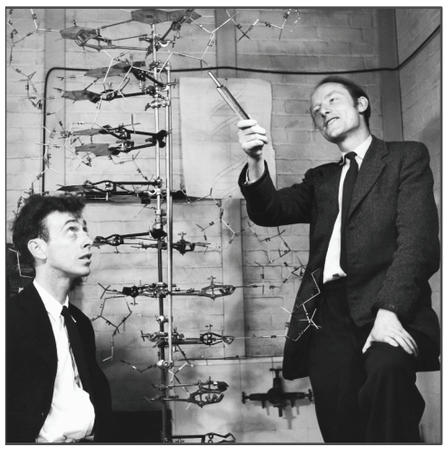


Understanding life based on molecular sequences.

5

Molecular Evolution

In the year 2000, then-President Clinton and Prime Minister Tony Blair jointly announced that the sequence of the human genome had largely been determined. There were a few missing parts to the sequence, but the string of nucleotides making up all 23 human chromosomes was largely known by the time of the announcement. Many journalists wrote this story up as if it was an unheralded thunderclap, a dramatic new turn in mankind's knowledge of itself.

But biologists had been working on genomic projects for years, often under the heading of "molecular evolution." The project of understanding life based on molecular sequences was conceived before 1950, and it received its greatest success with the development of the double-helix model for DNA by James Watson and Francis Crick in 1953. Rather than marking a beginning, the sequencing of the human genome was more the completion of a vision first glimpsed half a century earlier. Notably, Watson himself was the first to have guided the human genome sequencing project, before political enemies forced his resignation.

To understand the significance of the "genomic era," you need to understand molecular evolution. You need to learn what DNA sequences can, and cannot, tell you about life. You need some background concerning the information in genomic DNA—what is "junk," perhaps, and what is revealing.

We will cover all these topics in this chapter. First we will survey the overall structure of the genome, what it is made of, and which processes have contributed to its evolution. Then we show that much of gene evolution and genome evolution has been neutral, unimportant with respect to natural selection. In the final section, we deal with the role of selection in molecular evolution. By that point, you should begin to understand the true importance of sequencing the entire human genome. ❖

GENES AND GENOMES

5.1 The genome is not a huge library of information

DNA was established as the material of heredity in the 1940s and 1950s. The collection of all the DNA in the cell is called the **genome**. Until the 1970s, the common view of the genome was that it was a vast library (Figure 5.1A) of genetic information encoded by base pairs of DNA. Most multicellular animals and plants have billions of pairs of DNA nucleotides in each cell, enough for millions of genes. Therefore, biologists thought that there must be large numbers of genes, more than enough to specify physiological functions in great detail.

We now know that almost nothing about this view of the genome is correct. There is indeed a vast amount of DNA in many genomes, but most of that DNA does not code for amino acid sequences. This does not necessarily mean that the noncoding DNA is nonfunctional. Some of it is involved in the control of **genetic transcription**, the copying of the DNA sequence from the chromosome to *messenger RNA*. But any such information is secondary to protein encoding. Figures 5.1B and 5.1C show the contrast between the old and new views of genome structure.

The number of protein-coding genes is several orders of magnitude *smaller* than we used to think. Instead of millions of genes per genome, we know now that the number of genes ranges from a few thousand among bacteria to about 40,000 or less in vertebrates. Commonly studied genomes, like those of *Drosophila* or the nematode *Caenorhabditis* (Figure 5.1D), have 10,000 to 20,000 genetic loci. This is comparable to the number of parts in a modern car or airplane. In regard to gene numbers, genomes are extremely compact.

Methods for rapidly sequencing DNA were discovered in the 1970s. This finding led to a vast expansion of gene-sequencing activities. The first important result of this burst of sequencing activity was the discovery of DNA sequences within genes that did not code for amino acid sequences. Instead, as shown in Figure 5.1E, the noncoding segments of DNA are transcribed into **messenger RNA** (**mRNA**) and

