMP)

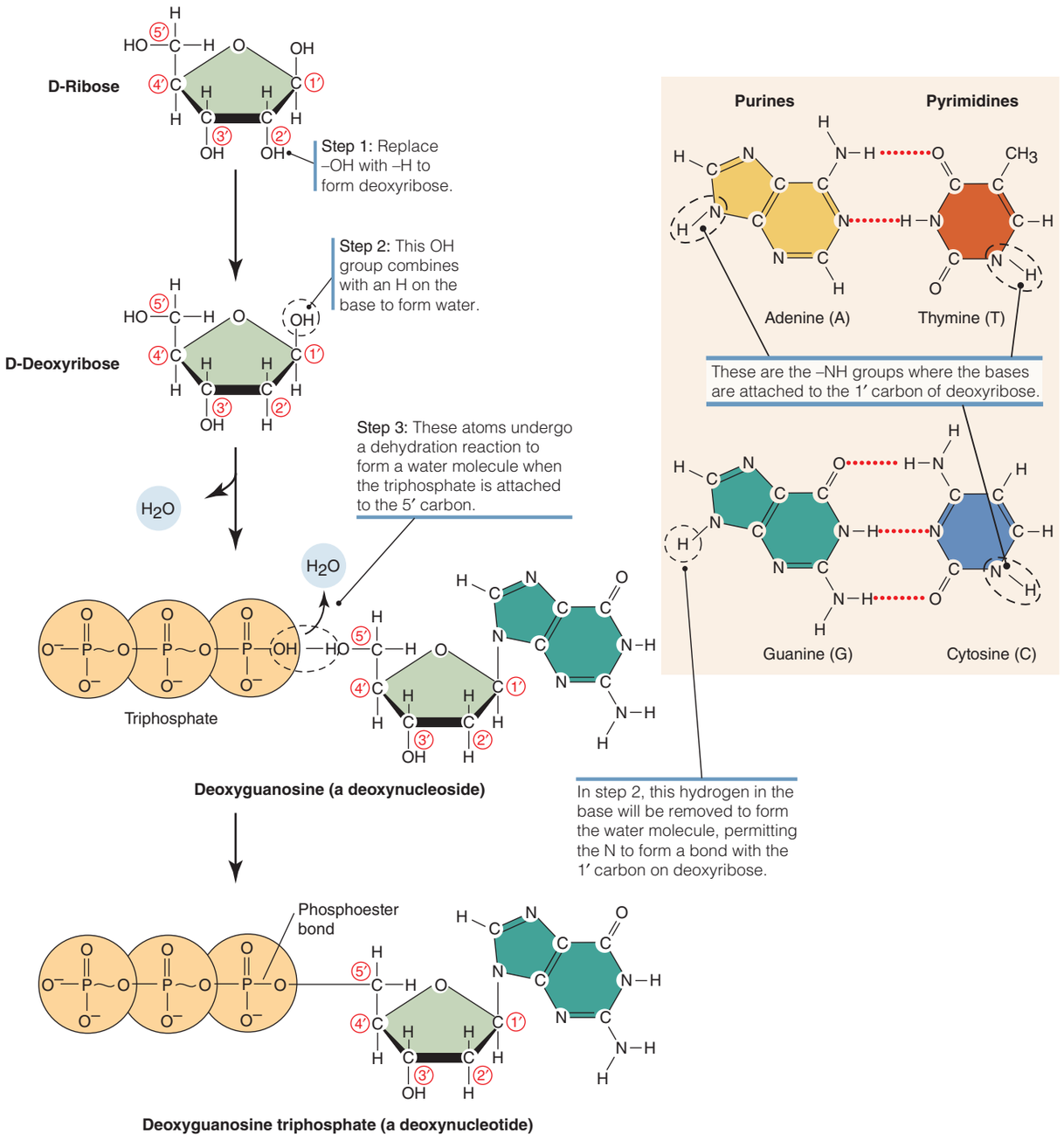


FIGURE 2-9 A stepwise method for drawing a deoxyribonucleotide.

A deoxyribonucleotide is a fairly simple structure. One good way to become familiar with this structure is to practice drawing it (use a pencil because some erasing will be required). There are three steps to this procedure, and they are outlined in **FIGURE 2-9**. Let's start by drawing ribose, a five-carbon sugar in a ring configuration (see Box 2-7). Be sure to number each of the carbons in this sugar in a clockwise fashion as shown.

- Remember that DNA is a *deoxy*ribonucleotide, so in step 1, we will take one oxygen away by replacing the hydroxyl group (–OH) on the 2′ carbon with hydrogen, yielding deoxyribose. Erase the hydroxyl group attached to the 2′ carbon, and replace it with a single hydrogen atom.

BOX 2-7 TIP

Study Figure 2-9 carefully; there is a lot of information there. Chapter 1 discusses the basic structure and nomenclature of sugars (see *Sugars Are Simple Carbohydrates*) and nucleotides (see *Nucleotides Are Complex Structures Containing a Sugar, a Phosphate Group, and a Base*). One easy way to remember the ring structure of ribose is to draw it like a cartoon house: the peak of the roof is the oxygen linking the 1 carbon and the 4 carbon, the 5 carbon represents the top of a chimney, and the 1 carbon is where the nitrogenous base is attached to the roof, like a flag on a pole.

- In step 2, we will attach a **base** to the deoxyribose. Draw a base near the 1′ carbon. We have four choices: two **purines** (adenine or guanine) or two **pyrimidines** (thymine or cytosine). Notice that all four of these bases contain a nitrogen atom bonded to a hydrogen atom that is pointing downward in the diagram; this is where we will attach the base to the ribose. We can join our base to the rest of our structure by creating a covalent bond between the 1′ carbon on the ribose molecule and the nitrogen atom on the base.

This reaction is called a **dehydration** (or **condensation**) reaction because it also yields a single water molecule as one of its products, as shown in Figure 2-9. The result is a deoxyribose sugar attached to a nitrogenous base; this structure is called a deoxyribonucleoside, or more simply, **deoxynucleoside** (see Box 2-8).

BOX 2-8 TIP

Many cell and molecular biologists use the word *base* as a shorthand term for the entire deoxyribonucleotide. A good example in this chapter is the term *base pair*. It's easy to get confused when we first encounter it. Watch out for the word *base,* and make sure to know which structure is being referred to (a deoxyribonucleotide or just the base portion of one) when we use that term.

Because four different bases can be joined to the deoxyribose, there are also four deoxynucleosides. Deoxyadenosine and deoxyguanosine contain the purines adenine and guanine, respectively, and deoxycytidine and deoxythymidine contain the pyrimidines cytosine and thymine (see Box 2-9).

BOX 2-9 TIP

Notice the difference in spelling between the base and the corresponding deoxynucleoside: although it would be convenient to simply put the term "deoxy" in front of the base to yield the corresponding name of the nucleoside, the rules of chemistry don't allow it. One way to remember the difference in spelling is that for the purines, one inserts the letters "os" before the "-ine" when referring to the nucleoside (or deoxynucleoside), and for the pyrimidines, one inserts the letters "id" before the "-ine." It helps that the word *pyrimidine* also follows this rule. Remember to think about this only *after* fully understanding the structures these words refer to.

- In step 3, we attach a triphosphate to the deoxyribose. Draw a triphosphate group near the 5′ carbon as shown. We can create a covalent bond between the 5′ carbon and the nearest phosphate by performing another dehydration reaction. The product of this reaction gets a new name: it is now called a deoxyribonucleotide, or simply **deoxynucleotide** (see Box 2-10).

BOX 2-10 TIP

Notice that the only difference in spelling between deoxynucleotide and deoxynucleoside is a single letter: s or t. Because "s" comes before "t" in the alphabet, it is easy to remember that the smaller structure contains the "s" and the bigger structure contains the "t."

There are two additional things to remember about deoxy(ribo) nucleotides:

- Deoxynucleotides have one, two, or three phosphate groups attached to the 5′ carbon, and each form gets its own name, as shown in **FIGURE 2-10**. For example, a deoxynucleotide that contains adenosine as its base can be called deoxyadenosine *mono*phosphate, deoxyadenosine *di*phosphate, or deoxyadenosine *tri*phosphate, depending on the number of phosphate groups attached to it. The bond linking the phosphate to the 5′ carbon is called a **phosphoester bond**. In Figure 2-9, we drew deoxyguanosine triphosphate. There is no deoxynucleotide that does not have a phosphate attached to it, nor can more than three phosphates be attached to a deoxynucleotide—the number is always one, two, or three. Because the formal names of deoxynucleotides are rather long, most of the time we

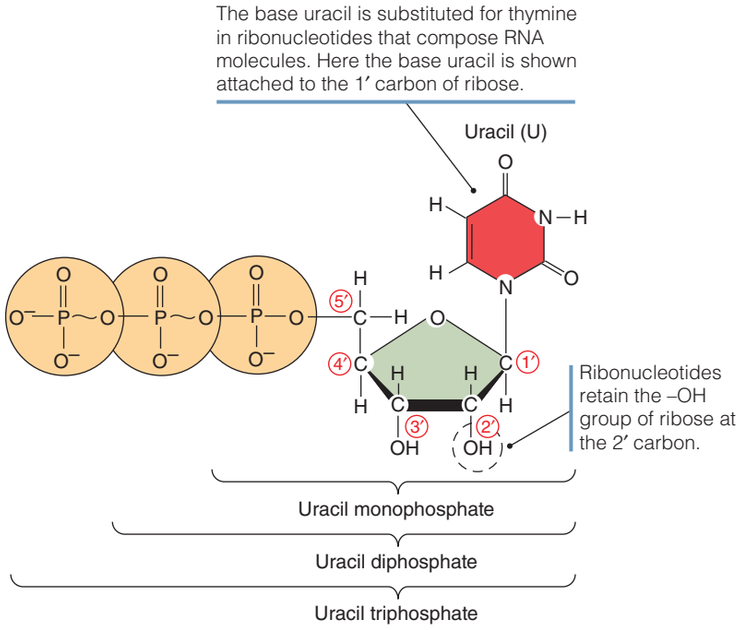


FIGURE 2-10 Distinctive features of ribonucleotides. The mono-, di-, and triphosphate forms of uracil triphosphate are indicated.

use abbreviations for them: dAMP, dADP, and dATP are the abbreviations for those that contain adenine, for example. Note that the lowercase "d" indicates that these are deoxynucleotides, to distinguish them from nucleotides that are not missing any oxygen atoms (such as those found in RNA). When we draw polymers of deoxynucleotides, we simplify the names even more by resorting to the familiar single-letter abbreviations (A, G, C, T).

- RNAs are composed of nucleotide subunits that closely resemble the deoxynucleotides in DNA (see Figure 2-10). The two most important differences between the deoxynucleotides found in DNA and the corresponding nucleotides found in RNAs are (1) the pyrimidine uracil is used in place of thymine, and (2) the nucleotides in RNAs contain ribose rather than deoxyribose (hence the absence of the term *deoxy* when naming these structures). Similar to their deoxyribonucleotide counterparts, AMP, ADP, and ATP are the abbreviations for adenine ribonucleotides containing one, two, or three phosphates. In addition to serving as building blocks for RNA, ATP and GTP are sources for metabolic energy and help control the function of proteins, as we will see in later chapters.

