CHAPTER FOUR

Genes, Genomes and DNA

History of DNA as the Genetic Material

How Much Genetic Information Is Necessary to Maintain Life?

Non-Coding DNA

Coding DNA May Be Present within Non-coding DNA

Repeated Sequences Are a Feature of DNA in Higher Organisms

Satellite DNA Is Non-coding DNA in the Form of Tandem Repeats

Minisatellites and VNTRs

Origin of Selfish DNA and Junk DNA

Palindromes, Inverted Repeats and Stem and Loop Structures

Multiple A-Tracts Cause DNA to Bend

Supercoiling is Necessary for Packaging of Bacterial DNA

Topoisomerases and DNA Gyrase

Catenated and Knotted DNA Must Be Corrected

Local Supercoiling

Supercoiling Affects DNA Structure

Alternative Helical Structures of DNA Occur

Histones Package DNA in Eukaryotes

Further Levels of DNA Packaging in Eukaryotes

Melting Separates DNA Strands; Cooling Anneals Them

History of DNA as the Genetic Material

Despite their complexity, living organisms obey the laws of chemistry.

Until early in the nineteenth century, it was believed that living matter was quite different from inanimate matter and was not subject to the normal laws of chemistry. In other words, organisms were thought to be made from chemical components unique to living creatures. Furthermore, there was supposedly a special vital force that mysteriously energized living creatures. Then, in 1828, Friedrich Wohler demonstrated the conversion in a test tube of ammonium cyanate, a laboratory chemical, to urea, a "living" molecule also generated by animals. This was the first demonstration that there was nothing magical about the chemistry of living matter.

Further experiments showed that the molecules found in living organisms were often very large and complex. Consequently, their complete chemical analysis was time consuming and is indeed, still continuing today. The de-mystification of life chemistry reached its peak in the 1930s when the Russian biochemist Alexander Oparin wrote a book outlining his proposal for the chemical origin of life. Although the nature of the genetic material was still unknown, Oparin put forward the idea that life, with its complex molecular composition, evolved from small molecules in the primeval ocean as a result of standard physical and chemical forces (see Ch. 20).

Avery found that purified DNA could carry genetic information from one strain of bacterium to another. This revealed that DNA was the genetic material.

Until the time of World War II, the chemical nature of the *inherited genetic information* remained very vague and elusive. DNA was actually discovered in 1869 by Frederich Miescher, who extracted it from the pus from infected wounds! However, it was nearly a century before its true significance was revealed by Oswald Avery. In 1944, Avery found that the virulent nature of some strains of bacteria that caused pneumonia could be transmitted to related harmless strains by a chemical extract. Avery purified the essential molecule and demonstrated that it was DNA, although he did not use the name "DNA," since its structure was then uncharacterized. When DNA from virulent strains was added to harmless strains, some took up the DNA and were "transformed" into virulent strains (see Ch. 18 for the mechanism of transformation). Avery concluded that the genes were made of DNA and that somehow genetic information was encoded in this molecule. Since DNA was known to have only half a dozen components, it had not been a leading competitor for the role of genetic material; it was viewed as too simple to encode the information for a living creature!

The question of how DNA, with only half a dozen components, could act as the genetic information was answered by James Watson and Francis Crick in 1953. Their now famous double helix provided a chemical basis for the genetic code and suggested

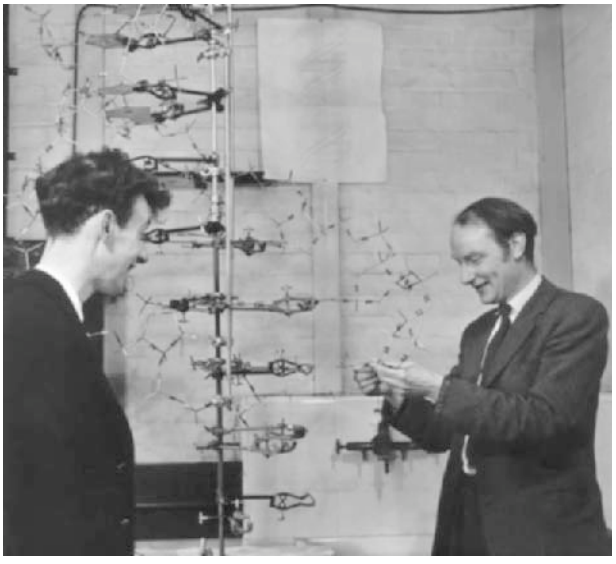


FIGURE 4.01 ***Watson and Crick in the 1950s***

James Watson (b.1928) at left and Francis Crick (b.1916), with their model of part of a DNA molecule in 1953. Courtesy of: A. Barrington Brown, Science Photo Library.

X-ray diffraction showed that two strands of DNA are twisted together forming a double helix.

a mechanism for DNA replication. In 1950 Maurice Wilkins and his assistant Raymond Gosling took the first images of DNA using X-ray diffraction. Gosling's work was continued by Rosalind Franklin who joined Wilkins' group the following year. Watson and Crick used a X-ray diffraction picture taken by Rosalind Franklin and Raymond Gosling in 1952 as the basis for their structural model. Rosalind Franklin died in 1958 of cancer aged 37, probably due to the effects of the X-rays. Unraveling the chemical basis for inheritance won Watson, Crick and Wilkins the Nobel Prize in Physiology or Medicine for 1962 "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".

This central finding underlies our whole understanding of how living cells operate and what life means. Since the discovery of the double helix, the genetic code has been worked out, and starting in 1995 with the bacterium *Haemophilus influenzae*, the DNA of a variety of organisms has been totally sequenced. As the third millennium begins, the human genome has been sequenced, but researchers are still working to assemble the data into complete contiguous sequences for each chromosome. This chapter will discuss how much genetic information is needed to operate a living cell and how that information is arranged on the DNA. Much of the information about these processes comes from studying bacteria, but the information frequently applies also to eukaryotes.

The Double Helix by James D. Watson Published in 1968 by Atheneum, New York

This book gives a personal account of the greatest biological advance of the 20th century—the unraveling of the structure of the DNA double helix by James Watson and Francis Crick. Like the bases of DNA, Watson and Crick formed a complementary pair. Crick, a physicist with an annoying laugh, was supposed to be working towards a Ph.D. on protein X-ray crystallography. Watson was a homeless American biologist, wandering around Europe with a post-doctoral fellowship, looking for something to do.

Despite spending much time carousing, the intrepid heroes, Crick and Watson, beat their elders to the finish line. Watson describes with relish how the great American chemist, Linus Pauling, placed the phosphate backbone of DNA down the middle, so failing to solve the structure. The data proving the phosphate backbone was on the outside of the double helix came from Rosalind Franklin, an X-ray crystallographer at London University. Of her Watson says, ". . . the best home for a feminist was in another person's lab."

The Director of the Cavendish Laboratory at Cambridge was Sir William Bragg, the august inventor of X-ray crystallography. Despite being depicted as a stuffy has-been who nearly threw Crick out for loud-mouthed insubordination, Bragg wrote the foreword to the book. After all, when younger scientists under your direction make the greatest discovery of the century, it is no time to bear a grudge!

The biographies of great scientists are usually exceedingly dull. Who cares, after all, what Darwin liked for breakfast or what size shoes Mendel wore? It is their discoveries, and how they changed the world, that are fascinating. "The Double Helix" is different. Biographers are generally minor figures, understandably hesitant to criticize major achievers. Watson, himself a big name, happily lacks such respect, and cheerfully castigates other top scientists. It is this honest portrayal of the flaws and fantasies of those involved in unraveling the DNA double helix that keeps the readers attention.

If your stomach can't stand any more sagas about caring investigators who work on into the early hours hoping that their discoveries will help sick children, this book is for you. Like most candid scientists, Watson and Crick did not work for the betterment of mankind; they did it for fun.

How Much Genetic Information Is Necessary to Maintain Life?

Bacteria typically carry all of their genes on a single circular chromosome. Occasional species of bacteria have two or more different chromosomes and a few bacteria even contain linear chromosomes (see Ch. 18). Higher organisms have linear chromosomes and the number ranges from a handful to over a thousand in a few flowering plants.

The simplest living cell probably needs around 200–300 genes.

The smallest eubacterial genome is that of *Mycoplasma genitalium*, which has one circular chromosome, consisting of 580,000 base pairs (**bp**) of DNA. Since the average gene is about 1,000 base pairs long, and there is little non-coding DNA between bacterial genes, *M. genitalium* should have approximately 500 genes. The precise number is disputed and ranges from 468 to 517. Extensive analysis of mutations suggests that around 300 of these genes are essential for the growth and reproduction of *M. genitalium*. Comparison with other small bacterial genomes suggests that around 250 genes may be the minimum essential for a living cell. (This estimate also takes into account those pathways needed for synthesis of all vital cell components, some of which are missing in *M. genitalium* because it is a parasitic bacterium.) The smallest prokaryotic genome belongs to *Nanoarchaeum equitans*, a marine archaebacterium that was discovered in 2002. *N. equitans* has about 15% less DNA than *M. genitalium* and may also be a parasite, as it cannot grow unless attached to the surface of other microorganisms. Despite having less DNA its genes are more closely spaced and in consequence *N. equitans* actually has more coding sequences than *M. genitalium*—approximately 550.

Most bacteria have a few thousand genes.

Although parasitic bacteria may have less than 1,000 genes, most free-living bacteria have 2,000 to 4,000 genes. Occasional bacteria with complex life cycles, such as *Myxococcus*, may have 9,000 to 10,000 genes. Free-living eukaryotes typically have from 6,000 to 50,000 genes (see Table 4.01). However, the parasitic eukaryote *Encephalitozoon cuniculi* (Protozoa, Microspora) has only 2.9 million base pairs (**Mbp**) of DNA, implying that it possesses no more than 3,000 genes—less than many bacteria.

Non-Coding DNA

The number of genes ranges from roughly 500 to 50,000—a 100-fold range. In contrast, the amount of DNA ranges from 0.5 Mbp to nearly 50,000 Mbp—a 100,000-fold range. This discrepancy is due to **non-coding DNA** in eukaryotes with larger DNA content. This, as its name indicates, is DNA whose base sequence is largely meaningless and does not encode useful genetic information, at least as far as we know currently.

Some regions of DNA contain useful genetic information, other regions do not.

In addition to segments of DNA that give rise to gene products (i.e., protein or RNA), DNA molecules contain many other regions, including both regulatory sites and non-coding regions. Any segment of DNA, whether coding or not, can be referred to as a **locus (plural, loci)**; that is, a location on a chromosome (or other molecule of DNA). Since any DNA sequence can occur in alternative versions, the term **allele** is used for these even if the DNA in question is non-coding.

Non-coding DNA accounts for the majority of the DNA in most higher animals and plants.

Although bacteria have relatively little non-coding DNA, eukaryotes have significant amounts. Even a relatively primitive eukaryote, such as yeast, has nearly 50 percent non-coding DNA. For example, yeast has about three times as much DNA as *E. coli* but only 1.5 times as many genes. Higher eukaryotes have even greater proportions of non-coding DNA. Mammals such as mice and men have an estimated

allele One particular version of a gene, or more broadly, a particular version of any locus on a molecule of DNA
bp Abbreviation for base pair(s)
locus (plural, loci) A place or location on a chromosome; it may be a genuine gene or just any site with variations in the DNA sequence that can be detected, like RFLPs or VNTRs
Mbp Megabase pairs or million base pairs
non-coding DNA DNA that does not code for proteins or functional RNA molecules

TABLE 4.01 Genome Sizes

Organism	Number of Genes	Amount of DNA (bp)	Number of Chromosomes
Viruses			
Bacteriophage MS2	4	3,600	1 (ssRNA)*
Tobacco Mosaic Virus	4	6,400	1 (ssRNA)*
ΦX174 bacteriophage	11	5,387	1 (ssDNA)
Influenza	12	13,500	8 (ssRNA)
T4 bacteriophage	200	165,000	1
Poxvirus	300	187,000	1
Bacteriophage G	680	498,000	1
Prokaryotes			
Mitochondrion (human)	37	16,569	1
Mitochondrion (*Arabidopsis*)	57	366,923	1
Chloroplast (*Arabidopsis*)	128	154,478	1
Nanoarchaeum equitans	550	490,000	1
Mycoplasma genitalium	480	580,000	1
Methanococcus	1,500	1.7 Mbp	1
Escherichia coli	4,000	4.6 Mbp	1
Myxococcus	9,000	9.5 Mbp	1
Eukaryotes (haploid genome)			
Encephalitozoon	2,000	2.5 Mbp	11
Saccharomyces	5,700	12.5 Mbp	16
Caenorhabditis	19,000	100 Mbp	6
Drosophila	12,000	140 Mbp	5
Homo sapiens	25,000	3,300 Mbp	23
Arabidopsis	25,000	115 Mbp	5
Oryza sativa (Rice)	45,000	430 Mbp	12

*ssRNA = single stranded RNA; ssDNA = single stranded DNA; all other genomes consist of double stranded DNA.

20,000 to 30,000 genes carried on a total of 300 Mbp of DNA. This means that just over 85 percent is non-coding. However, flowering plants, which are estimated to have roughly as many genes as mammals, possess 100-fold more DNA. Some amphibians, such as frogs and newts, possess almost as much.