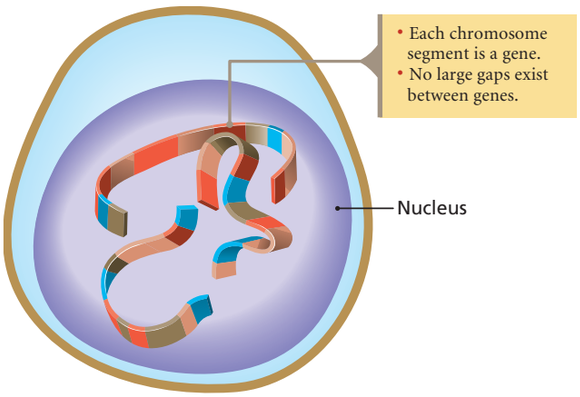


FIGURE 5.1B **Old View of the Genome, Still Accurate for Many Microbes**

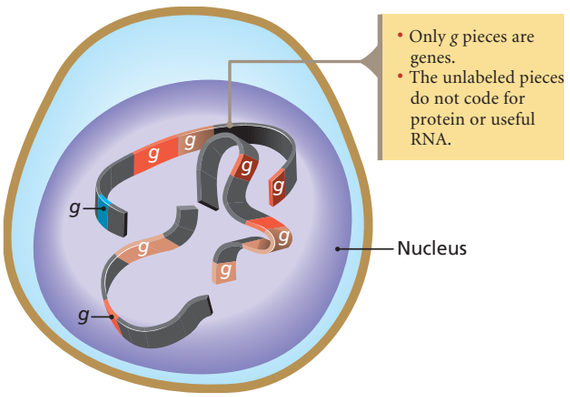


FIGURE 5.1C **New View of the Genome, Correct for Most Organisms, Especially Animals and Plants**