A Single Strand of DNA Is Held Together by Phosphodiester Bonds

Deoxynucleotides can be joined together linearly, via ester bonds, by performing dehydration reactions between the phosphate group of one and the hydroxyl group on the 3′ carbon of another (see Box 2-11). (We will discuss the exact mechanism of this reaction in much greater detail in Chapter 7.) The result is a string of deoxynucleotides linked together by an alternating sequence of phosphates and deoxyribose sugars (hence the familiar name "sugar-phosphate backbone") held together by **phosphodiester bonds**, with the nitrogenous bases extending out to the side, as shown in **FIGURE 2-11**. Note also that this string of deoxynucleotides has different structures at the two ends: no matter how long the string is, one end always has an unbound 5′ carbon (no additional nucleotides attached to its 5′ carbon), and the other end has an unbound 3′ carbon. These are called the 5′ and 3′ ends of DNA (see Box 2-12).

BOX 2-11 TIP: CHEMISTRY NOMENCLATURE

The definition of an ester bond is a covalent bond formed between an acid and an alcohol. In DNA and RNA, the phosphate group is the acid, and ribose or deoxyribose sugar is the alcohol. A single bond formed between a phosphate and a sugar is called a phosphoester bond. Two sugars attached to the same phosphate group form two phosphoester bonds, and these sugars are often described as being linked by a single phosphodiester bond.

Just as the simple building block (monomer) of DNA has a name, so too does the polymer: our strand is now called a deoxyribonucleic acid. Note that this is a generic name and that the sequence of deoxynucleotides in the strand makes no difference—any linear sequence of the four deoxynucleotides, arranged in this 5′ to 3′ fashion, is called DNA (see Box 2-13). When one wants to discuss both DNA and RNAs as a group, we use the more general term **nucleic acids**.

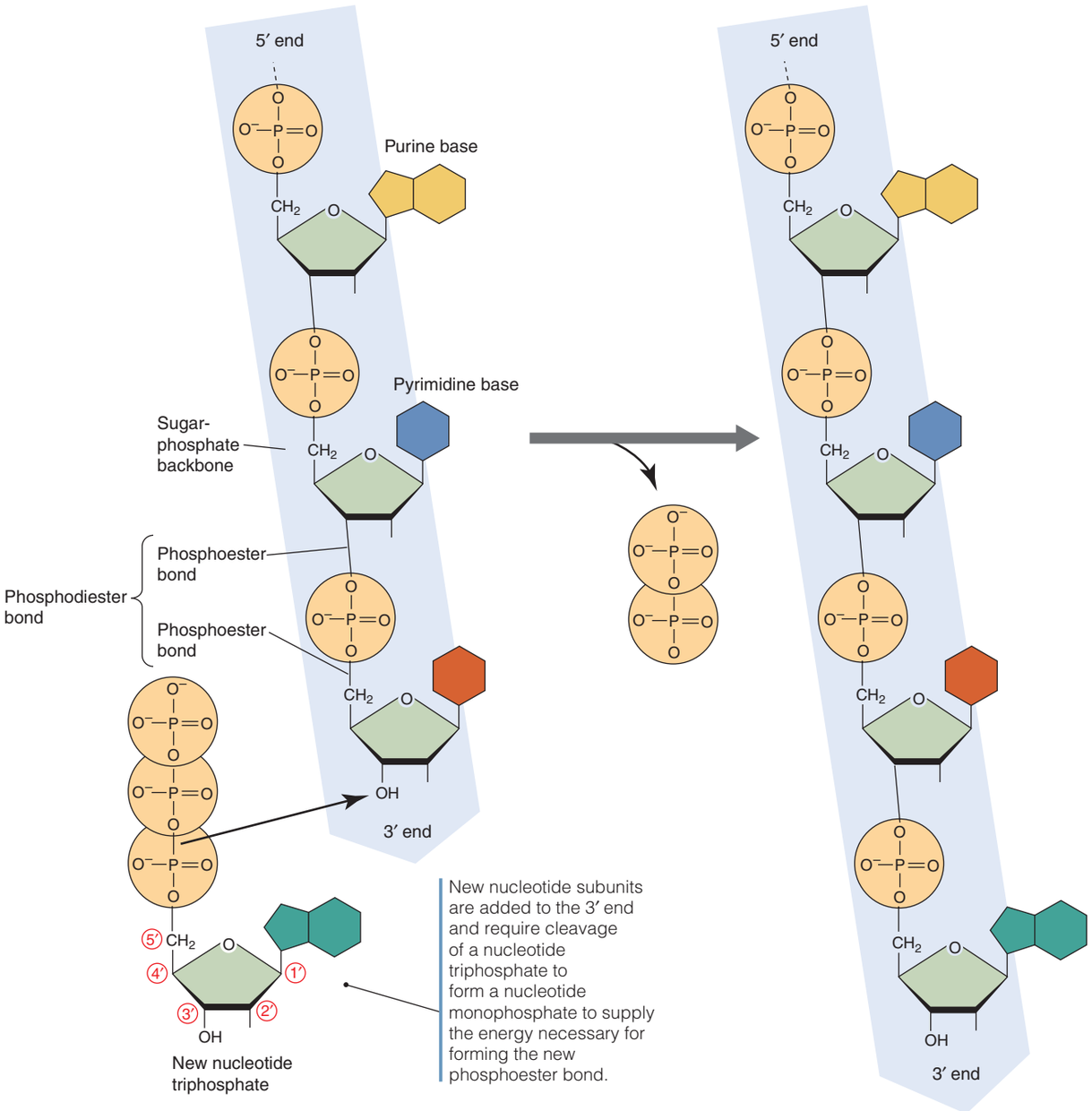


FIGURE 2-11 The general structure of a nucleic acid.

BOX 2-12 THE HOLIDAY LIGHTS ANALOGY

The linear string of nucleotides resembles a string of holiday lights: the wires represent the sugar-phosphate backbone, and the lightbulbs represent the bases. In this analogy, each bulb would be one of four colors to represent each of the four bases. One end of the wire would represent the 5′ end, and the other end would represent the 3′ end.

Single strands of DNA, such as the one shown in Figure 2-11, are not stable in cells. Cells stabilize DNA by folding/organizing it. To keep track of how it occurs, we'll identify five different levels and number them consecutively. Organization on this scale is a significant challenge for eukaryotic cells: storing

BOX 2-13 DNA SEQUENCING TECHNOLOGIES

One of the most important advances in biology in the past 50 years is the ability to "read" the linear sequence of nucleotides in DNA. The most commonly used "first-generation" technology for DNA sequencing was invented in 1977 by Frederick Sanger and his colleagues. The Sanger sequencing method relies on two simple principles: DNA cannot grow if the 3′ carbon on deoxyribose lacks a hydroxyl (–OH) group, and radioactive atoms can be detected on photographic film. During Sanger sequencing, multiple single strands of the piece of DNA being sequenced (the template) are mixed in a test tube with all of the cellular components necessary to replicate them, including all four deoxynucleotide triphosphates. As the chemical reaction takes place, a new piece of "complementary" DNA is synthesized to convert the single-stranded DNA template into a double-stranded helix. If one of the four types of deoxynucleotides being added lacks an –OH group on the 3′ end (these are called dideoxynucleotides because both the 2′ and 3′ carbons lack –OH groups), they can be inserted into the growing DNA chain, but no additional deoxynucleotides can be added to them. This shuts down DNA synthesis on that single template immediately. By "spiking" a test tube with a small amount of specific dideoxynucleotide (e.g., dideoxycytosine), biologists can conclude that each time DNA synthesis stops in this reaction, the last deoxynucleotide to be added must be this dideoxynucleotide. Performing this reaction on thousands of copies of a DNA template yields many different copies of double-stranded DNA, each of which contains a known dideoxynucleotide at its 3′ end. The phosphate groups in the newly added dideoxynucleotides contain radioactive phosphorus (P^{32}), which makes the newly synthesized DNA radioactive. These copies are separated, based on their length, using gel electrophoresis. The separated pieces are visualized by drying the gel on a sheet of nitrocellulose paper and placing it next to X-ray film; the radioactive phosphorus exposes the film, so when it is later developed, small lines (called bands) appear on the film. Each line represents a different length of DNA that contains a known dideoxynucleotide at the 3′ end.

When this procedure is repeated in four test tubes, each containing a small amount of one radioactive dideoxynucleotide (A, C, T, or G), the contents of all four tubes can be separated in different lanes of the same gel. The resulting pattern appearing on the developed X-ray film shows a series of bands that differ in length by one known dideoxynucleotide. Reading the gel band pattern from the smallest copy to the largest across all four lanes allows researchers to deduce the sequence of the newly synthesized DNA, from the 5′ end (smallest band) to the 3′ end (largest band). Because the two strands of a DNA helix are base-paired, researchers can also infer the sequence of the template that gave rise to the radioactive strands.

Later, radioactive dideoxynucleotides were replaced with fluorescently labeled versions, and these can be visualized with highly sensitive light detectors; radioactive gels were replaced with four-color readouts of DNA sequences. As technology improved, DNA amplification allowed hundreds of DNA templates to be sequenced at the same time, making it possible to sequence the entire genome of a cell with a robot. This technology produced the first complete human DNA sequence in 2001.

"Second-generation" DNA sequencing (also called next generation sequencing, or NGS) relies on immobilizing many copies of a single DNA template on a solid surface, adding the necessary proteins and replicating DNA, then washing the template with successive solutions containing only one type of deoxynucleotide triphosphate. If, for example, the template requires a T to be inserted in the newly synthesized strand at the 3′ end, washing the template with a solution containing dideoxythreonine triphosphate but no other deoxynucleotides will result in the cleavage of the triphosphate on deoxythreonine, generating a new 3′ end containing T and a pyrophosphate (two phosphates cleaved off the 5′ end). By measuring the amount of pyrophosphate generated after each wash with a known deoxynucleotide triphosphate solution, scientists can determine whether a given "base" was added to the growing DNA strand. Passing each type of "base" solution over the template in a repeated cycle yields a surge in pyrophosphate each time a base is added. Summing the bursts in pyrophosphate with each round of washes allows scientists to determine the order of base addition and, by extension, the 5′-to-3′ sequence of the growing DNA chain.

The advent of second-generation sequencing/NGS spawned a revolution in commercial DNA sequencing technologies. One of the most significant was replacing pyrophosphate detection with stepwise, fluorescence-based detection of bases as they are added to the growing 3′ end. This removes the need for repeated cyclical washes and allows scientists to observe the addition of bases by measuring the fluorescent color emitted by a "spot" of DNA template after each round of base addition. Now, millions of spots can be arrayed on a slide, exposed to all four types of fluorescently labeled deoxynucleotides at once, and the sequence of each spot is revealed by the release of color each time a new round of base addition is completed.

The field of second-generation DNA sequencing is expanding rapidly, with new approaches appearing every year. Unlike the situation in the days of Frederick Sanger (1918–2013), now many different types of DNA

sequencing are appearing in a highly competitive commercial market. The cost of sequencing an individual's entire genome is dropping so rapidly that scientists estimate that over 2 billion humans will have their genome sequenced by the mid-2020s.

FIGURE A Using NGS technology, scientists can read DNA sequences by measuring pyrophosphate release from specific nucleotides as DNA is synthesized.

genetic information, accumulated over billions of years of evolution, in the form of linear sequences of deoxyribonucleotides means that these sequences are very, very long. For example, the roughly 12 billion deoxyribonucleotides in an average human somatic cell form about 6 billion base pairs that, if laid end-to-end, would be a string over 2 meters long, hundreds of thousands of times the size of the cell containing it. If each of the deoxynucleotides was listed using its single-letter abbreviation, these letters would occupy more than a million pages of a typical book. How can all that information be packed into a single cell without it getting all jumbled up (see Box 2-14)? Here is where we describe Levels 1 through 5 of DNA organization.