In prokaryotes, almost all the non-coding DNA is found between genes as **intergenic DNA**. In eukaryotes, the situation is more complicated. Not only is non-coding DNA scattered throughout the eukaryotic chromosomes between the genes, but the actual genes themselves are often interrupted with non-coding DNA. These **intervening sequences** are known as **introns**, whereas the regions of the DNA that contain coding information are known as **exons**. Most eukaryotic genes consist of exons alternating with introns (Fig. 4.02).

In eukaryotes, the genes themselves are often interrupted by stretches of non-coding DNA.

In lower, single-celled eukaryotes, such as yeast, introns are relatively rare and often quite short. In contrast, in higher eukaryotes, most genes have introns and they

exon Segment of a gene that codes for protein and that is still present in the messenger RNA after processing is complete
intergenic DNA Non-coding DNA that lies between genes
intervening sequence An alternative name for an intron
intron Segment of a gene that does not code for protein but is transcribed and forms part of the primary transcript

FIGURE 4.02 ***Intervening Sequences Interrupt Eukaryotic Genes***

Regions of non-coding DNA between genes are called intergenic DNA. Non-coding regions that interrupt the coding regions of genes are called introns.

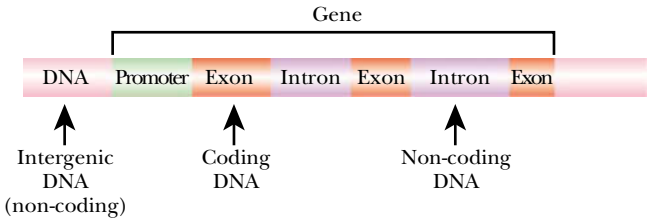
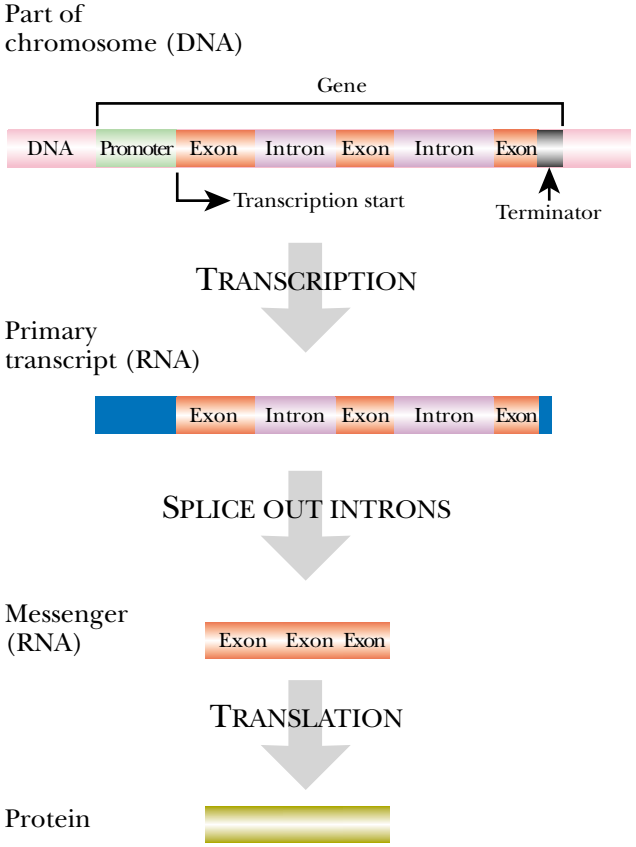


FIGURE 4.03 ***Removal of Introns Before Protein Synthesis***

A gene on the chromosomal DNA, consisting of a promoter, introns and exons, is transcribed to give the primary transcript, an RNA molecule containing both the introns and exons. Processing of the primary transcript removes the introns to leave an mRNA carrying only exons. The protein reflects only the information in the exons.



Genes that are interrupted must have the non-coding regions removed when messenger RNA is made.

are often longer than the exons. In some genes, the introns may occupy 90 percent or more of the DNA. For example, the mutated gene causing cystic fibrosis was found to occupy 250,000 base pairs and have 24 exons, which encoded a protein of 1,480 amino acids. Since 1,480 amino acids need only 4,440 base pairs to encode them, this means that scarcely 2 percent of the cystic fibrosis gene is actual coding DNA. The rest consists of intervening sequences—23 introns.

In order to synthesize the protein encoded by an interrupted gene, the introns must be removed at some stage. Splicing out of introns is accomplished after transcription at the mRNA stage within the nucleus. When a gene is expressed, the DNA is first transcribed to give a long RNA molecule, known as the **primary transcript**, that includes the introns. The primary transcript is then processed to remove the introns, yielding the mRNA (Fig. 4.03). Because this chapter is concerned with genome structure, the details of intron extraction will be deferred until Chapter 12.

Coding DNA May Be Present within Non-coding DNA

Introns are not totally absent from prokaryotes, but they are extremely rare. Moreover, there is usually only a single intron in a gene, unlike in eukaryotes where many

primary transcript The original RNA molecule obtained by transcription from a DNA template, before any processing or modification has occurred

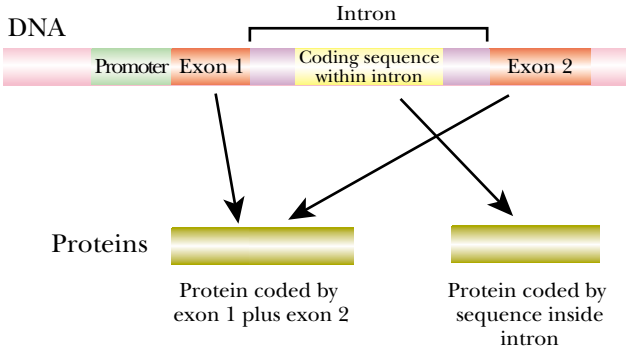


FIGURE 4.04 ***Intron Containing a Coding Sequence of Its Own***

An interesting situation can arise when a coding sequence has an intron within it that itself includes a coding sequence for another protein.

genes have multiple introns. Furthermore, most known examples are within the genes of bacterial viruses, rather than the chromosomal genes of bacteria themselves. For example, bacteriophage T4 possesses several introns, including one each in the genes encoding thymidylate synthase and ribonucleotide reductase. The T4 introns are homologous to the self-splicing introns of lower eukaryotes (for splicing mechanisms, see Ch. 12). This family of introns takes the complexity one step further, as there is a coding sequence for a separate protein located entirely within the intron (Fig. 4.04). This protein is concerned with survival of the intron.

In those rare cases where chromosomal genes of prokaryotes are interrupted, the genes often encode RNA molecules rather than proteins. Both tRNA and rRNA genes have been found with introns in both the eubacteria and archaebacteria. For example, one of the leucine tRNA genes of cyanobacteria (blue-green photosynthetic bacteria) and the corresponding gene in the DNA of chloroplasts contain self-splicing introns inserted at equivalent positions.

Repeated Sequences Are a Feature of DNA in Higher Organisms

Repeated sequences are frequently found in the DNA of higher organisms.

Most genes are present only as single copies. Such unique sequences account for almost all bacterial DNA. However, in higher organisms, unique sequences may comprise as little as 20 percent of the total DNA. For example, humans have 65 percent unique DNA, whereas frogs have only 22 percent. The rest of the DNA is made up of **repeated sequences (or repetitive sequences)** of one kind or another. Repeated sequence are what their name suggests, DNA sequences that are repeated many times throughout the genome. In some cases the repeat sequences follow each other directly—tandem repeats (see below), whereas others are spread separately around the genome—interspersed sequences. Some repeated sequences are genuine genes, but the majority consist of non-coding DNA.

Individual members of a family of repeated sequences are rarely identical in every base. Nonetheless, one may imagine an ideal, so-called **consensus sequence**, from which they are all derived by only minor alterations (Fig. 4.05). Such a consensus sequence is deduced in practice by examining many related individual sequences and including those bases most often found at each position. In other words, consensus sequences are found by comparing many sequences and taking the average.

In addition to multiply repeated sequences, eukaryotic cells also possess **pseudogenes**. These are defective duplicate copies of genuine genes, whose defects prevent them from being expressed. Pseudogenes are present in only one or two copies and may be next to the original, functional version of the gene or may be far away, even on a different chromosome. In quantitative terms, pseudogenes account for only a

consensus sequence Idealized base sequence consisting of the bases most often found at each position.
pseudogene Defective copy of a genuine gene
repeated sequences DNA sequences that exist in multiple copies
repetitive sequences Same as repeated sequences

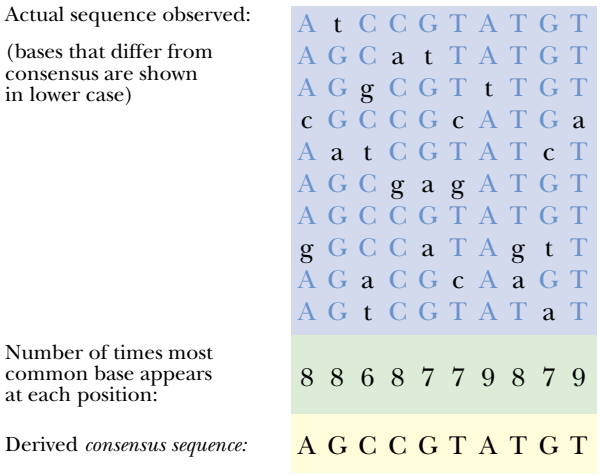


FIGURE 4.05 ***Deduction of Consensus Sequence***

The frequency of base appearances is used to derive a consensus sequence that is most representative of the series of related sequences shown.

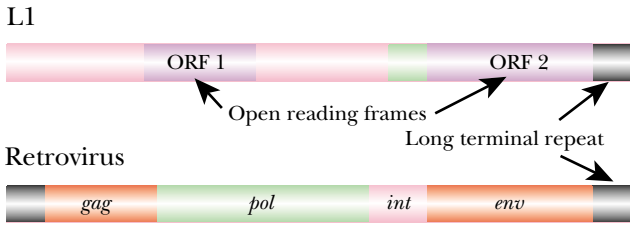


FIGURE 4.06 ***Structure of the LINE-1 Element***

An example of a LINE-1 or L1 element is shown. L1 contains blocks of DNA that show homology with the *pol* and LTR sequences of retroviruses, as well as two coding sequences or open reading frames (ORF1 and ORF2) involved in its own replication.

tiny fraction of the DNA. However, they are believed to be of great importance in molecular evolution as the precursors to new genes (see Ch. 20). Sometimes both copies of a duplicated gene remain functional and repeated duplication may even give families of related genes. The multiple copies gradually diverge to a greater or lesser extent as they adapt to carry out similar but related roles. Thus the repeated sequences due to a gene family are closely related but not absolutely identical.

Genes for ribosomal RNA are usually found in multiple copies. In higher organisms there may be thousands of copies.

Since each prokaryotic cell contains 10,000 or more ribosomes, it is not surprising that their DNA usually contains half a dozen copies of the genes for rRNA and tRNA. In the much larger eukaryotic cell, there are hundreds or thousands of copies of the rRNA and tRNA genes. Sequences present in hundreds or thousands of copies are referred to as **moderately repetitive sequences**. About 25 percent of human DNA falls into this category. This includes multiple copies of highly used genes, like those for rRNA as well as nonfunctional stretches of DNA that are repeated many times.

Much of the moderately repetitive non-coding DNA is formed of **LINEs**, which are "**Long INterspersed Elements**." They are thought to be derived from retrovirus-like ancestors. In mammalian genomes, there are 20,000–50,000 copies of the LINE-1 (L1) family (Fig. 4.06). A complete L1 element is around 7,000 bp and contains two coding sequences. However, most individual L1 elements are shorter and many contain sequence rearrangements that disrupt the coding sequences, rendering them nonfunctional.

About 7% of human DNA consists of repeats of the 300 bp Alu element.

Another 10 percent of human DNA consists of sequences present in hundreds of thousands to millions of copies. Much of this **highly repetitive DNA** consists of **SINEs**, or **Short INterspersed Elements**. These sequences are almost all nonfunctional as far

highly repetitive DNA DNA sequences that exist in hundreds of thousands of copies
LINE Long interspersed element
long interspersed element (LINE) Long sequence found in multiple copies that makes up much of the moderately repetitive DNA of mammals
moderately repetitive sequence DNA sequences that exist in thousands of copies (but less than a hundred thousand)
short interspersed element (SINE) Short sequence found in multiple copies that makes up much of the highly or moderately repetitive DNA of mammals
SINE Short interspersed element

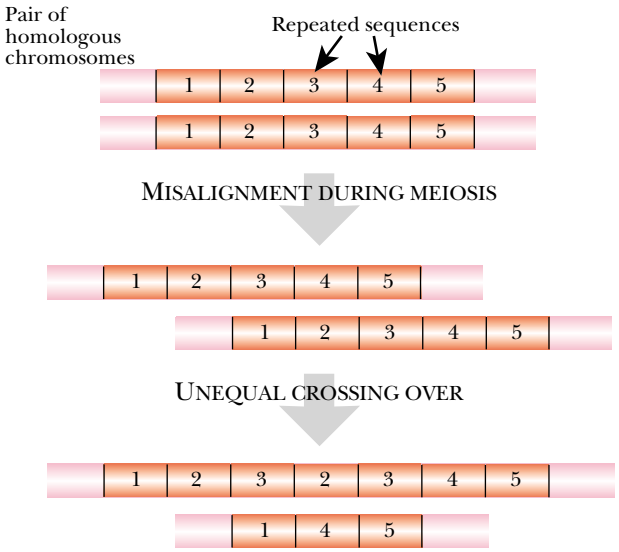


FIGURE 4.07 ***Unequal Crossover due to Misalignment***

A pair of homologous chromosomes contains repeated elements. Since repeated elements may be readily misaligned during meiosis, crossing over will sometimes occur in regions that are not comparable in each chromosome. The result is one longer and one shorter DNA fragment.

as is known. The best known SINE is the 300 base pair **Alu element**. It is named after the restriction enzyme *Alu I*, which cuts it at a single site 170 bp from the front of the sequence. (See Ch. 22 for restriction enzymes.) From 300,000 to 500,000 copies (per haploid genome) of the Alu element are scattered throughout human DNA. Though apparently useless, they make up 6 to 8 percent of a human's genetic information. They occur singly or in small clusters and the majority are mutated or incomplete.