BOX 2-14 THE LIBRARY OF CONGRESS ANALOGY, PART 4

In Section 2.3, we focus on how DNA information is stored in a highly organized, easy-to-use system. The library analogy applies quite well here: the books of information are stored on shelves, the shelves are arranged into rows, and the rows are assembled on floors of the building. Some books, because there is little or no demand for them, are stored away in shelves that are difficult to access. While the shelves do not encode much information themselves, they are essential for the books to be useful. Here, we will learn how cells build and arrange the protein/RNA "shelves" that support DNA.

Level 1: DNA Forms an Antiparallel Double Helix

The simplest form of stable DNA in cells, which we will call Level 1, is a double-stranded DNA molecule where the two strands run antiparallel to one another (one strand of 5′ to 3′ is alongside one that runs 3′ to 5′) and are held together by hydrogen bonds between oxygen and nitrogen atoms in the **complementary bases** to form base pairs. The absence of the –OH group on the 2′ carbon of deoxyribose allows the two DNA strands to twist around one another to form a helix (most RNA molecules never form a double-stranded helix). **FIGURE 2-12** shows three common representations of double-stranded DNA to highlight different structural features of the molecule. Figure 2-12a is a simple line drawing illustrating how complementary base pairing holds the

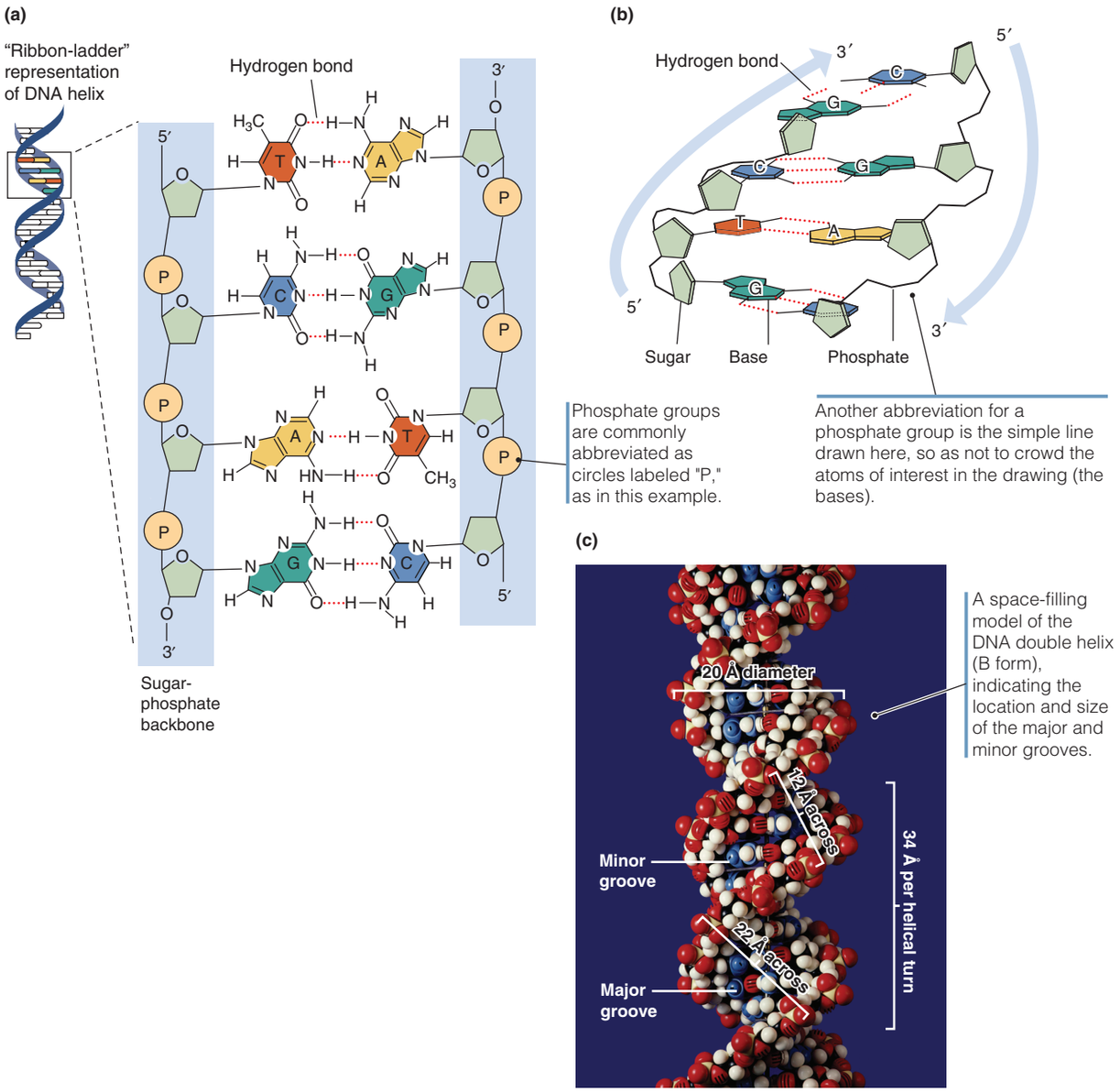


FIGURE 2-12 Level 1 of DNA organization is a double-stranded, antiparallel double helix held together by hydrogen bonds between base pairs. (a) A simple line drawing of the double-stranded double helix. A "ribbon-ladder" representation of the double helix is shown for reference. (b) A 3-D drawing showing the spatial arrangement of the nucleotides in a DNA double helix. (c) A space-filling model of the DNA double helix (B form), indicating the location and size of the major and minor grooves.

two strands of DNA together. Figure 2-12b uses a three-dimensional drawing to demonstrate that the hydrogen bonds between complementary bases stabilize the two strands in the double helix.

Practice drawing the ribbon-ladder form of DNA, as shown in Figure 2-12a. When doing so, make sure that the double helix contains two different grooves. The wider of these grooves is called the **major groove**, while the narrower is called the **minor groove**. These grooves are important because they form attachment sites for DNA binding proteins.

DNA binding proteins often contain finger-like structures that fit into these grooves; these structures allow DNA binding proteins to slide back and forth

in the grooves as they search for the specific sequence of deoxynucleotides they are targeting, as shown in Figure 2-3. Note that the twisting of the DNA strands results in a periodicity of approximately 3.4 nm (or 34 Å); this means that there are approximately 10.5 base pairs per turn of the helix. This matters because many DNA binding proteins bind to short DNA sequences (six base pairs or fewer). Because these sequences are shorter than a single turn of the helix, they seem more or less linear to DNA binding proteins, which makes them easy to detect.

Figure 2-12c is a space-filling model of the DNA double helix, and it illustrates one final, very important point that we will discuss in greater detail in later chapters. It shows that DNA is composed of a large number of atoms. This is important because changes in even a few atoms result in noticeable variations in the DNA structure. For example, a region of DNA encoded by several A–T base pairs will have a slightly different shape than one encoded by G–C base pairs, even though both pairs are perfectly aligned.

The fact that each base pair imparts its shape to the overall molecule means that two segments of DNA made up of different deoxynucleotide sequences also have slightly different shapes. It is this difference in shape that allows proteins to "know" which regions of DNA to bind to. Stated more simply, every different sequence of deoxyribonucleotides has a unique shape. Proteins that bind to specific sequences of DNA can, therefore, slide along a strand of DNA until they find the exact shape that fits their binding site. Even very minor changes in the atomic structure of deoxynucleotides can have profound effects on overall DNA shape: one common cause of DNA mutation is the loss of a single amino group ($-NH_2$) from the base of a single deoxynucleotide.

DNA Can Be Supercoiled to Form at Least Three Different Structures

Discovering the double-stranded, helical organization of DNA was one of the most significant advances in biology during the twentieth century. A considerable amount of the data used to deduce this structure came from crystals of DNA grown in the lab. These crystals also demonstrated that double-stranded DNA could form at least three different types of double helices. DNA adopts the configuration shown in **FIGURE 2-13**, called B-DNA, under conditions of high relative humidity (92%). A second configuration, called A-DNA, appears when DNA is crystallized under conditions of lower relative humidity. (The third type of double helix is called Z-DNA.) All three types of double

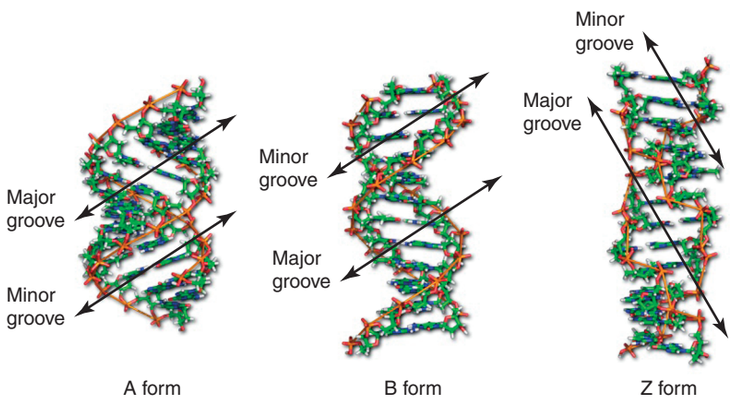


FIGURE 2-13 Three different forms of double-helical DNA (L to R: A form, B form, Z form). Note that all three forms have major and minor grooves.

Courtesy of Richard Wheeler.

helices have a major and a minor groove. Because the interior of a cell is entirely saturated with water, most of the DNA in a living cell likely adopts a shape very similar to B-DNA. Small regions of Z-DNA have been detected in cells, though it is still unclear what significance this conformation plays in chromosome function. Both A-DNA and B-DNA helices are "right-handed," meaning that if one holds a piece of it in front of one's eye like a telescope, the helix will appear to turn in a clockwise fashion. Z-DNA is twisted in the opposite ("left-handed") direction, so it has been suggested that Z-DNA regions may serve to reduce the amount of effort required to unwind B-DNA in areas of the chromosome that are frequently copied during transcription.

Level 2: DNA Is Bound to a Protein/RNA Scaffold

The next part of the strategy for packaging DNA in cells is to support it with an elaborate infrastructure made of proteins and RNA. The proteins and RNA don't store any information, but without them, the DNA would be hopelessly tangled and altogether useless. The complex formed by these proteins, RNA, and their associated DNA is called chromatin in the nuclei of eukaryotes, and nucleoid in mitochondria, chloroplasts, prokaryotes, and Archaea. Proteins account for at least 50% of the mass of chromatin and are likely as abundant in the nucleoid.

Double-Stranded DNA Is Wrapped Around Histone Proteins to Form a Small Particle

The best-known set of structural proteins belongs to the **histone** family. These proteins are found in all organisms and are thought to be some of the earliest proteins to appear during evolution. When associated with DNA, they form spools called **nucleosome core particles**, similar to those used to store thread, string, wire, and the like. This is Level 2 of DNA organization. In eukaryotic cells, these spools are composed of two copies each of four different histones (named H2A, H2B, H3, and H4). The histones contain many positively charged amino acids, which attract them to the negatively charged backbone of DNA in a way that is largely sequence independent. The double-stranded DNA molecule is wrapped around a histone "spool" approximately 1.7 times (167 base pairs). This spool, plus the short stretch (20–50 base pairs) of DNA called **linker DNA** that lies between spools, is called a **nucleosome**, as shown in **FIGURE 2-14**. A linear arrangement of several nucleosomes forms a structure that looks like a string of beads (**FIGURE 2-15**). The addition of a "linker" histone, either H1 or H5, "pins" the DNA to the core particle, resulting in a structure called a **chromatosome** (see Figure 2-14). Similar structures are found in prokaryotic cells, where the DNA is wrapped around a different set of histone protein spools. Wrapping DNA in this fashion causes the DNA double-helical strand to become shorter and thicker: the length is reduced by approximately seven-fold, and the width increases from 2 nm to about 11 nm. These "beads-on-a-string" structures also contain proteins in addition to histones in both prokaryotes and eukaryotes.

Does wrapping DNA around a spool have any negative consequences? Remember that the goal is to compact DNA without compromising a cell's ability to access the genetic information. Because so much of the DNA comes into contact with the spool, most of it is inaccessible to other DNA-binding proteins, thereby negating the beneficial effects of these spools. In eukaryotes, DNA can be partially unwrapped from the nucleosome by members of the **SWI/SNF** family of proteins. These proteins use ATP energy to move the core particle a short distance along the DNA, thereby freeing up any base-pair

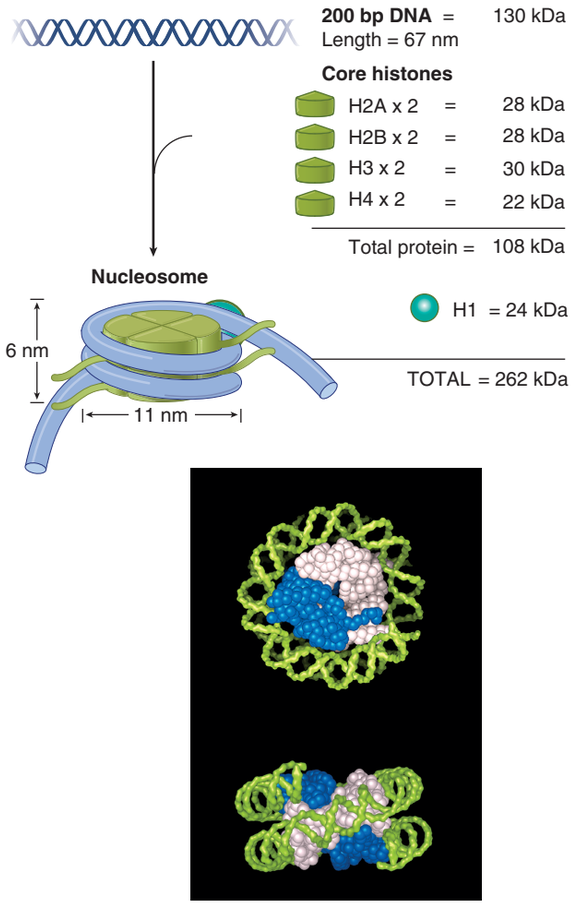


FIGURE 2-14 Two representations of the chromatosome. The top panel shows the octameric arrangement of histones in a nucleosome core particle, with DNA wrapped around it. Histone H1 pins the DNA to the core particle, forming a chromatosome. The lower panel shows a computer model of DNA wrapped around the nucleosome core particle.

Photos courtesy of E. N. Moudrianakis, John Hopkins University.

sequences that may have been buried in the core particle. At least two other families of proteins participate in this type of **chromatin remodeling** as well; this is illustrated in Figure 2-15.

Histone Modifiers Control the Structure of Nucleosomes

Chromatin remodeling, illustrated in **FIGURE 2-16**, is the term used to describe the process of displacing histones to control access to DNA. In some cases, this means that cells simply slide nucleosomes from one position in chromatin to another, which can change the spacing between nucleosomes. In other cases, entire nucleosomes are temporarily removed from a particular region of DNA. Large ATP-consuming complexes are primarily responsible for performing these remodeling activities. Humans have at least eight different complexes, and the best-known complex is called SWI/SNF.

Level 3: DNA Is Twisted to Form Fibers

The beads-on-a-string structure we see in the microscope only appears when we break cells open and spread out their contents. In intact cells, chromatosomes are clustered together in a highly ordered fashion to form a series of similar configurations, all of which are called the **30-nm fiber**.

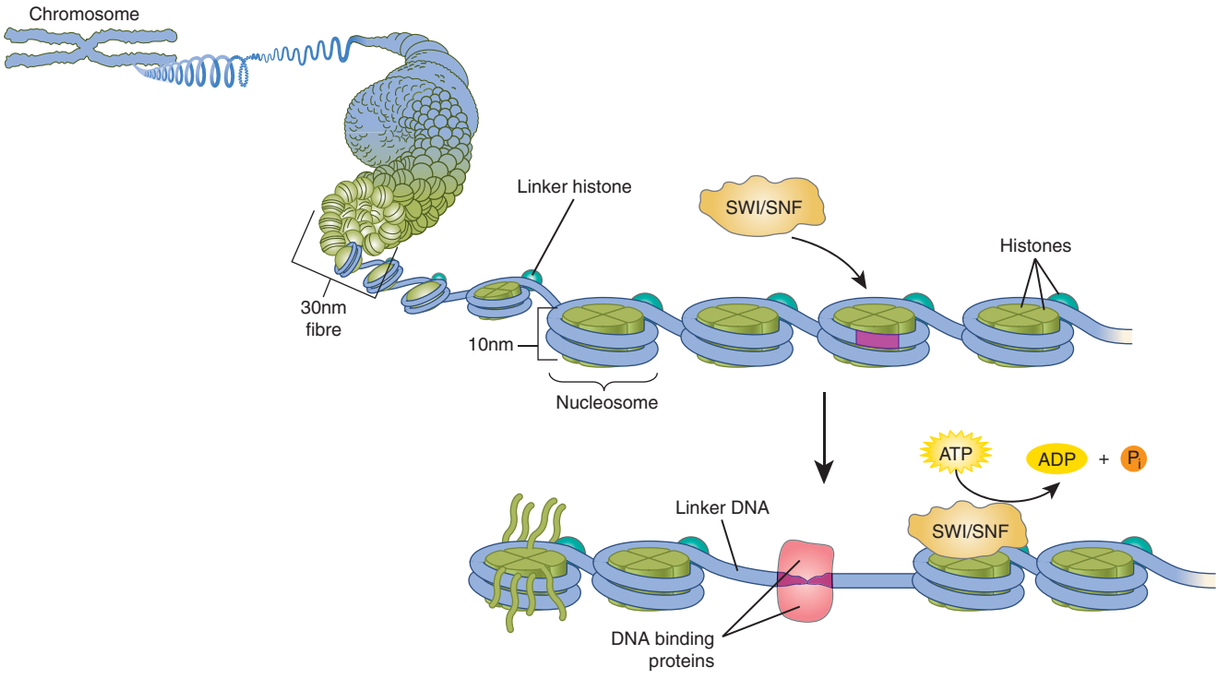


FIGURE 2-15 Nucleosomes are linked by short segments of linker DNA to form a "beads-on-a-string" arrangement. Proteins named switch/sucrose nonfermentable (SWI/SNF) can slide the core particles to widen and narrow the gaps between nucleosomes.

BOX 2-15 FAQ: WHAT IS THE DIFFERENCE BETWEEN A NUCLEOSOME AND A CHROMATOSOME?

In simple terms, a chromatosome is a nucleosome attached to a linker histone. The important distinction to keep in mind is that all of the mechanisms discussed here about controlling DNA access focus on the shape of the nucleosome, not the chromatosome. This means that loosening of DNA around the spool, sliding DNA over the surface of the spool, and even removing the spool altogether takes place *when the linker histone is not present*, and thus this is all happening to a nucleosome. After the modification has occurred and the linker histone returns to bind the nucleosome, it again becomes a chromatosome. DNA reforms a 30-nm fiber when chromatosomes bind together via their lnker histones.

(In prokaryotes, similar **40-nm fibers** form from clusters of the "beads" in the genophore.) These fibers represent Level 3 of DNA organization. Multiple configurations have the same name because scientists are still not entirely sure how many of the configurations are present in living cells (the electron microscope, which is a common tool for observing chromatin, cannot be used to view living cells). These fibers will form spontaneously in a test tube if the salt concentration of the buffer is kept low enough, demonstrating that no additional proteins or metabolic energy are required.

The 30-nm fiber is held together by electrostatic interactions between different histones. For example, a negatively charged region on histone H4 binds to a positive region in a histone H2A/H2B complex in another nucleosome,

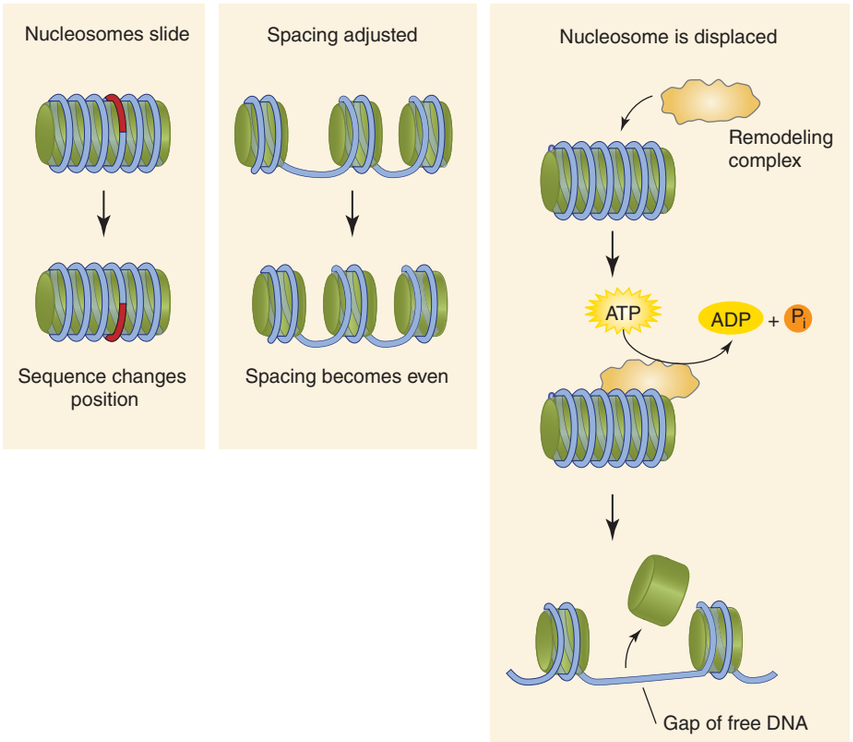


FIGURE 2-16 Chromatin remodeling is a modification of histones to permit their removal or sliding along a piece of DNA. Opening up gaps between nucleosomes permits RNA polymerases and transcription factors to bind to promoters.

drawing the two nucleosomes together and further compacting the chromosome. In addition, linker histones bind to each other, as well as to other proteins that serve as bridges between chromatosomes. This results in additional shortening and thickening of the chromosome and increases the packing density to about 42 times that of double-stranded DNA alone.

Level 4: DNA Fibers Attach to a Protein-RNA Scaffold

At Level 4 of DNA organization, the 30-nm fibers are attached to a protein-RNA **scaffold** (also called a **matrix**) that keeps them organized (**FIGURE 2-17**). Specifically, the fibers are attached to the scaffold at intervals of 10–30 μm, thereby forming so-called **loop domains** of approximately 60 kilobases in length. Exactly how these loops are attached to the scaffold is not known, but the attachment occurs at DNA sequences called MARs or SARs (for *matrix* or *scaffold attachment regions,* respectively). These sequences typically contain a large number of A–T base pairs but are otherwise not very similar. Many proteins have been isolated from the chromosome scaffold, and the structure is sensitive to enzymes that digest an unknown form of RNA, but it is not yet clear how these molecules assemble to form the mature structure. Loop domains are approximately 750-fold more compact than B-DNA. The prokaryotic chromosome is thought to consist entirely of Level 1–4 structures.