Most mammals contain SINEs topographically related to the Alu element. However, the original sequence, as found in mice, hamsters, etc., is only 130 bp long. For example, the mouse contains about 50,000 copies of the Alu-related **B1 element**. The human Alu element possesses two tandem repeats of this ancestral 130 bp B1 sequence plus an extra, unrelated 31 bp insertion of obscure origin. The mouse has less than 100,000 copies of the B1 element, so it would be classified as moderately repetitive DNA, whereas humans have more than 100,000 copies of the related Alu element, which is therefore classified as highly repetitive DNA. Clearly, the division into "moderately" repetitive and "highly" repetitive DNA is somewhat arbitrary.

Satellite DNA Is Non-coding DNA in the Form of Tandem Repeats

Tandem repeats cluster together forming regions of inert satellite DNA.

Unlike the LINEs and SINEs, which by definition are scattered throughout the genome, a significant amount of highly repetitive DNA in eukaryotic cells is found as long clusters of **tandem repeats**. This is also known as **satellite DNA**. Tandem means that the repeated sequences are next to each other in the DNA without gaps between. The amount of satellite DNA is highly variable. In mammals such as the mouse, satellite DNA accounts for about 8 percent of the DNA, whereas in the fruit fly, *Drosophila*, it comprises nearly 50 percent.

Long series of tandem repeats tend to misalign when pairs of chromosomes line up for recombination during meiosis. **Unequal crossing over** will then produce one shorter and one longer segment of repetitive DNA (Fig. 4.07). Thus, the exact number of tandem repeats varies from individual to individual within the same population.

Alu element An example of a SINE, a particular short DNA sequence found in many copies on the chromosomes of humans and other primates
B1 element An example of a SINE found in mice; the precursor sequence from which the human Alu element evolved
satellite DNA Highly repetitive DNA of eukaryotic cells that is found as long clusters of tandem repeats and is permanently coiled tightly into heterochromatin
tandem repeats Repeated sequences of DNA (or RNA) that lie next to each other
unequal crossing over Crossing over in which the two segments that cross over are of different lengths; often due to misalignment during pairing of DNA strands

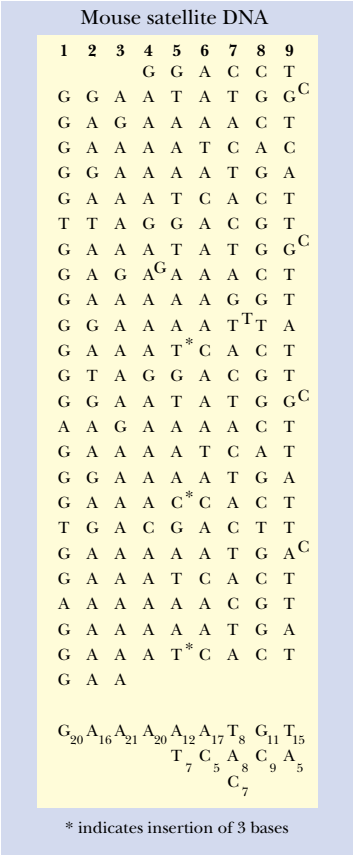


FIGURE 4.08 ***Repeating Motifs in Mouse Satellite DNA***

Variations in the consensus 9 bp satellite DNA sequence GAAAAATGT are shown.

Person to person variation in the overall length of short tandem repeats allows individual identification and is used in forensic analysis.

TABLE 4.02 Distribution of a 64 bp Human VNTR

% of population	# of repeats
7	18
11	16
43	14
36	13
4	10

In insects, the repeating sequences of satellite DNA are very short and consist of only one or very few different sequences. Thus in *Drosophila virilis* a 7 bp repeat with a consensus sequence of ACAAACT accounts for almost all of the satellite DNA. About half of the repeats have the consensus itself and the rest differ by one or, rarely, two bases. Satellite sequences vary considerably from one organism to another. The more commonly used *Drosophila melanogaster* has more complex satellite DNA that includes the 7 bp sequence just described as well as other 5, 10 and 12 bp repeats. In mammals, the satellite sequences are relatively complex. Although there is an overall 9 bp consensus in the mouse, there is much more variation among the repeats (Fig. 4.08).

Satellite DNA is inert and is permanently coiled tightly into what is known as **heterochromatin**. The heterochromatin is located around the **centromeres** of the chromosomes, suggesting that it serves some structural role. Note, however, that these satellite DNA sequences are quite distinct from the **centromere sequences**, which are needed for attachment of the spindle fibers during cell division.

Minisatellites and VNTRs

Segments of DNA consisting of short tandem repeats, but in much fewer copies than satellites, are known as **mini-satellites** or **VNTRs** (**Variable Number of Tandem Repeats**). Typically there may be from five to 50 tandem repeats in a VNTR. In mammals, VNTRs are common and are scattered over the genome, although they tend to be found close to the telomeres.

Due to unequal crossing over, the number of repeats in a given VNTR varies among individuals. Although VNTRs are non-coding DNA and not true genes, nonetheless the different versions are referred to as alleles. For example, Table 4.02 shows the distribution of one human VNTR of 64 bp among the population.

Some hyper-variable VNTRs may have as many as 1,000 different alleles and give unique patterns for almost every individual. This quantitative variation may be used for the identification of individuals by **DNA fingerprinting**.

Origin of Selfish DNA and Junk DNA

It is thought that most of the repetitive and non-coding DNA found in the chromosomes of eukaryotes is useless to the organism concerned. Such useless DNA is sometimes referred to as **junk DNA**. Where did it come from? It is thought that many

centromere Structure found on a chromosome and used to build and organize microtubules during mitosis
centromere sequence (CEN) A recognition sequence found at the centromere and needed for attachment of the spindle fibers
DNA fingerprint Individually unique pattern due to multiple bands of DNA produced using restriction enzymes, separated by electrophoresis and usually visualized by Southern blotting
heterochromatin Highly condensed form of chromatin that is genetically inert
junk DNA Defective selfish DNA that is of no use to the host cell it inhabits and which can no longer move or express its genes
mini-satellite Another term for a VNTR (variable number tandem repeats)
VNTR See variable number tandem repeats
variable number tandem repeats (VNTR) Cluster of tandemly repeated sequences in the DNA, whose number of repeats differs from one individual to another

DNA that consists of very large numbers of tandem repeats may well have a base composition different from that of the genome as a whole. If so, the satellite DNA will have a different buoyant density from the rest of the DNA, as this property depends on the base composition. DNA may be fractionated according to density by ultracentrifugation in a gradient of the heavy metal salt, cesium chloride (CsCl). Each fraction of DNA forms a band at the position corresponding to its own density. If the %GC varies by 5 percent or more, separate bands are obtained. When mouse DNA is run on a CsCl density gradient, two DNA bands are seen (Fig. 4.09). One contains 92 percent of the DNA with a density of 1.701 gm/cm^3 and the smaller, satellite band contains 8 percent of the DNA with a density of 1.690 gm/cm^3. Satellite DNA was originally defined by this density separation. However, in cases where the average satellite DNA base composition is close to that of the genome as a whole, the satellite DNA cannot be physically separated using a density gradient.

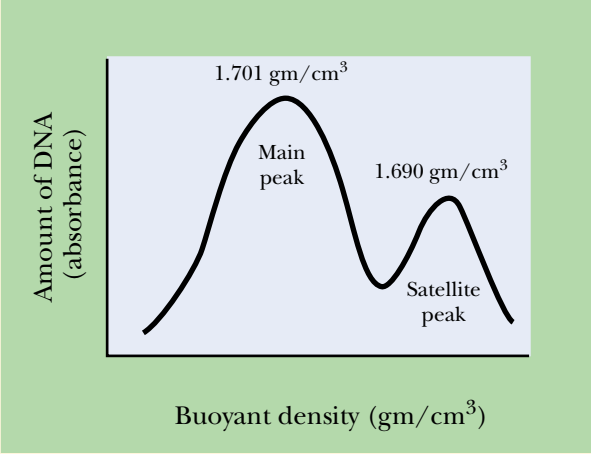


FIGURE 4.09 ***Density Gradient Centrifugation and Satellite Bands***

A cesium chloride gradient will reveal two (or more) bands of fragmented DNA if these differ in density. In this case, the lighter DNA contains sequences that are primarily satellite DNA.

Much of our genome consists of the defunct remains of viruses and transposable elements.

repetitive sequences and other non-coding DNA, including even some introns, may have originated from viral DNA that was inserted into the chromosome of the eukaryotic host cell. Retroviruses, in particular, have played a large role in generating such insertions. In addition, transposable elements (mobile DNA; see Ch. 15) are probably responsible for generating a significant fraction of the repetitive DNA.

The development of repetitive DNA would involve two processes. The original sequences of either viral DNA or transposable elements must have been intact and functional to insert in the first place. Both types of element might replicate and duplicated copies would then be inserted at more locations in the chromosomes of the infected cell. Thus, the numbers of these parasitic sequences would increase. In addition, many of these sequences would mutate, both by base changes and deletions. The result would be a family of related sequences, most of which are no longer functional (Fig. 4.10). Transposable elements that are solely concerned with their own survival and replication, rather than benefiting the host cell in any way are referred to as **selfish DNA**.

Selfish DNA, of no benefit to the host cell, accumulates in the genomes of slow-growing, multi-cellular organisms.

selfish DNA A sequence of DNA that manages to replicate but which is of no use to the host cell it inhabits

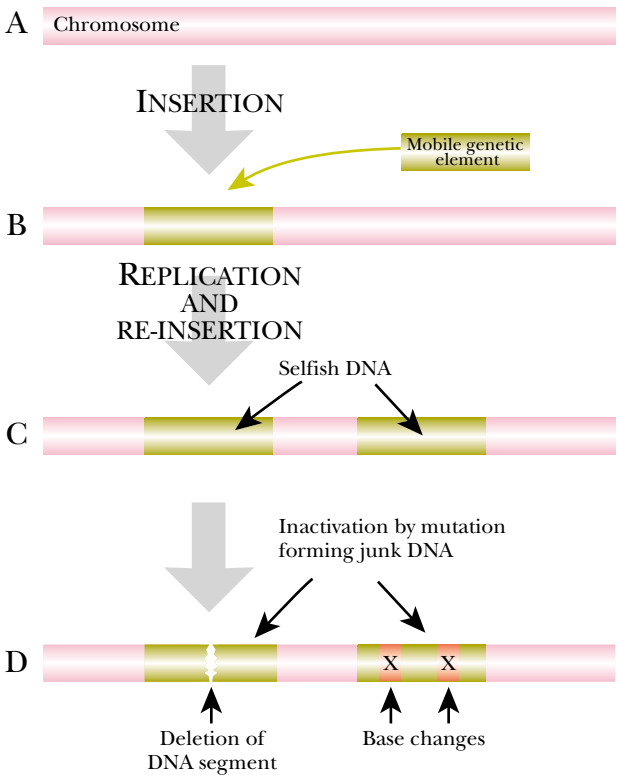


FIGURE 4.10 ***Origins of Selfish DNA and Junk DNA***

1. Insertion of originally mobile DNA; 2. Replication of the inserted DNA; and 3. Deletion and mutation of the inserted DNA appear to be the likely steps in affecting how much junk DNA resides in the genome.

Selfish DNA proliferates through the genome and may be regarded as a subcellular parasite infecting the host chromosomes. The accumulation of selfish DNA depends on two opposing processes, the replication and re-insertion of selfish DNA and its spontaneous deletion. In rapidly dividing, single-celled organisms, such as bacteria, selfish DNA tends to be eliminated, whereas in slowly dividing, multi-cellular organisms it has more opportunity to accumulate. Although originally useless, some defunct selfish DNA sequences may have been put to use by the host cell in non-coding roles such as helping to maintain chromosome structure.

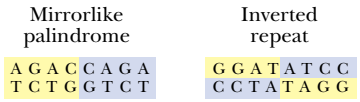


FIGURE 4.11 ***Palindromes and Inverted Repeats***

A mirror-like palindrome and an inverted repeat are shown. Similar colors indicate palindromic or inverted sequences.

Regulatory proteins often bind to DNA at inverted repeat sequences.