Level 5: Chromatin Is Packaged into Highly Condensed Chromosomes

Eukaryotic cells adopt an additional means for organizing DNA that most prokaryotes do not use: they cut their DNA up into several chromosomes. Human

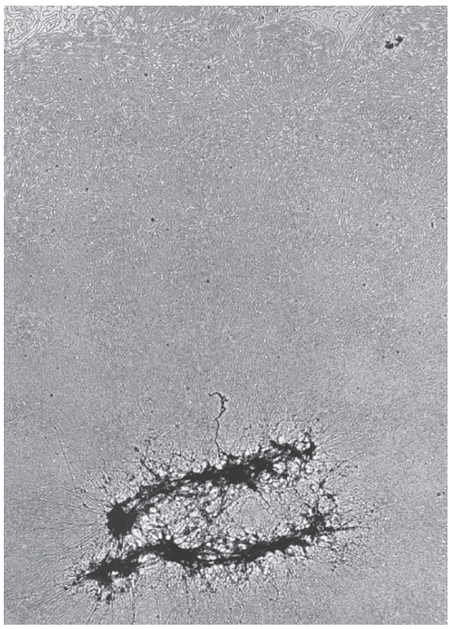


FIGURE 2-17 Level 4 of DNA organization can be seen in this electron micrograph of loop domains projecting outward from the protein/RNA scaffold (dark material at the bottom).

beings are considered to be diploid, meaning they contain two copies of each chromosome that range in length from approximately 47 million base pairs to nearly 250 million base pairs. These are organized as two copies each of 22 autosomes and 1 pair of sex chromosomes. We can tell by simply looking in a microscope that chromosomes are very large bundles of DNA with distinctive shapes that change throughout a cell's life. During mitosis, these chromosomes condense to form their familiar X-shaped structures, and they decondense once mitosis is complete, as shown in **FIGURE 2-18**. This condensation/decondensation is analogous to packing one's belongings into a small, compact space (e.g., a suitcase for a trip) and then unpacking once the journey is complete.

These observations reveal two important things. First, DNA organization is dynamic: chromosomes can be tightly bundled or loosely bundled as necessary during a cell's lifetime. Second, this implies that there must be some machinery responsible for controlling this bundling. Consider how important this machinery is. During mitosis, a eukaryotic chromosome can be condensed into a structure that is about 15,000 to 20,000 times shorter than its unwound length. Changes in the length of a chromosome result from a complex series of folding and twisting events designed to prevent the individual strands from becoming tangled.

Heterochromatin Is a Form of Tightly Packed DNA in Eukaryotic Cells

Collectively, Levels 1–4 of DNA organization are called **euchromatin** in eukaryotes because they all share one important property: they can be easily accessed by proteins responsible for replicating the chromosomes in preparation for cell division or by proteins responsible for reading a strand of DNA to make RNA (i.e., transcription). In other words, DNA sequences organized

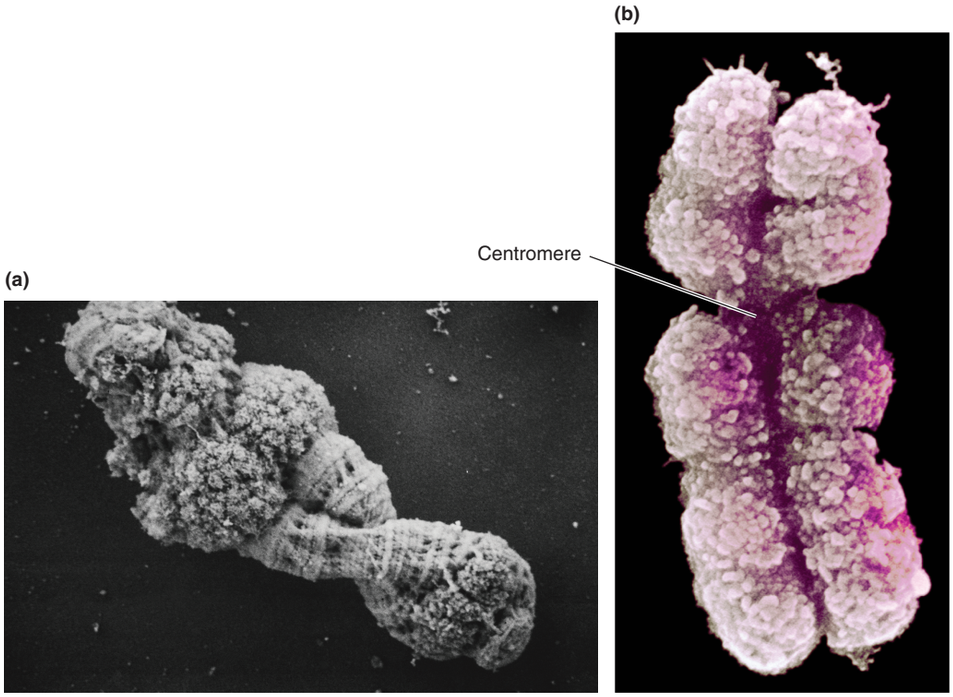


FIGURE 2-18 (a) Chromosomes at different levels of DNA compaction. These chromosomes have banded regions containing tightly wound chromatin and loose "puffs" where genes are easier to access. (b) This chromosome is fully condensed, as it would look during mitosis.

as a form of euchromatin are *easy to use* (**FIGURE 2-19**). This helps explain why prokaryotic cells can alter their gene expression patterns fairly rapidly, compared to most eukaryotes: their entire DNA is easily accessible.

Eukaryotes go even further in compacting their chromosomes. Advancing to Level 5 is a very big step for these cells. Any portion of a chromosome that condenses past the point of loop domains becomes essentially inactive. Unlike the case in prokaryotes, however, a considerable amount of DNA in eukaryotic chromosomes is actually rather useless to a cell. This DNA may contain genes or fractions of genes that were important for our distant ancestors but are no longer useful. Or it may contain genes that are useful only during early embryonic development or in specialized cells (e.g., humans form gill-like structures only during early development; usually only nerve and muscle cells make neurotransmitter receptors while only bone cells make bone proteins). Once a cell commits to a specific developmental fate (e.g., nerve, muscle, or bone), it no longer needs access to at least some of the genes required for other fates (e.g., skin or liver). A nerve cell can afford to "pack up" the portions of its DNA containing instructions necessary for functioning as a liver cell because it does not ever expect to use them.

Cohesins and Condensins Help Control the Packaging State of Chromatin

The additional condensation of DNA is accomplished by twisting loop domains into shorter and thicker filaments. Several different structures are possible, depending on the extent of twisting used. The degree of this compaction has been estimated to be between 250- and 20,000-fold. We will break this range into two parts: those areas of condensed DNA found in cells when

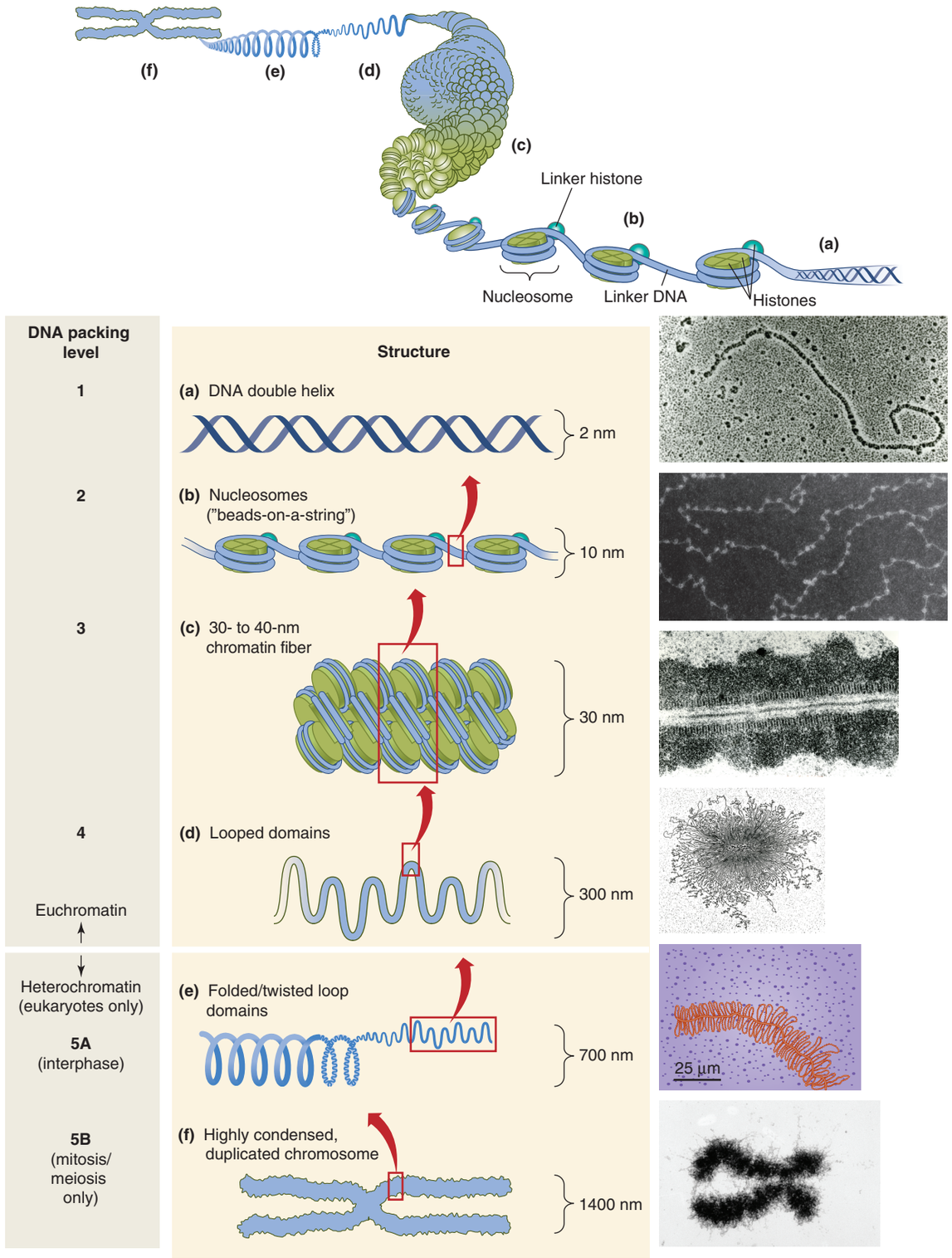


FIGURE 2-19 The five different levels of DNA organization in eukaryotic cells. Note that two types of Level 5 (heterochromatin) are shown: 5A is found in nondividing (interphase) cells, and 5B is found only in cells undergoing cell division (mitosis or meiosis).

Photo (a) © Science Source; Photo (b) © Fancy Tapis/Shutterstock; Photo (c) Courtesy of Barbara Hamkalo, University of California, Irvine; Photo (d) Courtesy of Bruno Zimm and Ruth Kavenoff. Used with permission of Georgianna Zimm, University of California, San Diego; Photo (f) Courtesy of the Cell Image Library.

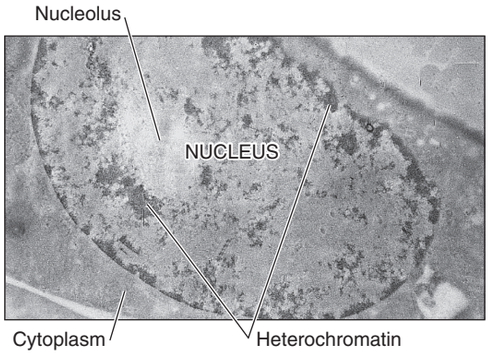


FIGURE 2-20 Heterochromatin appears as dark patches in the nucleus during interphase. This is Level 5A of DNA organization.

Photo courtesy of Edmund Puvion, Centre National de la Recherche Scientifique.

they are not actively dividing (a period also called interphase) are Level 5A; more compact chromosomes are required for cells to undergo the mitotic or meiotic phase of cell division, and this extra degree of compaction is Level 5B. Regardless of their size, all of these Level 5 structures are called **heterochromatin**, both to differentiate them from euchromatin and because they appear as blobs of varying darkness in an electron microscope (**FIGURE 2-20**). Like the condensation of chromatosomes to form 30-nm fibers, this condensation requires additional proteins.

Two good examples are those belonging to the **SMC** (*s*tructural *m*aintenance of *c*hromosomes) family of proteins. One group, called **condensins**, is responsible for general chromosome structure, along with chromosome condensation during the prophase period of cell division (see *Prophase Prepares the Cell for Division* in Chapter 7). A second group, called **cohesins**, plays an important role in condensation of yeast chromosomes during cell division and regulates access to genes in virtually all eukaryotes.

FIGURE 2-21 shows models of how they might accomplish this condensation. Cohesins bind two strands of DNA together by forming a ring-shaped structure that encloses Level 1 and 2 forms of DNA, and condensins are thought to condense DNA by forming similar ring-shaped structures enclosing Level 3 loops of DNA. While it is known that DNA condensation requires ATP energy, it is still not entirely clear exactly how that ATP is used. It is possible that these proteins hydrolyze ATP to gather the DNA into these rings.

RNAs Play a Critical Role in Compacting Chromatin

Unfortunately, our understanding of how chromosomes are compacted is hampered by the same teamwork between molecules that forms our first principle of cell biology (discussed in Chapter 1). As chromosomes condense, the molecular teams responsible for regulating this process grow in both size and complexity. A very common method for deciphering how a molecular team works is to break it into its constituent parts and then reassemble it to figure out how each part interacts with the others. For condensed chromosomes, even breaking the scaffold into simpler parts is very difficult: it is resistant to most chemicals and remains largely intact even when almost everything else in a cell is broken apart.

What we have learned so far is that a significant portion of the nuclear scaffold is made up of RNA molecules and that most of these belong to the seventh class of RNAs discussed in Section 2.2. These RNAs play two important

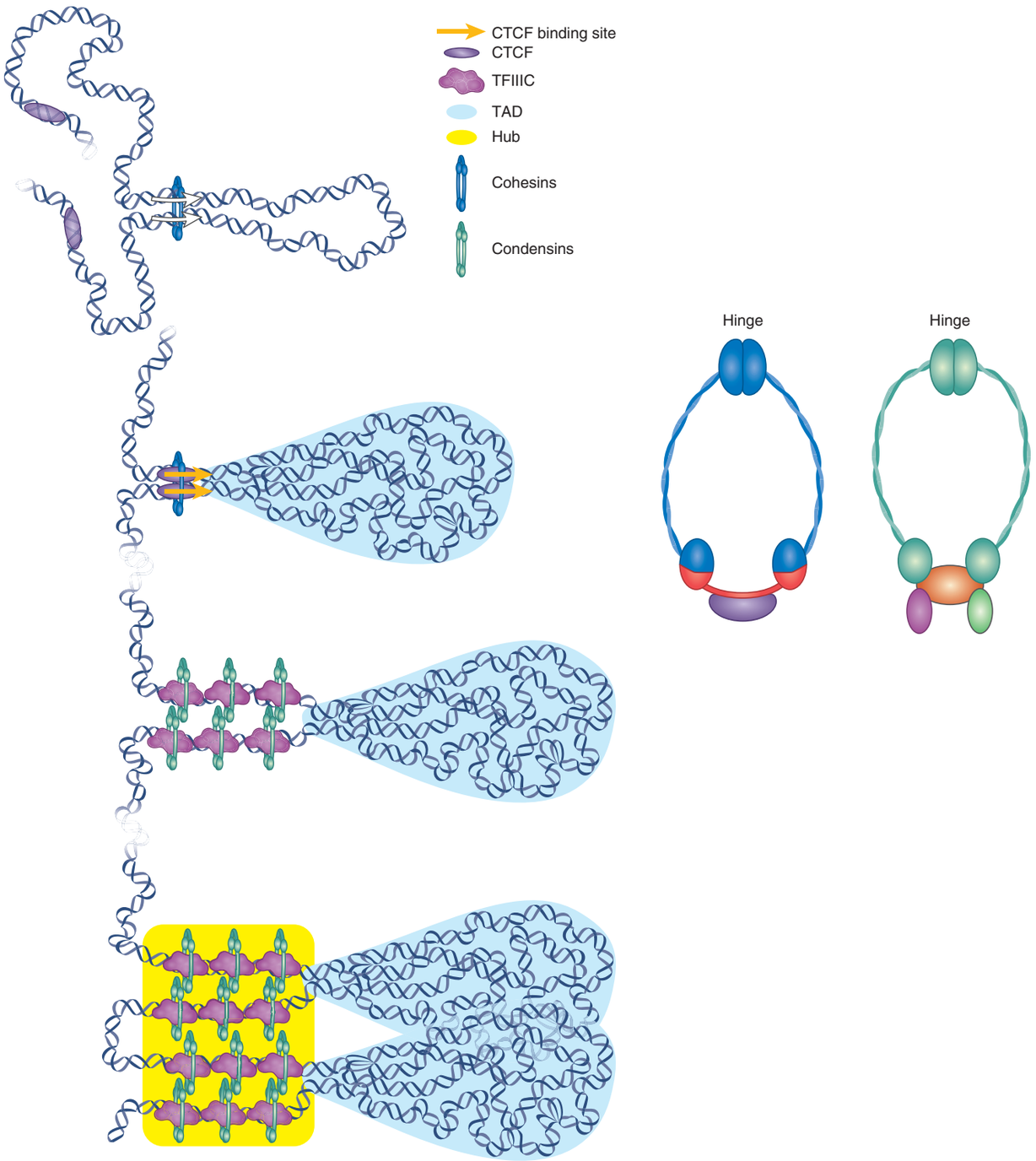


FIGURE 2-21 Cohesins and condensins control the spatial arrangement of chromatin.

roles in forming the scaffold and controlling the compaction of chromosomes. First, they fold back on themselves to form double-stranded RNA "loop-stem" structures that are stabilized by hydrogen bonds between bases. This makes them unexpectedly resistant to physical or chemical disruption; the stem-loop regions bind directly to proteins within the scaffold and help hold them in place. Second, these RNAs serve as binding partners for additional RNAs and proteins that trigger DNA condensation. One important class of these proteins

is histone-modifying proteins (discussed in Section 2.4); the modifications to the histones, in turn, control the degree of compaction of the scaffold and thereby help determine whether a region of DNA is accessible for gene transcription. Other proteins link the scaffold to the inner surface of the nucleus, helping to stabilize it.

One of the best-studied nuclear scaffold RNAs is called X-inactive-specific transcript (Xist) (see Box 2-16). Xist is both necessary and sufficient to compress an X chromosome into a permanently inactive structure (often abbreviated as Xi) called a Barr body. Female mammals contain two copies of the X chromosome, and it is important that one of these be inactivated in every cell soon after fertilization; failure to silence one of the X chromosomes results in overexpression of sex-specific genes and failure to complete embryonic development. Xist is the consummate team player: scientists estimate it has at least thirty protein binding partners, forming hundreds of potential combinations (i.e., different teams).

BOX 2-16 CASE STUDY: DUCHENNE MUSCULAR DYSTROPHY IN A FEMALE PATIENT

Muscular dystrophies are a group of disorders that share clinical characteristics of progressive muscular weakness. Duchenne muscular dystrophy (DMD) is the most common type of muscular dystrophy. It includes delays in muscle development, wheelchair confinement, and cardiac and respiratory problems that prematurely end the patient's life. DMD is an X chromosome–linked recessive disorder. Because girls inherit one X chromosome from each parent, they nearly always have at least one healthy X chromosome and therefore rarely exhibit DMD symptoms. Individuals who carry one healthy and one mutated version of the same gene are called carriers because they can transmit the mutated gene to their children without being affected themselves. If boys inherit a mutated X chromosome from their mother, they must develop DMD because they lack a healthy X chromosome to compensate for the mutated gene; boys cannot be carriers of X chromosome–linked disorders.

In this clinical case, we have a family of phenotypically normal parents and their three children: one son and a set of female identical twins. Identical twins arise from a single fertilized egg and therefore have identical genes. All three children appeared healthy from birth to adolescence. At age 16, one of the girls began displaying the classic clinical symptoms for DMD: muscle weakness and frequent tripping and falling while walking or running. The pediatrician did not suspect DMD since the patient was a girl. As the disease progressed, the affected sister was tested, and the diagnosis was confirmed as DMD. Genetic analysis of blood samples and skin biopsies confirmed that both girls were carriers for DMD. Everyone was puzzled by the test results. Why was only one sister affected by the mutation?

The most likely reason for this outcome is that X chromosome inactivation differed between the two girls. During early embryonic development, Xist RNA randomly inactivated one of the X chromosomes in each cell to form condensed, inactive Barr bodies in both girls. Note that either the functioning or the mutated X chromosome had an equal chance of being inactivated in each cell. In one girl, the mutated X chromosome was inactivated in most of her cells, while the opposite occurred in her sister. Determining what causes Xist RNA to select which chromosome to silence and which to leave unaffected remains a significant challenge. Despite recent progress, how Xist RNA localizes and interacts with the X chromosome is still not completely understood.

Study Questions

1. If we assume that the DMD mutation did not spontaneously occur in both twins, explain why the mother, but not the father, must be a carrier of the DMD mutation.
2. In addition to being coated with Xist RNA, what other changes occur on the inactivated X chromosome to ensure it remains compact and transcriptionally inactive? Search the internet for additional clues.
3. Given your answer to question 2, what are the challenges facing scientists who might try to develop a means of reversing X chromosome inactivation in DMD patients?

Despite this tremendous variation, scientists have arrived at some consensus as to how Xist functions. The *Xist* gene is on the X chromosome, and Xist transcripts bind to a protein in the X chromosome scaffold called scaffold attachment factor A (SAF-A), forming a "coat" that spreads out from the *Xist* gene until the entire X chromosome is covered. After SAF-A binds to Xist, SAF-A changes shape to attract additional proteins that further compact the chromosome.

Concept Check 2

Apply the five levels of DNA organization to the Library of Congress analogy: What would be reasonable approximations of these levels for books on a shelf? What happens in a library that distinguishes Level 4 and Level 5? What parts of the DNA organization do not readily fit into this analogy?

2.4 Cells Chemically Modify DNA and Its Scaffold to Control Packaging

Key Concepts

- In addition to changing the physical organization of DNA, cells control DNA packing by chemically modifying DNA and the proteins in the DNA scaffold, including chromatin.
- The best-known chemical modifications target Level 1 and 2 DNA organization.
- Modifications of Level 1 and 2 DNA organizations can impact higher-level organization as well, including silencing of DNA in regions of heterochromatin.
- Level 1 chemical modifications occur directly on the DNA double helix; the most common modification is the addition of a methyl group to deoxycytosines, which suppresses transcription.
- Level 2 modifications occur primarily on histones and can affect chromatin condensation, gene transcription, and nucleosome assembly.

DNA and the proteins responsible for packaging it into Levels 1–5 of organizations can be chemically modified to change their structure and function (see *Cells Chemically Modify Proteins to Control Their Shape and Function* in Chapter 3). This is an important way for cells to more carefully control which regions of a chromosome are available for sharing information and which are effectively closed off. (In our Library of Congress analogy, this is equivalent to prominently displaying some texts while placing others in storage.)