Palindromes, Inverted Repeats and Stem and Loop Structures

Palindromes are words or phrases that read the same backwards as forwards. In the case of DNA, which is double stranded, two types of palindromes are theoretically possible. **Mirror-like palindromes** are like those of ordinary text, but involve two strands for DNA. However, in practice, the **inverted repeat** type of palindrome is much more common and of major biological significance. In an inverted repeat, the sequence reads the same forwards on one strand as it reads backwards on the complementary strand (Fig. 4.11).

Inverted repeats are extremely important as recognition sites on the DNA for the binding of a variety of proteins. Many regulatory proteins recognize inverted repeats, as do most restriction and modification enzymes (see Ch 22). In such cases, the inverted repeat usually remains as normal double helical DNA and does not need to be distorted by supercoiling. The term "direct repeat" refers to the situation where the repeated sequences point in the same direction and are on the same strand.

inverted repeat Sequence of DNA that is the same when read forwards as when read backwards, but on the other complementary strand. One type of palindrome
mirrorlike palindrome Sequence of DNA that is the same when read forwards and backwards on the same strand. One type of palindrome
palindrome A sequence that reads the same backwards as forwards

TABLE 4.03 Components of the Eukaryotic Genome

(Numbers of copies given is for the human genome.)

Unique sequences	
Protein encoding genes—comprising upstream regulatory region, exons and introns	
Genes encoding non-translated RNA (snRNA, snoRNA, 7SL RNA, telomerase RNA, Xist RNA, a variety of small regulatory RNAs)	
Non-repetitive intragenic non-coding DNA	
Interspersed Repetitive DNA	
Pseudogenes	
Short Interspersed Elements (SINEs)	
Alu element (300 bp)	~1,000,000 copies
MIR families (average ~130 bp) (mammalian-wide interspersed repeat)	~400,000 copies
Long Interspersed Elements (LINEs)	
LINE-1 family (average ~800 bp)	~200,000–500,000 copies
LINE-2 family (average ~250 bp)	~270,000 copies
Retrovirus like elements (500–1300 bp)	~250,000 copies
DNA transposons (variable; average ~250 bp)	~200,000 copies
Tandem Repetitive DNA	
Ribosomal RNA genes	5 clusters of about 50 tandem repeats on 5 different chromosomes
Transfer RNA genes	multiple copies plus several pseudogenes
Telomere sequences	several kb of a 6 bp tandem repeat
Mini-satellites (= VNTRs)	blocks of 0.1 to 20 kbp of short tandem repeats (5–50 bp), most located close to telomeres
Centromere sequence (α-satellite DNA)	171 bp repeat, binds centromere proteins
Satellite DNA	blocks of 100 kbp or longer of tandem repeats of 20 to 200 bp, most located close to centromeres
Mega-satellite DNA	blocks of 100 kbp or longer of tandem repeats of 1 to 5 kbp, various locations

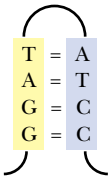


FIGURE 4.12 ***A Hairpin***

If a single strand of DNA containing inverted repeats is folded back upon itself, base pairing occurs forming a hairpin structure.

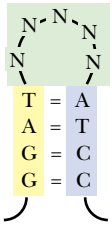


FIGURE 4.13 ***Stem and Loop Motif***

If inverted repeats are separated by a few bases, a stem and loop structure results. The loop contains unpaired bases (NNN).

Consider just a single strand of the inverted repeat sequence. Note that the right and left halves of the sequence on each single strand must be complementary to each other. Thus, such a sequence (e.g., GGATATCC) can be folded into a **hairpin** whose two halves are held together by base pairing (Fig. 4.12).

The U-turn at the top of the hairpin is possible, but energetically unfavorable. In practice, normally a few unpaired bases (shown as N = any base, in the diagram below) are found forming a loop at the top of the base paired stem—the so-called **stem and loop** motif. Such a stem and loop can form from one strand of any inverted repeat that has a few extra bases in the middle (Fig. 4.13).

Multiple A-Tracts Cause DNA to Bend

A DNA sequence that contains several runs of A residues (three to five nucleotides long) separated by 10 bp forms bends. Note that the spacing of the A-tracts corresponds to one turn of the double helix. Bending occurs at the 3′-end of the runs of As

hairpin A double stranded base-paired structure formed by folding a single strand of DNA or RNA back upon itself
stem and loop Structure made by folding an inverted repeat sequence

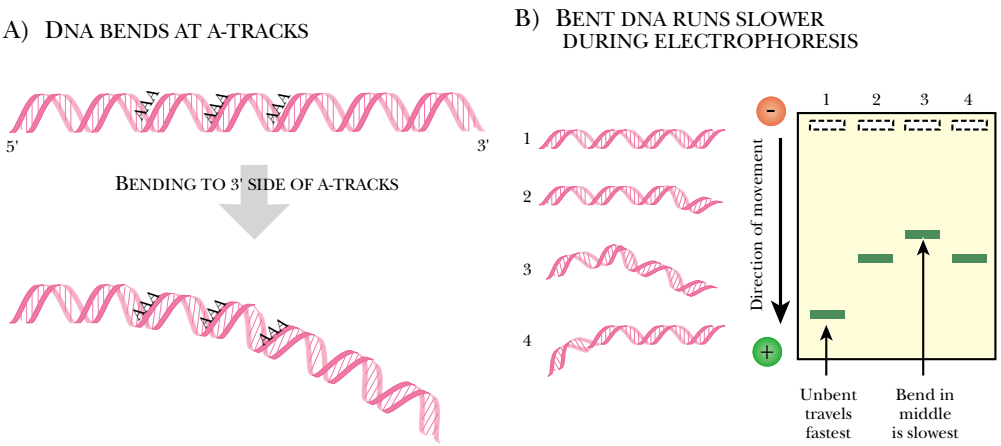


FIGURE 4.14 ***DNA Bending Due to Multiple A Tracts***

(A) Bending of DNA occurs to the 3′-side of A-tracts. (B) Such bending decreases the speed at which DNA travels during electrophoresis. Indeed, the mobility of a DNA molecule of a given length varies depending on the location of bent regions within the molecule. Bends in the middle have greater effect than those close to the ends.

(Fig. 4.14). **Bent DNA** moves more slowly during gel electrophoresis than unbent DNA of the same length (see Ch 21 for electrophoresis).

Bent DNA is found at the origins of replication of some viruses and of yeast chromosomes. It is thought to help the binding of the proteins that initiate DNA replication (see Ch. 5). In addition to "naturally" bent DNA, certain regulatory proteins also bend DNA into U-turns when activating transcription (see Chapters 9 and 10).

Supercoiling is Necessary for Packaging of Bacterial DNA

Bacterial DNA is 1000 times longer than the cell that contains it. The DNA must be supercoiled in order to fit into the cell.

An average bacterial cell is about one millionth of a meter long. The length of the single DNA molecule needed to carry the 4,000 or so genes of a bacterial cell is about one millimeter! Thus, a stretched out bacterial chromosome is a thousand times longer than a bacterial cell. The double helical DNA inside a cell must be **supercoiled** to make it more compact. The DNA, which is already a double helix, is twisted again, as shown in Figure 4.15. The original double helix has a right-handed twist but the supercoils twist in the opposite sense; that is, they are left-handed or **"negative" supercoils**. There is roughly one supercoil every 200 nucleotides in typical bacterial DNA. Negative (rather than positive) supercoiling helps promote the unwinding and strand separation necessary during replication and transcription. [Eukaryotic DNA is also negatively supercoiled, however the mechanism is rather different and involves coiling it around histone proteins as discussed below.]

DNA gyrase puts negative supercoils into the bacterial chromosome.

Negative supercoils are introduced into the bacterial chromosome by DNA gyrase. In the absence of topoisomerase I and topoisomerase IV, the DNA becomes hyper-negatively supercoiled. The steady-state level of supercoiling in *Escherichia coli* is maintained by a balance between topoisomerase IV, acting in concert with topoisomerase I, to remove excess negative supercoils and thus acting in opposition to DNA gyrase. A typical bacterial chromosome contains approximately 50 giant loops of supercoiled DNA arranged around a protein scaffold. In Figure 4.16, the single line represents a double helix of DNA and the helixes are the supercoils.

bent DNA Double helical DNA that is bent due to several runs of As
negative supercoiling Supercoiling with a left handed or counterclockwise twist
supercoiling Higher level coiling of DNA that is already a double helix

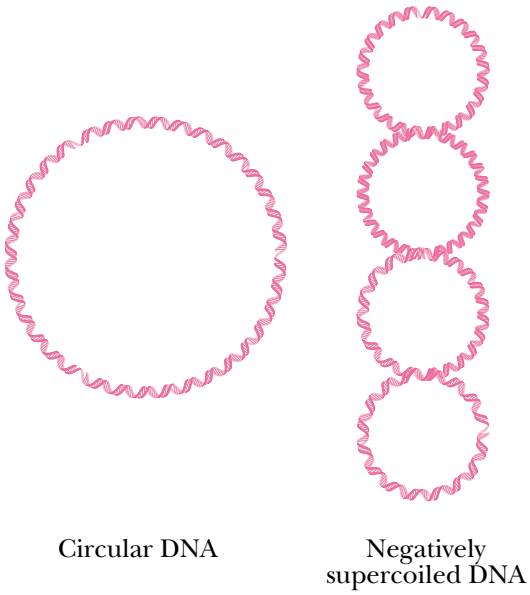


FIGURE 4.15 ***Supercoiling of DNA***

Bacterial DNA is negatively supercoiled in addition to the twisting imposed by the double helix.

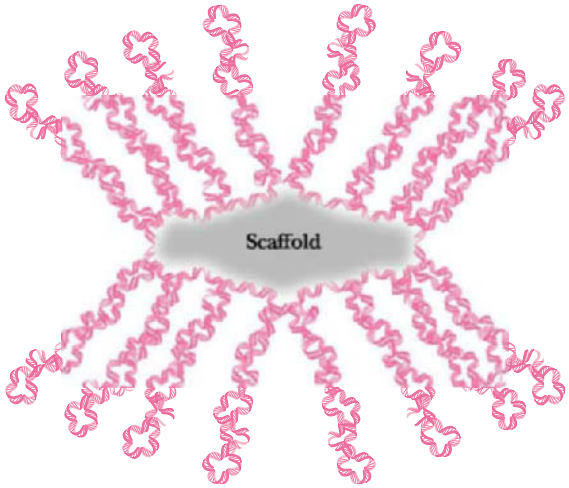


FIGURE 4.16 ***Supercoiling of the Bacterial Chromosome***

Supercoiling of bacterial DNA results in giant loops of supercoiled DNA extending from a central scaffold.

The bacterial chromosome consists of about 50 giant supercoiled loops of DNA.

Bacterial chromosomes and plasmids are double stranded circular DNA molecules and are often referred to as **covalently closed circular DNA**, or **cccDNA**. If one strand of a double stranded circle is nicked, the supercoiling can unravel. Such a molecule is known as an **open circle**.

Topoisomerases and DNA Gyrase

The total amount of twisting present in a DNA molecule is referred to as the **linking number (L)**. This is the sum of the contributions due to the double helix plus the supercoiling. [The number of double helical turns is sometimes known as the **twist**, **T**, and the number of superhelical turns as the **writhe** or **writhing number**, **W**. In this terminology, the linking number, L, is the sum of the twist plus the writhe (L = T + W).]

covalently closed circular DNA (cccDNA) Circular DNA with no nicks in either strand
linking number (L) The sum of the superhelical turns (the writhe, W) plus the double helical turns (the twist, T)
open circle Circular DNA with one strand nicked and hence with no supercoiling
twist, T The number of double helical turns in a molecule of DNA (or double-stranded RNA)
writhe Same as writhing number, W
writhing number, W The number of supercoils in a molecule of DNA (or double-stranded RNA)

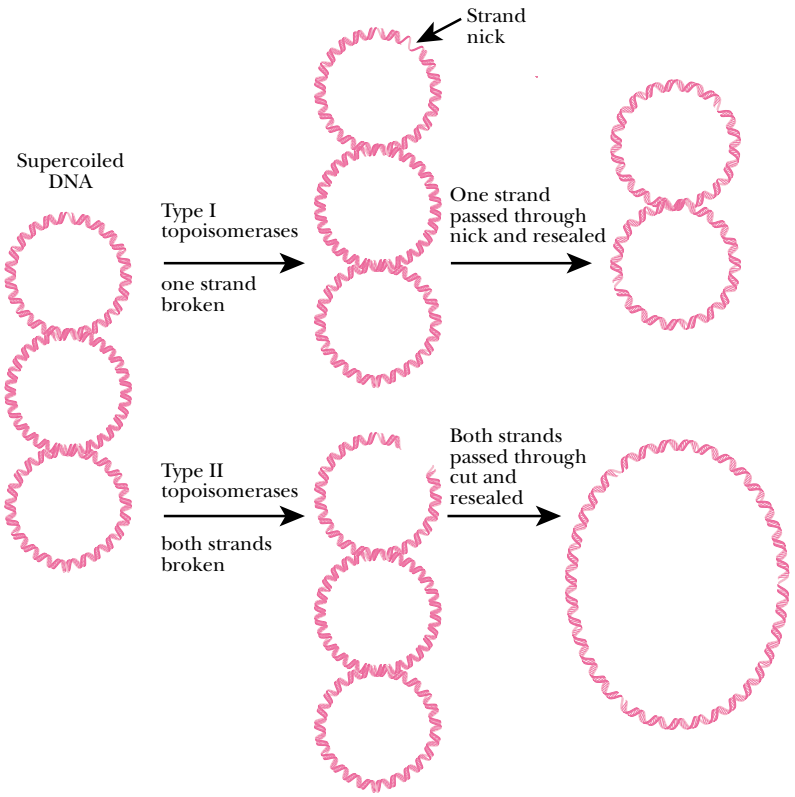


FIGURE 4.17 ***Mechanism of Type I and II Topoisomerases***

The difference in action between topoisomerases of Type I and Type II is in the breakage of strands. Type I breaks only one strand, while Type II breaks both strands. When one strand is broken, the other strand is passed through the break to undo one supercoil. When two strands are broken, double stranded DNA is passed through the break and the supercoiling is reduced by two. After uncoiling, the breaks are rejoined.

Enzymes known as topoisomerases change the level of supercoiling.

The same circular DNA molecule can have different numbers of supercoils. These forms are known as topological isomers, or **topoisomers**. The enzymes that insert or remove supercoils are therefore named **topoisomerases**. **Type I topoisomerases** break only one strand of DNA, which changes the linking number in steps of one. In contrast, **type II topoisomerases** (including **DNA gyrases**) break both strands of the DNA and pass another part of the double helix through the gap. This changes the linking number in steps of two (Fig. 4.17).

DNA gyrase, a type II topoisomerase, introduces negative supercoils into closed circular molecules of DNA, such as plasmids or the bacterial chromosome. Gyrase works by cutting both strands of the DNA, introducing a supertwist and rejoining the DNA strands. Gyrase can generate 1,000 supercoils per minute. As each supertwist is introduced, gyrase changes conformation to an inactive form. Reactivation requires energy, provided by breakdown of ATP. DNA gyrase can also remove negative supercoils (but not positive ones) without using ATP, but this occurs ten times more slowly.

Ciprofloxacin kills bacteria by inhibiting DNA gyrase. It is harmless to animals as they do not use DNA gyrase for compacting their DNA.

DNA gyrase is a tetramer of two different subunits. The GyrA subunit cuts and rejoins the DNA and the GyrB subunit is responsible for providing energy by ATP hydrolysis. DNA gyrase is inhibited by **quinolone antibiotics**, such as **nalidixic acid** and their fluorinated derivatives such as **norfloxacin** and **ciprofloxacin**, which bind to the GyrA protein. An inactive complex is formed in which GyrA protein is inserted into the DNA double helix and covalently attached to the 5′-ends of both broken DNA strands. **Novobiocin** also inhibits gyrase by binding to the GyrB protein and preventing it from binding ATP.

ciprofloxacin A fluoroquinolone antibiotic that inhibits DNA gyrase
DNA gyrase An enzyme that introduces negative supercoils into DNA, a member of the type II topoisomerase family
nalidixic acid A quinolone antibiotic that inhibits DNA gyrase
norfloxacin A fluoroquinolone antibiotic that inhibits DNA gyrase
novobiocin An antibiotic that inhibits type II topoisomerases, especially DNA gyrase, by binding to the B-subunit
quinolone antibiotics A family of antibiotics, including nalidixic acid, norfloxacin and ciprofloxacin, that inhibit DNA gyrase and other type II topoisomerases by binding to the A-subunit
topoisomerase Enzyme that alters the level of supercoiling or catenation of DNA (i.e. changes the topological conformation)
topoisomers Isomeric forms that differ in topology—i.e. their level of supercoiling or catenation
type I topoisomerase Topoisomerase that cuts a single strand of DNA and therefore changes the linking number by one
type II topoisomerase Topoisomerase that cuts both strands of DNA and therefore changes the linking number by two

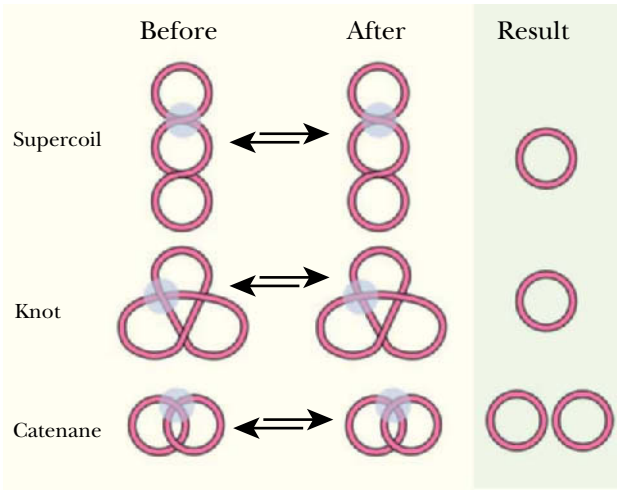


FIGURE 4.18 ***Unlinking of Catenanes by Topoisomerase IV***

Topoisomerases may uncoil, unknot or unlink DNA as well as carrying out the coiling, knotting or interlinking of DNA. Topoisomerases act (at the locations shaded blue) by cutting both strands of the DNA at one location and passing another region of the DNA through the gap.

Catenated and Knotted DNA Must Be Corrected

Circular molecules of DNA may become interlocked during replication or recombination. Such structures are called **catenanes**. The circles may be liberated by certain type II topoisomerases, such as topoisomerase IV (Fig. 4.18) of *E. coli* and related enzymes. Circular DNA molecules may also form knots. Type II topoisomerases can both create and untie knots. Like DNA gyrases, these enzymes are tetramers of two different subunits, one for cutting the DNA and the other for energy coupling. Like gyrase, topoisomerase IV is inhibited by quinolone antibiotics.

Local Supercoiling

When DNA is replicated or when genes are expressed, the double helix must first be unwound. This is aided by the negative supercoiling of the chromosome. However, as the replication apparatus proceeds along a double helix of DNA, it creates positive supercoiling ahead of itself. Similarly, during transcription, when RNA polymerase proceeds along a DNA molecule, it also creates positive supercoiling ahead of itself. For replication and transcription to proceed more than a short distance, DNA gyrase must insert negative supercoils to cancel out the positive ones. Behind the moving replication and transcription apparatus, a corresponding wave of negative supercoiling is generated. Excess negative supercoils are removed by topoisomerase I.

As a result, at any given instant, the extent of supercoiling varies greatly in any particular region of the chromosome. It has been suggested that supercoiling might regulate gene expression. However, only rare examples are known; thus transcription of the gene for DNA gyrase in *E. coli* is regulated by supercoiling. More often, the opposite is the case. Local supercoiling depends largely on the balance between transcription and the restoration of normal supercoiling by gyrase and topoisomerase.

Supercoiling Affects DNA Structure

Supercoiling places DNA under physical strain. This may lead to the appearance of alterations in DNA structure that serve to relieve the strain. Three of these alternate forms are **cruciform structures**; the left-handed double helix, or **Z-DNA**; and the triple helix, or **H-DNA**. All of these structures depend on certain special characteristics within the DNA sequence, as well as supercoiling stress.

catenane Structure in which two or more circles of DNA are interlocked
cruciform structure Cross shaped structure in double stranded DNA (or RNA) formed from an inverted repeat
H-DNA A form of DNA consisting of a triple helix. Its formation is promoted by acid conditions and by runs of purine bases
Z-DNA An alternative form of DNA double helix with left-handed turns and 12 base pairs per turn

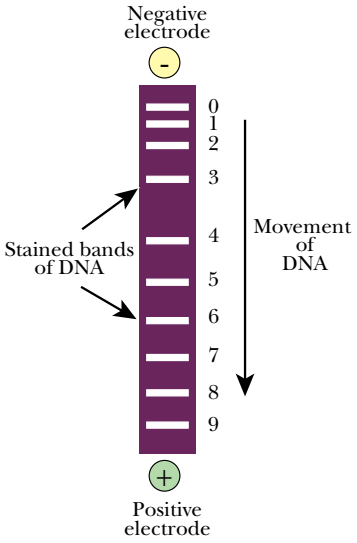


FIGURE 4.19 ***Separation of Supercoiled DNA by Electrophoresis***

Supercoiled DNA molecules, all of identical sequence, were electophoresed to reveal multiple bands, with each band differing in the number of supercoils. The number of supercoils is shown beside the band. Zero refers to open circular DNA, which is not supercoiled at all.

Separation of Topoisomers by Electrophoresis

DNA molecules of different sizes are routinely separated by electrophoresis on agarose gels (see Ch. 21). The mobility of a DNA molecule in such a gel depends on both its molecular weight and its conformation. Heavier molecules travel more slowly but among molecules of the same molecular weight, those that are more compact move faster. Consequently, for a given DNA molecule, supercoiled cccDNA moves faster than open circular DNA, which in turn moves faster than linearized DNA.

For small circular molecules of DNA, it is even possible to separate topoisomers with different numbers of superhelical twists. The more superhelical twists, the more compact the molecule is and the faster it moves in an electrophoretic field (Fig 4.19).

When the DNA gyrase of an *E. coli* strain containing small plasmids is inhibited, it is possible to isolate the plasmids and demonstrate that they are now positively supercoiled. Topoisomerase I continues to remove the excess negative supercoils but the DNA gyrase no longer cancels out the surplus positive supercoils, which therefore accumulate.

The cruciform (cross-like) structures are formed when the strands in a double stranded DNA palindrome are separated and formed into two stem and loop structures opposite each other (Fig. 4.20). The probability that cruciform structures will form increases with the level of negative supercoiling and the length of the inverted repeat. In practice, the four to eight base sequences recognized by most regulatory proteins and restriction enzymes are too short to yield stable cruciform structures. Palindromes of 15 to 20 base pairs will produce cruciform structures. Their existence can be demonstrated because they allow single strand specific nucleases to cut the double helix. (A nuclease is an enzyme that cuts nucleic acid strands; see Ch. 22) Cutting occurs within the small single stranded loop at the top of each hairpin. Cruciform structures partially straighten supercoiled DNA and so the molecule is not folded up as compactly and thus travels slower during gel electrophoresis.

Alternative helical structures are sometimes found for DNA, in addition to the common form, made famous by Watson and Crick and known as B-DNA.

Alternative Helical Structures of DNA Occur

Several double helical structures are actually possible for DNA. Watson and Crick described the most stable and most common of these. It is right handed with 10 base pairs per turn of the helix. The grooves running down the helix are different in depth and referred to as the major and minor grooves. The standard Watson and Crick double helix is referred to as the **B-form** or as **B-DNA** to distinguish it from the other helical forms, **A-DNA**, and Z-DNA (Fig. 4.21). Most of these structures apply not only to double stranded DNA (dsDNA), but also to RNA when it is double stranded.

The A-form helix is found in dsRNA or DNA/RNA hybrids. It has 11 bp per turn—one more than B-DNA.

The **A-form** of the double helix is shorter and fatter than the B-form and has 11 base pairs per helical turn. In the A-form, the bases tilt away from the axis, the minor groove becomes broader and shallower, and the major groove becomes narrower and deeper. Double-stranded RNA or hybrids with one RNA and one DNA strand usually form an A-helix. The extra hydroxyl group, at the 2′ position of ribose, prevents double-stranded RNA from forming a B-helix. Double-stranded DNA tends to form an A-helix only at a high salt concentration or when it is dehydrated. The tendency to form an A-helix also depends on the sequence. The physiological relevance, if any, of A-DNA is obscure. However, double-stranded regions of RNA exist in this form *in vivo*.

A-DNA A rare alternative form of double stranded helical DNA
A-form An alternative form of the double helix, with 11 base pairs per turn, often found for double stranded RNA, but rarely for DNA
B-form or B-DNA The normal form of the DNA double helix, as originally described by Watson and Crick

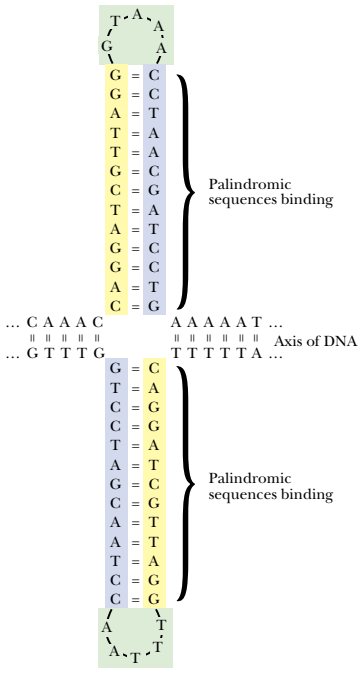


FIGURE 4.20 ***Cruciform Structure Formed from an Inverted Repeat***

Because the DNA is palindromic, the strands can separate and base pair with themselves to form lateral cruciform extensions.