Chemical Modifications at Level 1 and Level 2 Can Affect DNA Packing Across All Levels of DNA Organization

One of the most well-studied effects of these modifications is the formation of heterochromatin (Level 5 packaging) and subsequent repression of gene expression, often called **gene silencing**. This can occur by at least two mechanisms. In the first mechanism, shown in **FIGURE 2-22**, methyl groups are

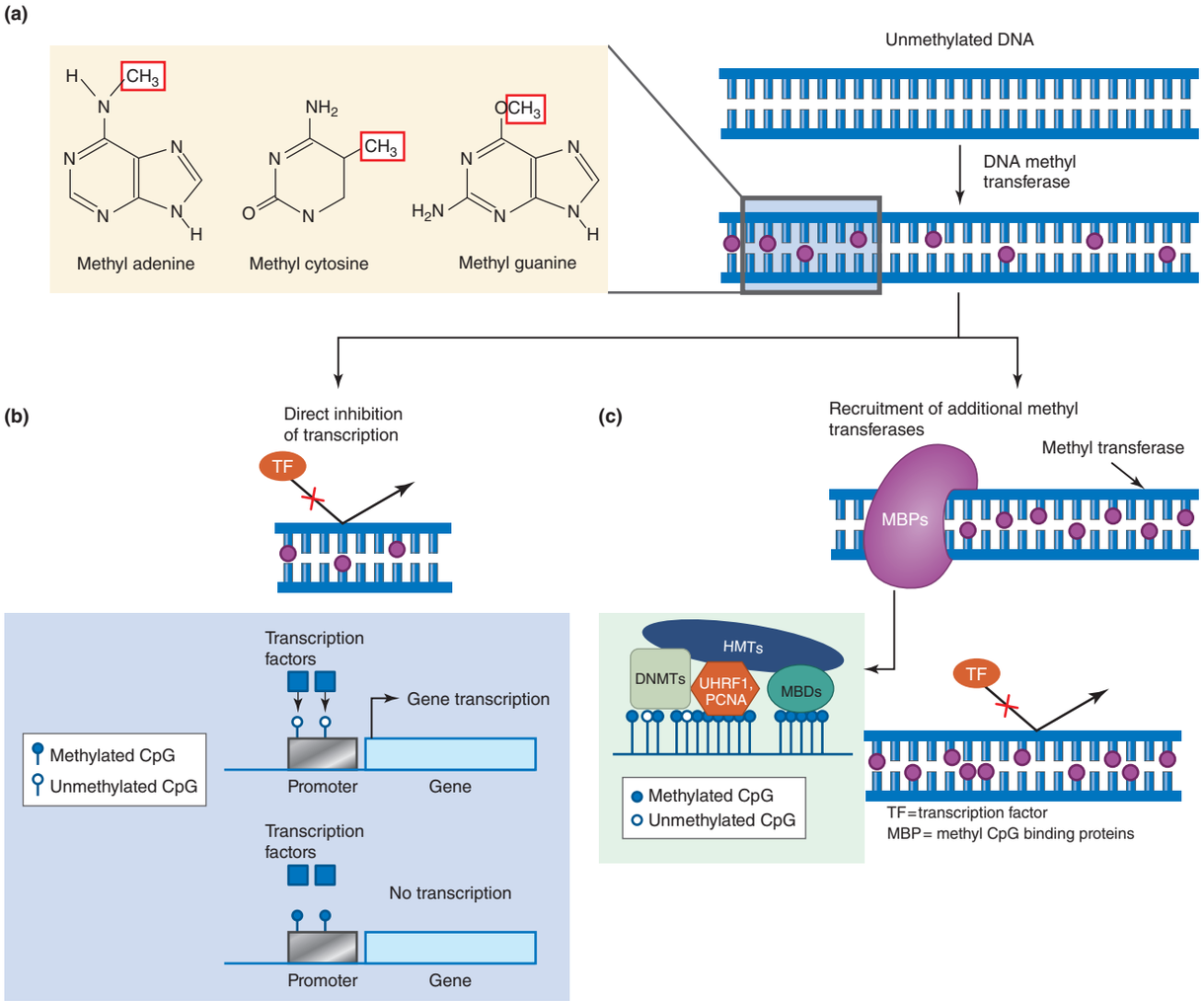


FIGURE 2-22 Methylation of DNA triggers gene silencing.

added directly to the bases adenine (in prokaryotes) or cytosine and guanine (in eukaryotes) in the DNA, a Level 1 process called **DNA methylation**. In mammals, these modifications occur most commonly on deoxycytosines that are adjacent to deoxyguanosines in a DNA strand. This is often abbreviated as CpG, with the p representing the phosphodiester bond holding the two deoxynucleotides together. (Regions of DNA that have a high proportion of CpG sequences are often called CpG islands.) DNA methylation silences gene expression because it directly prevents the binding of proteins that are required for transcription to take place (see *Transcription Factors Promote the Expression of Genes* in Chapter 12).

A second common way to silence genes is to modify histone proteins in the nucleosome. A variety of proteins are capable of attaching relatively small molecules (methyl groups, acetyl groups, or phosphate groups) to the tails of histones, which causes them to change their shape, as shown in **FIGURE 2-23**. This change in shape alters the function of the proteins as well (see *Cells Chemically Modify Proteins to Control Their Shape and Function* in Chapter 3). **FIGURE 2-24** shows an example of how histone modification can silence genes. An enzyme called histone deacetylase (HDAC) removes an acetyl group from histone H3;

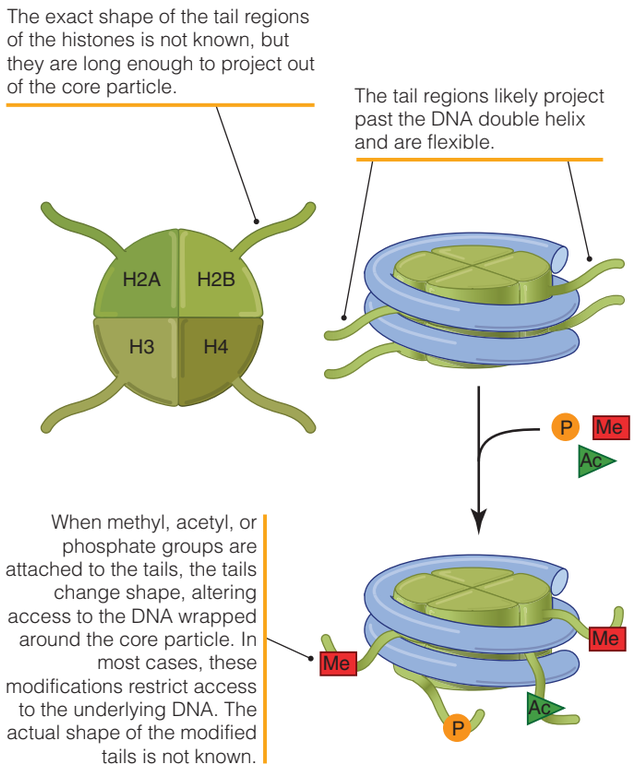


FIGURE 2-23 Chemical modification of histone tails changes the shape of chromatosomes.

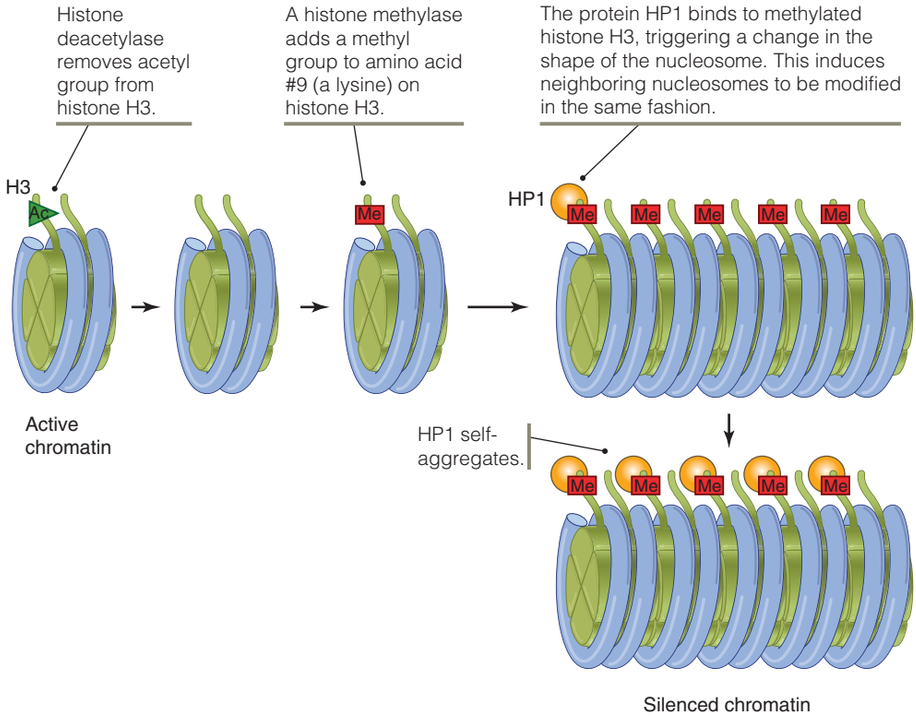


FIGURE 2-24 Histone modification can silence DNA to form heterochromatin.

histone acetylation often activates transcription, so removal of the acetyl group partially inhibits translation. Following this, a protein called histone methyltransferase attaches a methyl group to the ninth amino acid (a lysine, abbreviated K) on histone H3 (this is sometimes abbreviated as H3K9me). Finally, a third protein (the names vary in different species; in mammals, it is called heterochromatin protein 1, or HP1) attaches to the newly methylated histone H3. When this histone is deacetylated and methylated, its shape changes. This triggers a conformational change in the entire core particle, such that the DNA attached to the core particle can no longer be transcribed; it is now silenced. As additional nucleosomes next to the newly silenced section undergo the same modifications, the silencing can extend to larger stretches of DNA.

Modification of a chromatosome can help ensure that a region of DNA is not accessible. This is shown in **FIGURE 2-25**, when a protein called Rap1 binds to a portion of DNA. The resulting change in Rap1's shape allows two additional proteins named Sir3 and Sir4 to attach to it, and they in turn bind to histones H3 and H4. This creates a change in the shape of the core particle and facilitates the binding of additional Sir3/Sir4 complexes to adjacent nucleosomes. Note that in both strategies, changing the structure of the nucleosome core particle by altering the configuration of histones is a key step.

Some Regions of Eukaryotic Chromosomes Are Always Silenced

In eukaryotes, portions of the chromosomes are never active and are, therefore, called constitutive heterochromatin (in biology, *constitutive* means "constantly produced"). A minimal amount of transcription, such as for the siRNAs mentioned previously, takes place in these regions, but none of the resulting RNA transcripts is ever translated. Instead, these regions play important roles in maintaining the structure and organization of a chromosome during mitosis. For example, the centromere region of the chromosome

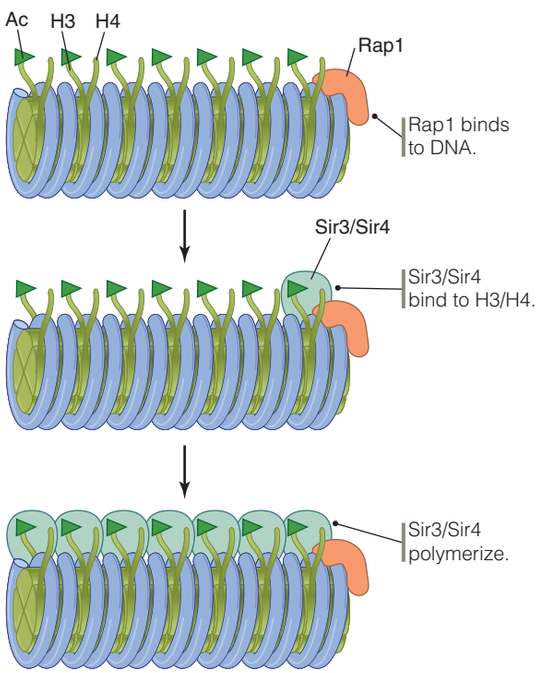


FIGURE 2-25 Rap1, Sir3, and Sir4 can silence DNA to form heterochromatin.

is essential for proper attachment of the chromosome to the microtubule spindle during mitosis, and the telomere regions protect the ends of the chromosome from damage. Both of these regions bind to many different proteins but only unwind completely during DNA replication. They are discussed in greater detail in Chapter 7.