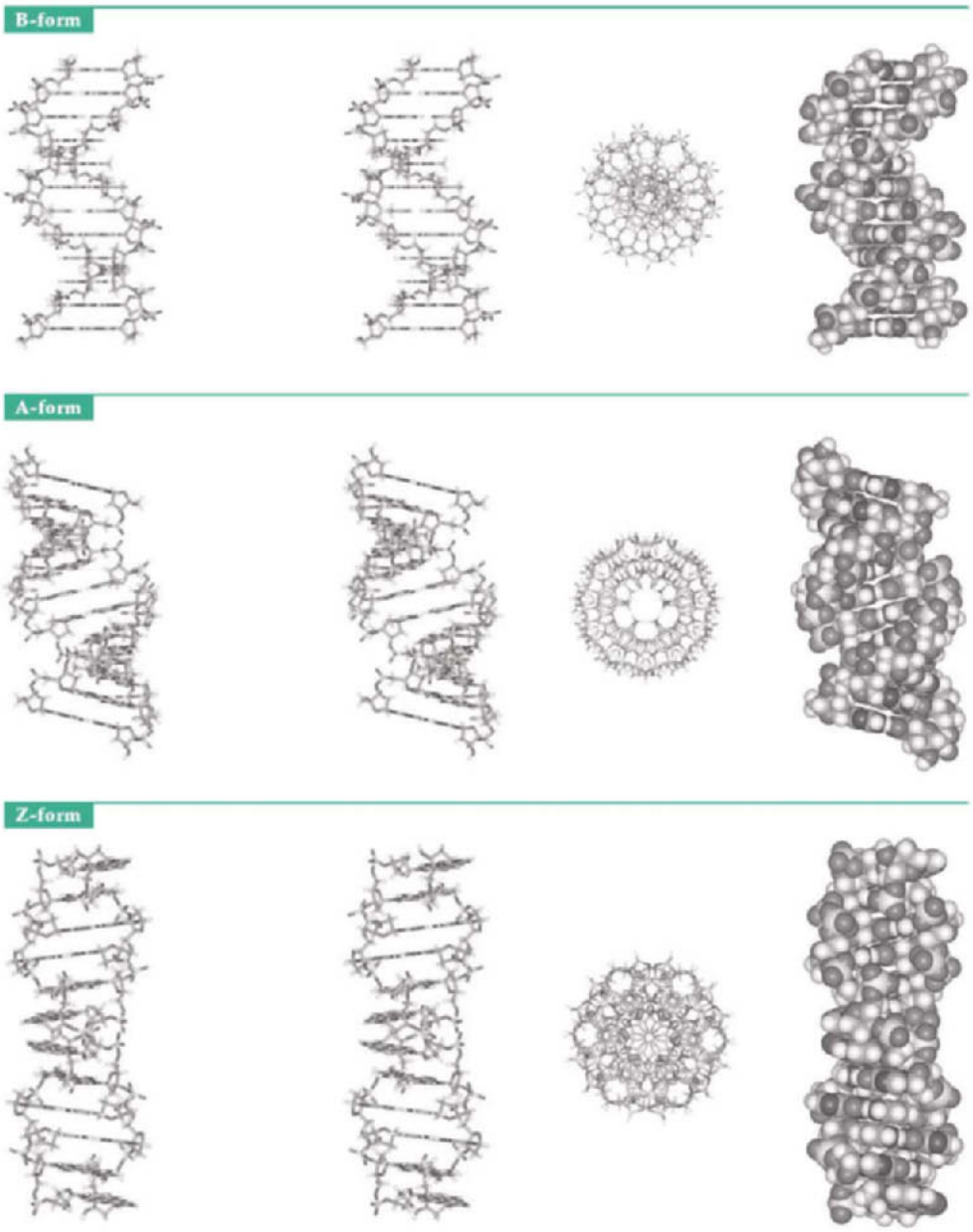


FIGURE 4.21 ***Comparison of B-DNA, A-DNA and Z-DNA***

Several structurally different versions of the double helix exist. Shown here are the normal Watson-Crick double helix, the B-form, together with the rarer A-form and Z-DNA form. From left to right: stereoscopic skeleton pairs, end view down the helix axis, and space filling models. Courtesy of Tamar Schlick.

Z-DNA is a left-handed double helix with 12 base pairs per turn.

Z-DNA is a left-handed double helix with 12 base pairs per turn. It is thus longer and thinner than B-DNA and its sugar phosphate backbone forms a zigzag line rather than a smooth helical curve (Fig. 4.21). High salt favors Z-DNA as it decreases repulsion between the negatively charged phosphates of the DNA backbone. Z-DNA is formed in regions of DNA that contain large numbers of alternating GC or GT pairs, such as:

```
GCGCGCGCGCGCGC  or  GTGTGTGTGTGTGT
CGCGCGCGCGCGCG  or  CACACACACACACA
```

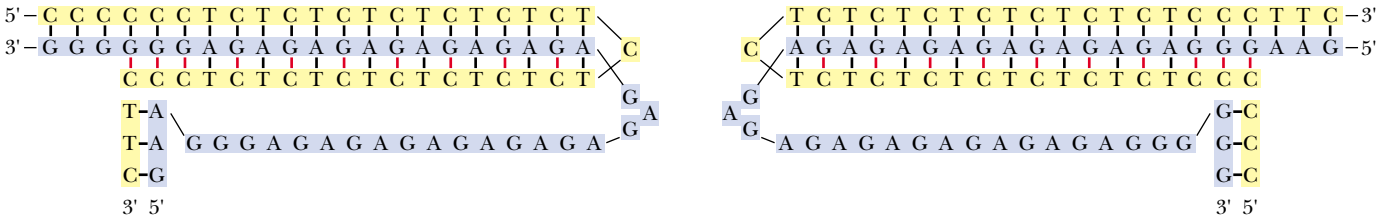


FIGURE 4.22 ***Structure of H-DNA***

This triple helix is formed by GA- and TC-rich regions of a plasmid and is composed of triads of bases.

Such tracts may be abbreviated as $(GC)_n \cdot (GC)_n$ and $(GT)_n \cdot (AC)_n$. Note that the sequence of each individual strand is written in the 5′- to 3′-direction. That Z-DNA depends specifically on G (not A) alternating with C or T is shown by the fact that DNA with many alternating AT pairs forms cruciform structures, not Z-DNA.

Since Z-DNA is a left-handed helix, its appearance in part of a DNA molecule helps to remove supercoiling stress. As negative supercoiling increases, the tendency for GC- or GT-rich tracts of DNA to take the **Z-form** increases. Its existence can be demonstrated in small plasmids by changes in electrophoretic mobility. Single-strand specific nucleases can cut DNA at the junction between segments of Z-DNA and normal B-DNA.

Short artificial segments of double stranded DNA made solely of repeating GC units take the Z-form even in the absence of supercoiling, provided that the salt concentration is high. (This was how Z-DNA was originally discovered.) Antibodies to Z-DNA can be made by immunizing animals with such linear (GC)n fragments. These antibodies can then be used to show the presence of Z-form (helix) regions in natural DNA, provided it is highly supercoiled.

It has been suggested that regions of Z-helix may be specifically recognized by certain enzymes. An example is the RNA editing enzyme ADAR1, which modifies bases in dsRNA (see Ch. 12 for the details of RNA processing). ADAR1 stands for adenosine deaminase (RNA) type I and it removes the amino group of adenosine, so converting it to **inosine**. It requires a double stranded RNA substrate that, in practice, is formed by folding an intron back onto the neighboring exon. ADAR1 contains separate binding motifs for both DNA and for dsRNA. It has been postulated that the DNA binding domain recognizes Z-DNA because base modification must occur before cutting and splicing of the introns and exons. When RNA polymerase moves along, transcribing DNA into RNA, it generates positive supercoils ahead and negative supercoils behind. Negative supercoiling induces the formation of Z-DNA, especially in GC- or GT-regions. Consequently, a zone of Z-DNA will be found just behind the RNA polymerase. Binding to Z-DNA would ensure that ADAR1 works on newly synthesized RNA.

H-DNA is even more peculiar—it is not a double but a *triple* helix. It depends on long tracts of purines in one strand and, consequently, only pyrimidines in the other strand; e.g.:

```
GGGGGGGGGGGGGG  or  GAGAGAGAGAGAGAG
CCCCCCCCCCCCCC  or  CTCTCTCTCTCTCTC
```

Two such segments are required and may interact forming H-DNA when the DNA is highly supercoiled (Fig. 4.22). In addition, the overall region must be a mirror-like palindrome. H-DNA contains a triple helix, consisting of one purine-rich strand and two pyrimidine-rich strands. The other purine-rich strand is displaced and left unpaired.

inosine A purine nucleoside, found most often in transfer RNA, that contains the unusual base hypoxanthine
Z-form An alternative form of double helix with left-handed turns and 12 base pairs per turn. Both DNA and dsRNA may be found in the Z-form

H-DNA involves strange "sideways" base pairing to form a triple helix.

In H-DNA, adenine pairs with two thymines and guanine pairs with two cytosines. In each case, one pairing is normal, the other sideways (so-called **Hoogsteen base pairs**—hence the name H-DNA). Furthermore, to form the C=G=C triangle, an extra proton (H^+) is needed for one of the hydrogen bonds. Consequently, formation of H-DNA is promoted by acidic conditions. High acidity also tends to protonate the phosphate groups of the DNA backbone, so decreasing their negative charges. This reduces the repulsion between the three strands and helps form a triple helix.

Despite these complex sequence requirements, computer searches of natural DNA have shown that potential sequences that might form triplex H-DNA are much more frequent than expected on a random basis. These are called **PIT (potential intra-strand triplex)** elements. For example, the *E. coli* genome contains 25 copies of a 37 base PIT element. Isolated PIT element DNA does form a stable triplex even at neutral pH. Not surprisingly, the presence of artificial triplexes has been shown to block transcription. This suggests that H-DNA does have some real biological function, although what this is remains obscure.

Histones Package DNA in Eukaryotes

Eukaryotes package their DNA by winding it around specialized proteins, known as histones.

Plants and animals have vastly more DNA than bacteria and must fold this DNA to fit into the cell nucleus. Typical bacteria carry approximately 4,000 genes on a single chromosome, which is about one millimeter long. The chromosome is thus 1,000 times longer than the bacterial cell in which it fits. Eukaryotic chromosomes may be as much as a centimeter long and must be folded up to fit into the cell nucleus, which is five microns across, a necessity for a 2,000-fold shortening. However, eukaryotic chromosomes are not circular, and instead of supercoiling using DNA gyrase, the mechanism of packaging involves winding the DNA around special proteins, the **histones**.

Eukaryotic DNA starts folding by coiling around the histones, positively charged proteins that neutralize the negative charge of the DNA itself. DNA with histones bound to it was named **chromatin** when it was first discovered in chromosomes. Chromatin consists of roughly spherical subunits, the **nucleosomes**, each containing approximately 200 bp of DNA and nine histones, two each of H2A, H2B, H3 and H4 and one of H1 (Fig. 4.23).

200 bp of DNA is wound around nine histones, making up a particle known as a nucleosome. Eukaryotic chromosomes consist of long strings of nucleosomes that are folded up further.

The eight paired histones cluster together and two coils of DNA are wrapped around them. Each coil of DNA is approximately 80 bp long so that this core particle accommodates 160 bp of DNA overall. The remaining 40 bp or so of DNA forms a linker region between neighboring core particles that may vary somewhat in length (from 10 to 100 bp) depending on the DNA sequence. The ninth histone, H1, joins each core particle to the next.

The DNA in the linker region is relatively exposed and can be cut with nucleases specific for dsDNA; these nucleases make double stranded cuts. In practice, micrococcal nuclease is often used to perform this job. Cutting occurs in three stages. First, single nucleosomes with 200 bp of DNA are released, then the linker region is cut off, leaving about 165 bp. Finally the ends of the DNA wound around the core are nibbled away, leaving about 146 bp that are fully protected by the core particle from further digestion.

The core histones, H2A, H2B, H3 and H4, are small, roughly spherical proteins with 102 to 135 amino acids. However, the linker histone, H1, is longer, having about 220 amino acids. H1 has two arms extending from its central spherical domain. The central part of H1 binds to its own nucleosome and the two arms bind to the nucleosomes on either side (Fig. 4.24).

chromatin Complex of DNA plus protein which constitutes eukaryotic chromosomes
histone Special positively charged protein that binds to DNA and helps to maintain the structure of chromosomes in eukaryotes
Hoogsteen base pair A type of nonstandard base pair found in triplex DNA, in which a pyrimidine is bound sideways on to a purine
nucleosome Subunit of a eukaryotic chromosome consisting of DNA coiled around histone proteins
potential intrastrand triplex (PIT) Stretch of DNA that might be expected from its sequence to form H-type triplex DNA

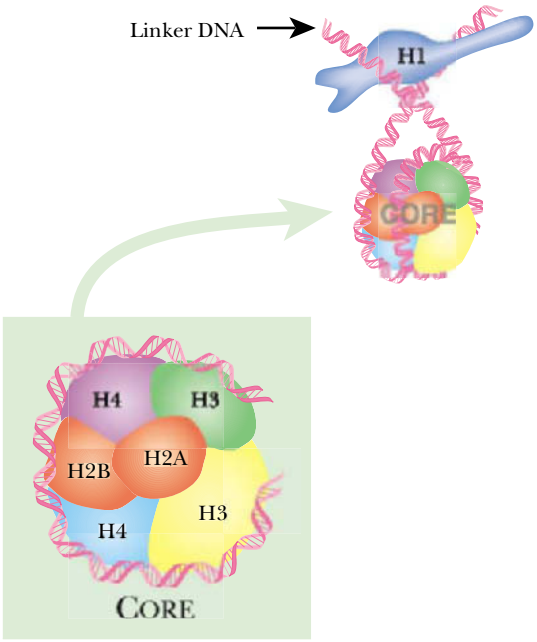


FIGURE 4.23 ***Nucleosomes and Histones***

The basic unit in the folding of eukaryotic DNA is the nucleosome as shown here. A nucleosome is composed of eight histones comprising a core and one separate histone (H1) at the site where the wrapped DNA diverges. The enlarged region shows the packing of histones in the core. The H3-H4 tetramer dictates the shape of the core. Only one of the H2A and H2B dimers is shown; the other is on the other side, hidden from view.

The core histones have a body of about 80 amino acids and a tail of 20 amino acids at the N-terminal end. This tail contains several lysine residues that may have acetyl groups added or removed (Fig 4.25). This is thought to partly control the state of DNA packaging and hence of gene expression. Thus, in active chromatin, the core histones are highly acetylated (see Ch. 10 for further discussion).

Further Levels of DNA Packaging in Eukaryotes

Eukaryotic DNA is so long that it needs several successive levels of folding to fit into the nucleus.

DNA covered with histones and twisted into a series of nucleosomes resembles a string of beads and is sometimes called the "beads on a string" form. However, the folding process continues. The chain of nucleosomes is wound into a giant helical structure with six nucleosomes per turn. It is now known as the **30 nanometer fiber** (Fig. 4.26). In turn, these fibers are looped back and forth. The loops vary in size, averaging about 50 of the helical turns (i.e., about 300 nucleosomes) per loop. The ends of the loops are attached to a protein scaffold, or chromosome axis.

Further folding of chromosomes occurs in preparation for cell division. The precise nature of this is uncertain, but condensed mitotic chromosomes are 50,000 times shorter than fully extended DNA. Highly condensed chromatin is known as **heterochromatin**, appearing dense in the light and electron microscope. In this form it cannot be transcribed (see Ch. 10 for discussion). [Note that some regions of DNA (e.g., satellite DNA near the centromeres) are always found as heterochromatin whereas active regions of the genome condense into heterochromatin during cell division.] An overall summary of DNA folding is presented in Figure 4.27.

Between cell divisions, regions of heterochromatin persist around the centromere and at the ends of the chromosome. These regions include the satellite DNA discussed above and make up about 10 percent of the chromosome. The rest of the chromatin, the **euchromatin**, is in the more extended form shown as a string of beads in panels B and C of Figure 4.27. About 10 percent of this euchromatin is even less condensed and is either being transcribed or is accessible for transcription in the near future (see Ch 10 for details). This is the "active chromatin." During both replication and transcription, the histones are temporarily displaced from short regions of the DNA. After the synthetic enzymes have passed by, the histone cores reassemble on the DNA.

30 nanometer fiber Chain of nucleosomes that is arranged helically, approximately 30 nm in diameter
euchromatin Normal chromatin, as opposed to heterochromatin
heterochromatin A highly condensed form of chromatin that cannot be transcribed because it cannot be accessed by RNA polymerase

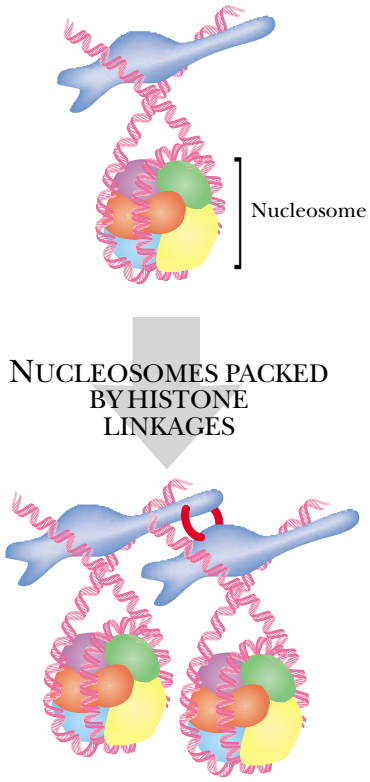


FIGURE 4.24 ***Histone H1 Links Nucleosomes***

The positioning of H1 (blue) above the DNA wrapped around the core particles allows one H1 to bind to another along a linear chain of nucleosomes. This helps in the tighter packing of the nucleosomes.

TABLE 4.04 Summary of Chromosome Folding

Level of Folding	Consists of	Base Pairs per Turn
DNA double helix	nucleotides	10
Nucleosomes	200 base pairs each	100
30 nanometer fiber	6 nucleosomes per turn	1,200
Loops	50 helical turns per loop	60,000
Chromatid	2,000 loops	

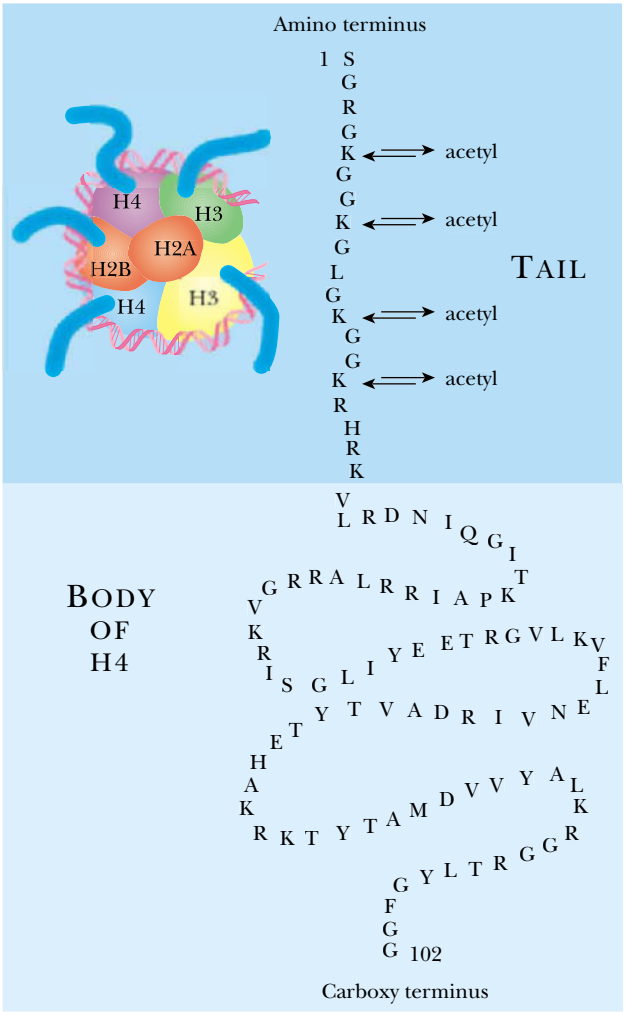


FIGURE 4.25 ***Histone Tails May Be Acetylated***

The N-terminal domains of some of the histone proteins are free for acetylation as indicated by "acetyl." The single letter system for naming amino acids is used.

Histones Are Highly Conserved and Originated Among the Archaebacteria

Of all known proteins, the eukaryotic core histones, especially H3 and H4, are the most highly conserved during evolution. For example, only two amino acids are different, out of 102, between the H4 of cows and peas. The linker histone, H1, is more variable in composition.

Typical bacteria (i.e., the eubacteria) do not possess histones. [Large numbers of "histone-like proteins" are found bound to bacterial chromosomes. Despite the name, these are not homologous in sequence to true histones nor do they form nucleosomes for packaging DNA.] However, some members of the genetically distinct lineage of archaebacteria (e.g., the methane bacteria), do possess histones. Archaeal histones vary significantly from each other. They are 65–70 amino acids long and are missing the tails characteristic of eukaryotic histones. Archaeal nucleosomes accommodate a little under 80 bp of DNA and contain a tetramer of the archaeal histone. They are probably homologous to the $(H3 + H4)_2$ tetramers found in the core of the eukaryotic nucleosome.

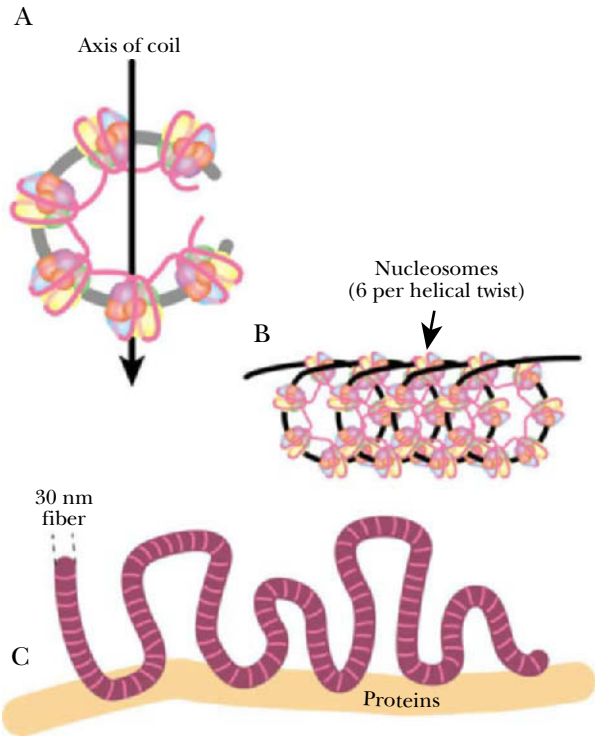


FIGURE 4.26 ***Looping of 30 Nanometer Fiber on Chromosome Axis***

A) A chain of nucleosomes is coiled further with six nucleosomes forming each turn. B) The coiled nucleosomes form a helix, known as a 30 nm fiber. C) The 30 nm fibers form loops that are periodically anchored to a protein scaffolding.

When chromosomes are visible under the light microscope, it is because they are present in a cell that has been caught in the act of dividing. Between cell divisions, most of the DNA is less condensed. It consists of a single, extended molecule of double helical DNA and does not look at all like typical chromosome pictures. Just before cell division, the DNA condenses and folds up, as described above. The typical metaphase chromosome, seen in most pictures, has replicated its DNA some time previously, and is about to divide into two daughter chromosomes, as shown in Figure 4.28. It therefore consists of two identical double helical DNA molecules that are still held together at the centromere. These are known as **chromatids**. Note that between cell

chromatid Single double-helical DNA molecule making up whole or half of a chromosome. A chromatid also contains histones and other DNA-associated proteins.

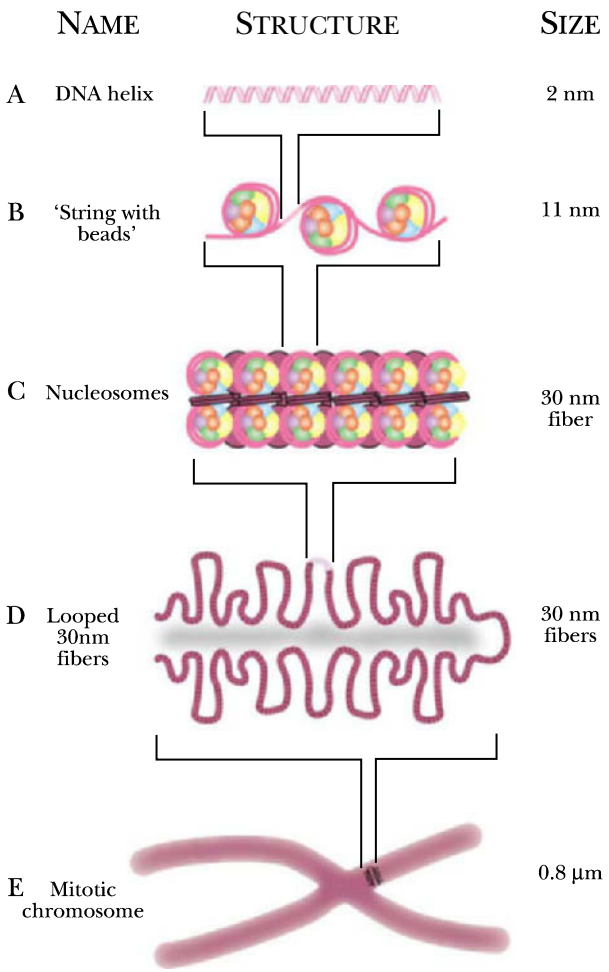


FIGURE 4.27 ***Summary of the Folding of DNA in Eukaryotic Chromosomes***

The DNA helix (A) is wrapped around (B) eight histones (the core). The linker DNA regions unite the nucleosomes to give a "string with beads." This in turn is coiled helically (C) (not clearly indicated) to form a 30 nm fiber. The 30 nm fibers are further folded by looping and attachment to a protein scaffold. Finally, during mitosis the DNA is folded yet again to yield very thick chromosomes.