Recently, scientists discovered another ATP-dependent group of proteins that play a crucial role in chromosome condensation prior to mitosis. These proteins are responsible for reading highly repetitive sequences of DNA to synthesize siRNAs. This finding came as something of a surprise because until then most researchers believed that siRNAs only controlled gene transcription. While the exact mechanism has yet to be determined, the siRNA-forming proteins recruit and bind to still another group of proteins that modify histones. This, too, is an essential step in heterochromatin formation.

Concept Check 3

This section discusses altering the physical form of DNA as a means of controlling access to genetic information. A common expression is that humans are currently living in an age of information overload. Are these two ideas at all related? Can cells suffer from information overload? What would be the consequences if they did? How do cells control access to this information without being overwhelmed? Use the vocabulary in this section to answer these questions.

2.5 Chapter Summary

To remain alive, cells must do two things: respond appropriately to external signals and internal programs and maintain their internal environment. Nearly all of the molecules responsible for these activities are either RNA or proteins, and the instructions for creating these molecules are stored in a relatively simple polymer, DNA. Each time a cell divides, the daughter cells inherit a copy of the parental cell DNA, which is slightly modified in each successive generation of cells by mistakes (mutations) made during the replication process. The instructions in DNA are organized into units called genes; cells read these genes to produce several types of RNAs (by transcription) and proteins (by translation of messenger RNA). Mutated genes produce altered RNAs and proteins when they are transcribed and translated, and it is these differences in RNAs and proteins that yield the variation in cellular phenotype that is acted upon by natural selection in each generation of cells and organisms. DNA is, therefore, the heritable material acted on by evolution.

Because DNA encodes the instructions for producing all of the RNAs and proteins a cell will require during its lifetime (and may also encode additional unused genes), it is typically an enormous molecule relative to the cell that harbors it. The complete DNA molecule, called a chromosome, is made up of combinations of four subunits, called deoxyribonucleotides, which form two antiparallel strands held together by hydrogen bonds. One of these two strands contains the coding sequence of a gene. To ensure that the genes can be easily accessed while also compacting them enough to fit into a cell, DNA is supported by an elaborate protein/RNA scaffold, called genophore in prokaryotes and chromatin in eukaryotes. Histone proteins are at the heart of these

scaffolds, and DNA wraps around "spools" of histones. Chemical modifications of histones and DNA bases play an important role in controlling which sections of DNA molecules are read by the transcription machinery. In some eukaryotes, portions of the chromosomes are condensed and modified so that they are not capable of undergoing transcription. These regions are called heterochromatin, to distinguish them from the transcriptionally accessible regions called euchromatin. Prokaryotes do not form heterochromatin.

Chapter Study Questions

1. How many phosphoester bonds, phosphodiester bonds, and hydrogen bonds are shown in Figure 2-12a? Do not count bonds formed at the ends of the strands with atoms not shown.
2. Explain how mutations in a cell's DNA can impact the proteins made by that cell's progeny (i.e., daughter cells).
3. Propose a model for how/why DNA binding proteins can "read" DNA sequences in double-stranded DNA.
4. During heterochromatin formation, how is DNA "compaction" different from DNA "silencing"?
5. Briefly describe the functions of the protein scaffold in DNA organization.
6. Why is it important that the two strands of the DNA double helix are held together by hydrogen bonds instead of (much stronger) covalent bonds?
7. Have a careful look at the structure of the core particle in Figures 2-14, 2-15, and 2-16. Based on our discussion of the function of this structure, explain three advantages provided by its molecular composition and shape.
8. If the phenotype of the cell is a product of the genes it expresses at a given time, why don't cells simply use a binary, on-off mechanism for controlling gene expression? In other words, why do eukaryotic cells use at least five levels of DNA organization instead of two?
9. Explain why DNA must always be synthesized in the 5′ to 3′ direction. Why is this important for current DNA sequencing technologies?
10. Based on your understanding of DNA structure, function, and replication, propose an explanation for why genomes of organisms tend to grow in size rather than shrink or remain constant over evolutionary time. How do organisms manage the useless genes they inherit from their ancestors?

Multiple-Choice Questions

1. What is a gene?
   - A. A sequence of nucleotides that wraps around a histone "spool."
   - B. A sequence of nucleotides that includes the coding sequence for an RNA molecule, plus the sequences that control the timing and amount of RNAs generated from the coding sequence.
   - C. A sequence of nucleotides that encodes all of the polypeptides necessary to form a protein with quaternary structure.
   - D. A sequence of nucleotides that binds to proteins.
   - E. A sequence of nucleotides that encodes the alpha helix portion of each polypeptide.

2. A common assay (method) for isolating DNA binding proteins from cells is called chromatin immunoprecipitation (ChIP). The product of this experimental method is often a list of proteins and the nucleotide sequences (200–1,000 base pairs) attached to them. Which one of the following questions could be answered using a ChIP assay?
   A. How many cells are in this sample?
   B. How many nucleosomes does this sample of DNA have in it?
   C. How big/long are the chromosomes in this sample?
   D. Which DNA binding proteins bind most tightly to DNA?
   E. Which nucleotide sequences do these proteins bind to?

3. Which statement best explains how natural selection promotes changes in DNA sequences in successive generations of cells?
   A. Natural selection increases the mutation rate in cells, thereby increasing the genetic diversity of each successive cell generation. This diversity is essential for ensuring that the optimal cell phenotype will emerge in each generation. Therefore, natural selection directly alters DNA sequences.
   B. Natural selection determines the size and shape of histone proteins. These, in turn, determine the shape of nucleosomes and have a direct impact on the packing state of DNA in cells. When histone composition of nucleosomes changes, the DNA sequences bound to these nucleosomes change as well.
   C. Natural selection increases the rate of histone modification, thereby giving rise to a more diverse range of DNA packing in cells. This increased diversity, in turn, increases the chances that DNA will be damaged during replication, giving rise to changes in DNA sequences in successive generations of cells.
   D. Natural selection increases the probability of survival in cells that express proteins best suited to allow cells to prosper in their environment. Because a cell's external environment is never static, natural selection favors slightly different cells in successive generations. One of the major factors determining the phenotype of cells is the structure and function of the proteins they express, and these are directly influenced by the sequence of DNA encoding their genes.
   E. Natural selection reflects a cell's relative ability to overcome deleterious mutations by repairing them and allowing cells to reproduce. Thus, natural selection promotes changes in the DNA in successive generations of cells by favoring cells with the best DNA repair mechanisms; eventually, future generations of cells will be able to repair virtually all DNA damage.

4. Assume you analyze the hemoglobin gene of a sickle-cell patient and an unaffected (control) individual and find that both have single base-pair mutations in this gene. Which statement best explains why the mutation in the control individual is not causing sickle-cell disease?
   A. The mutation in the sickle-cell patient occurs in a sequence of DNA that encodes the amino acid sequence of the hemoglobin protein; the mutation in the unaffected individual occurs in a region of the hemoglobin gene that does not.
   B. The mutation in the sickle-cell patient was inherited from her parents; the mutation in the unaffected individual arose spontaneously when the individual was a child.

C. The mutation in the sickle-cell patient is in a region of DNA wrapped around a nucleosome core particle; the mutation in the unaffected individual is not.
D. The mutation in the sickle-cell patient is caused by deletion of a single nucleotide in the gene sequence; the mutation in the unaffected individual is caused by the addition of a single nucleotide in the gene sequence.
E. The mutation in the sickle-cell patient is found in euchromatin; the mutation in the unaffected individual is found in heterochromatin.

5. All nucleotides are composed of _____.
   A. a base, DNA, and amino acids
   B. a base, a sugar, and a scaffold
   C. a base, a sugar, and a phosphate group
   D. mRNA, rRNA, and tRNA
   E. mRNA, rRNA, tRNA, siRNA, and miRNA

6. Choose the best answer that differentiates a nucleosome from a chromatosome.
   A. A nucleosome is located only in the nucleus; a chromatosome is located in the nucleus, mitochondria, and chloroplasts.
   B. A nucleosome consists of 167 base pairs of double-stranded DNA wrapped around eight histones; a chromatosome consists of a nucleosome plus linker DNA and a linker histone.
   C. A chromatosome is identical to a nucleosome, except a chromatosome lacks a linker histone.
   D. A nucleosome is a 167-base-pair-long double-stranded DNA with a linker histone.
   E. A and B are both correct.

7. In double-stranded DNA, which kinds of bonds hold one complementary strand to the other?
   A. Hydrogen bonds
   B. Covalent bonds
   C. Phosphodiester bonds
   D. Ionic bonds
   E. Hydrophobic and hydrophilic bonds

8. Which one of the following statements is true?
   A. The deoxynucleotides comprising both strands in DNA are oriented unidirectionally, in a parallel and complementary fashion, with deoxynucleotides on opposite strands bound by double or triple hydrogen bonds.
   B. The building blocks of DNA are called deoxynucleotides because they lack oxygen on their 5′ end when compared to building blocks of RNA.
   C. DNA is composed of two strands of amino acid monomers organized antiparallel to each other, with one strand oriented 3′ to 5′, while the other strand is oriented 5′ to 3′.
   D. DNA is a macromolecule made up of a nitrogenous base and a phosphate backbone, held together by hydrogen bonds between deoxyribose sugars.
   E. None of the above statements is correct.

9. During dinucleotide polymerization in DNA, RNA, and oligonucleotides, covalent bonds are formed between neighboring nucleotides. Which carbons form these bonds?
   A. 5′ carbon on the sugars and 3′ carbon on the nitrogenous bases
   B. 1′ carbon and 2′ carbon on the sugars
   C. 3′ carbon and 5′ carbon of the sugars, linked via a phosphate group
   D. 3′ carbon and 5′ carbon of the sugars in DNA and 2′ carbon and 5′ carbon of the sugars in RNA
   E. 3′ carbon of the sugars and a phosphorus atom
10. Under conditions of high relative humidity, DNA adopts the _____ configuration.
   A. Z
   B. 3′
   C. B
   D. 5′
   E. dideoxy

References

Heather JM, Chain B. (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107(1): 1–8. doi:[10.1016/j.ygeno.2015.11.003]

Panda D, Molla KA, Baig MJ, Swain A, Behera D, Dash M. (2018) DNA as a digital information storage device: hope or hype? 3 Biotech. 8(5): 239. doi: 10.1007/s13205-018-1246-7

Scherrer K. (2018) Primary transcripts: from the discovery of RNA processing to current concepts of gene expression—Review. Exp Cell Res 2018 September 25. pii: S0014-4827(18)30948-0. doi: 10.1016/j.yexcr.2018.09.011

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, Waterston RH. (2017) DNA sequencing at 40: past, present and future. Nature 550(7676): 345–353. doi: 10.1038/nature24286

Allshire RC, Madhani HD. (2018) Ten principles of heterochromatin formation and function. Nat Rev Mol Cell Biol. 2018 Apr;19(4):229–244. doi: 10.1038/nrm.2017.119