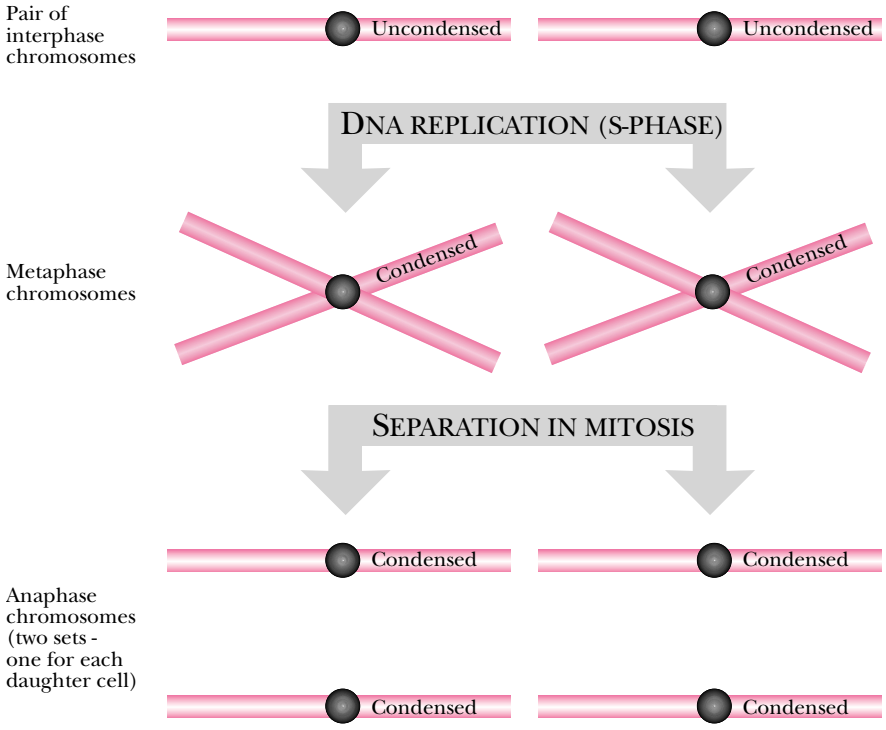


FIGURE 4.28 ***Interphase and Metaphase Chromosomes***

Between rounds of cell division, chromosomes consist of single chromatids, and are referred to as interphase chromosomes. Before the next cell division, the DNA is replicated and each chromosome consists of two DNA molecules or chromatids linked at the centromere. Just prior to mitosis, condensation occurs, making the chromosomes (and chromatids) visible. The chromosomes are best viewed while spread out during the middle part (metaphase) of mitosis. Each daughter cell will acquire one of the chromatids and the process begins anew.

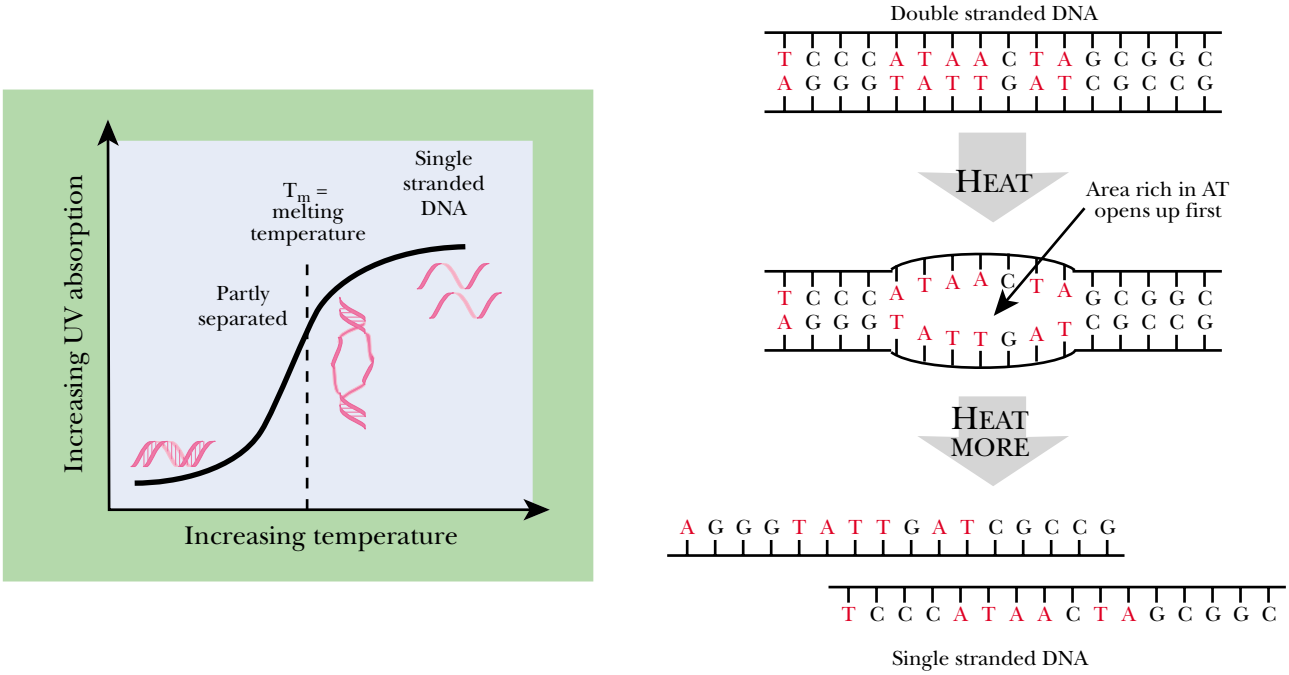


FIGURE 4.29 ***Melting of DNA***

DNA strands separate, or "melt," with increasing temperature. This curve shows the measurement of DNA separation by ultraviolet absorption. As the temperature increases more UV is absorbed by the individual strands. The Tm or melting temperature is the point at which half of the double stranded DNA has separated. During the melting process regions rich in A=T melt first since these basepairs have only two hydrogen bonds.

divisions and in non-dividing cells each chromosome consists of only a single chromatid. The term chromatid is mostly needed only to avoid ambiguity when describing chromosomes in process of division. In addition, there are a few unusual cases where giant chromosomes with multiple chromatids are found in certain organisms (e.g. in the salivary glands of flies).

Melting Separates DNA Strands; Cooling Anneals Them

Heating breaks hydrogen bonds and eventually causes the two strands of a DNA double helix to separate—the DNA "melts".

Hydrogen bonds are rather weak, but since a molecule of DNA usually contains millions of base pairs, the added effect of millions of weak bonds is strong enough to keep the two strands together (Fig. 4.29). When DNA is heated, the hydrogen bonds begin to break and the two strands will eventually separate if the temperature rises high enough. This is referred to as "**melting**" or **denaturation** and each DNA molecule has a **melting temperature** (Tm) that depends on its base composition. Therefore, the melting temperature of a DNA molecule is defined strictly as the temperature at the halfway point on the melting curve, as this is more accurate than trying to guess where exactly melting is complete.

The melting temperature is affected by the pH and salt concentration of the solution, so these must be standardized if comparisons are to be made. Extremes of pH disrupt hydrogen bonds. A highly alkaline pH deprotonates the bases which abolishes their ability to form hydrogen bonds and at pH > 11.3 DNA is fully denatured. Con-

denaturation When used of proteins or other biological polymers, refers to the loss of correct 3-D structure
melting When used of DNA, refers to its separation into two strands as a result of heating
melting temperature (Tm) The temperature at which the two strands of a DNA molecule are half unpaired

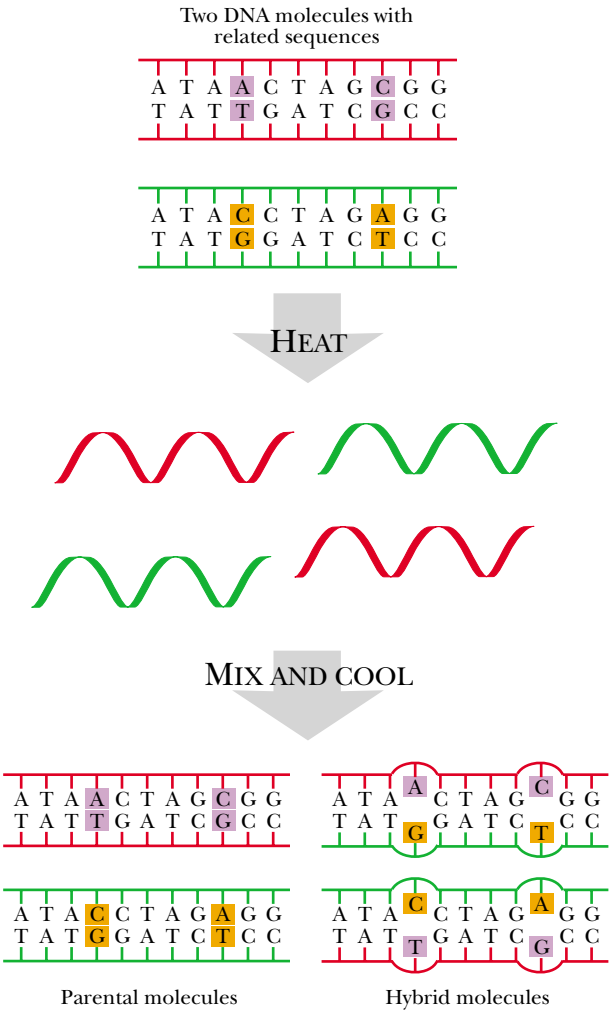


FIGURE 4.30 ***Annealing of DNA***

When pure DNA is heated and cooled the two strands pair up again, i.e., they re-anneal. When DNA from two related sources is heated and re-annealed, hybrid strands may form. The likelihood of hybridization depends on how closely related the two sequences are.

versely, a very low pH causes excessive protonation, which also prevents hydrogen bonding. When DNA is deliberately denatured by pH, alkaline treatment is used because unlike acid, this does not affect the glycosidic bonds between bases and deoxyribose. DNA is relatively more stable at higher ionic concentrations. This is because ions suppress the electrostatic repulsion between the negatively charged phosphate groups on the backbone and hence exert a stabilizing effect. In pure water, DNA will melt even at room temperature.

Melted DNA absorbs more UV light than double-helical DNA.

A spectrophotometer detects the amount absorbed when light is passed through a solution containing DNA. This is compared with the light absorbed by a solution containing no DNA to determine the amount absorbed by the DNA itself. Melting is followed by measuring the absorption of ultraviolet (UV) light at a wavelength of 260 nm (the wavelength of maximum absorption), since disordered DNA absorbs more UV light than a double helix (see Ch. 21).

The more GC base pairs (with three hydrogen bonds) the higher the melting temperature for DNA.

Overall, the higher the proportion of GC base pairs, the higher the melting temperature of a DNA molecule. This is because AT base pairs are weaker, as they have only two hydrogen bonds, as opposed to GC pairs, which have three. In addition, the stacking of GC base pairs with their neighbors is also more favorable than for AT base pairs. In the early days of molecular biology, melting temperatures were used to estimate the percentage of GC versus AT in samples of DNA. DNA base compositions are often cited as the **GC ratio**. The GC content (% G + C) is calculated from the fractional composition of bases as follows:

GC ratio The amount of G plus C relative to all four bases in a sample of DNA. The GC ratio is usually expressed as a percentage

$$\frac{G+C}{A+T+G+C} \times 100\%$$

GC contents for the DNA from different bacterial species vary from 20 percent to 80 percent, with *E. coli* having a ratio of 50 percent. Despite this, there is no correlation between GC content and optimum growth temperature. Apparently this is because the genomes of bacteria are circular DNA molecules with no free ends and this greatly hinders unraveling at elevated temperatures. In fact, small circular DNA molecules, like plasmids, may remain base paired up to 110–120°C. In contrast, the range of GC contents for animals (which have linear chromosomes) is much narrower, from approximately 35–45 percent, with humans having 40.3 percent GC.

As a DNA molecule melts, regions with a high local concentration of AT pairs will melt earlier and GC-rich regions will stay double stranded longer. When DNA is replicated, the two strands must first be pulled apart at a region known as the origin of replication (see Ch. 5). The DNA double helix must also be opened up when genes are transcribed to make mRNA molecules. In both cases, AT-rich tracts are found where the DNA double helix will be opened up more readily.

Upon cooling, the bases in the separated strands of DNA can pair up again and the double helix can re-form.

If the single strands of a melted DNA molecule are cooled, the single DNA strands will recognize their partners by base pairing and the double stranded DNA will re-form. This is referred to as **annealing** or **renaturation**. For proper annealing, the DNA must be cooled slowly to allow the single strands time to find the correct partners. Furthermore, the temperature should remain moderately high to disrupt random H-bond formation over regions of just one or a few bases. A temperature 20–25°C below the Tm is suitable. If DNA from two different, but related, sources is melted and reannealed, **hybrid DNA** molecules may be obtained (Fig. 4.30).

Hybrid DNA molecules may be formed by heating and cooling a mixture of two different, but related, DNA molecules.

Hybridization of DNA and/or RNA was originally used to estimate the relatedness of different organisms, especially bacteria where the amount of DNA is relatively small, in the days before direct sequencing of DNA became routine. Other uses for hybridization include detection of specific gene sequences and gene cloning. Several extremely useful techniques that are still current and are based on the hybridization of DNA and/or RNA are described in detail in Chapter 21.

annealing The re-pairing of separated single strands of DNA to form a double helix
hybrid DNA Artificial double stranded DNA molecule made by pairing two single strands from two different sources
hybridization Pairing of single strands of DNA or RNA from two different (but related) sources to give a hybrid double helix
renaturation Re-annealing of single-stranded DNA or refolding of a denatured protein to give the original natural 3-D structure