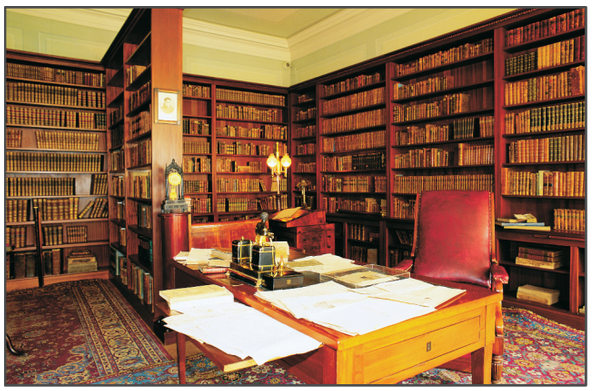


FIGURE 5.1A **A Library of Books**

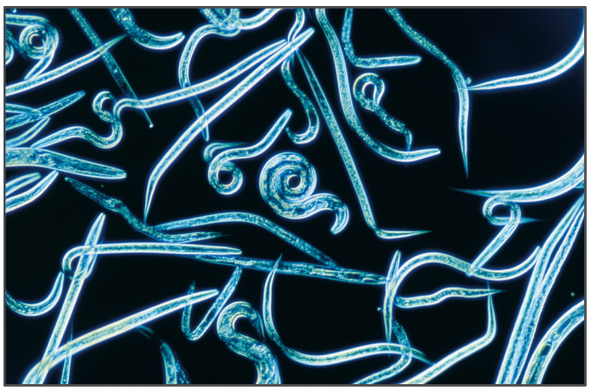


FIGURE 5.1D **The Nematode *Caenorhabditis elegans***

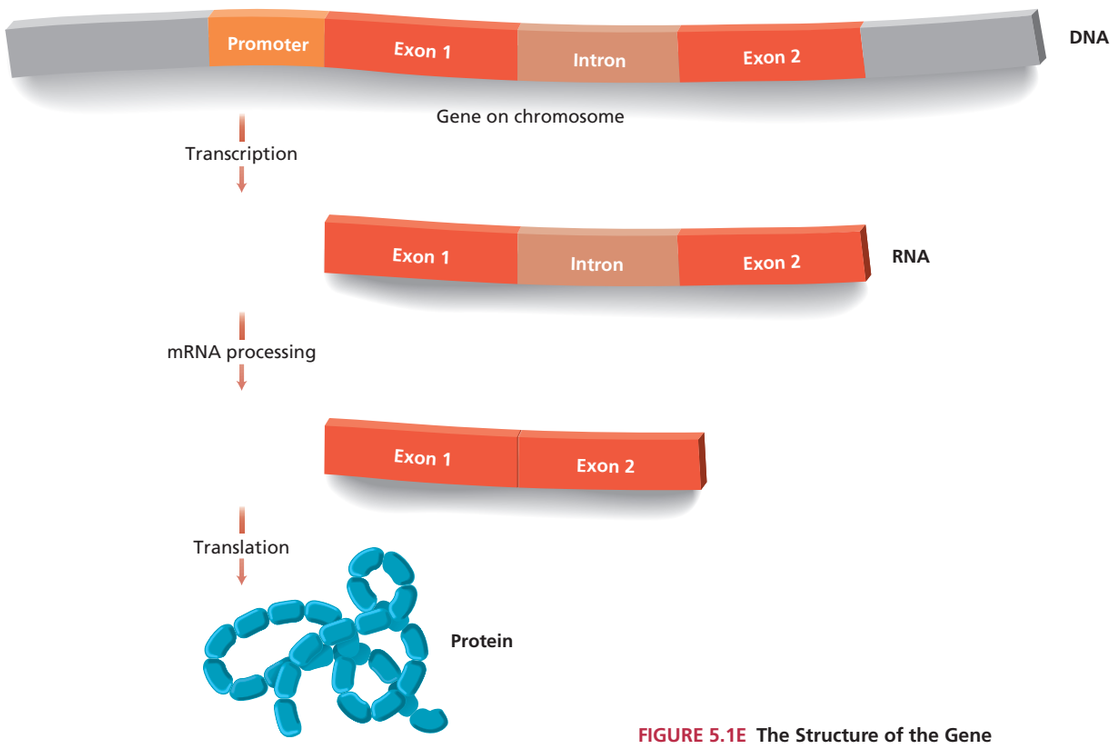


FIGURE 5.1E The Structure of the Gene

then excised from messenger RNA before it is used to guide **translation,** the assembly of the amino acid sequences of proteins from the coded instructions of the messenger RNA. The noncoding DNA sequences within genes are called **introns**. The coding DNA sequences are called **exons**.

Among the DNA found within introns and other noncoding regions were **transposable elements**—DNA sequences that move around within genomes. This movement does not follow any simple genetic rules (Figure 5.1F). Transposable elements are even thought to move from species to species.

Both introns and transposable elements were major anomalies for the old view of the genome as a well-organized library of functional information. They suggested that the "texture" of the genome was like Swiss cheese—full of holes and lacking in structural rigidity.

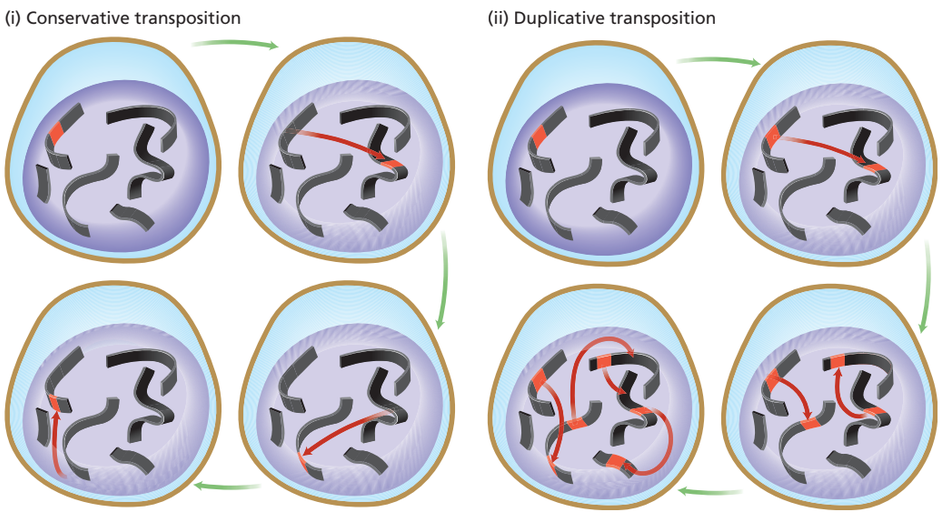


FIGURE 5.1F Two Types of Transposition With conservative transposition (i), the numbers of transposable elements does not increase. Duplicative transposition (ii) permits copy number increases.

5.2 The eukaryotic gene is a complex structure with many nucleotides that do not code for amino acids

Although introns are common features of eukaryotic gene structure, they have a number of features that are unlike coding DNA. Introns are not universal. Most of the genes of bacteria entirely lack introns. There is nothing about genetic function that appears to require introns. In itself, this tells us that introns are not functional parts of genes *in the same way* that coding sequences are.

The DNA flanking the exons of eukaryotic genes plays a functional role. Both the DNA that precedes the transcribed portion of genes and the DNA that follows influence the functioning of proteins. In short, this DNA has molecular-genetic regulatory functions. The DNA that comes just before genes plays an important role in determining the situations in which transcription occurs, such as starvation, development, aging, and so on. The most important regulators of transcription are DNA sequences called **promoters.** The DNA that follows exons also plays a role in the stability of the RNA transcript before it is processed to remove introns and then used for translation. The additional regulatory DNA sequences extend what can be considered the gene beyond the exons that code for amino acids.

Eukaryotic genes—comprising exons, introns, and flanking regulatory sequences—can be very long, containing thousands of nucleotides. In addition, related genes may be clustered together. Such gene clusters can have complex interactions, and the DNA flanking a gene cluster can have regulatory functions.

Introns are highly variable in their location. As shown in Figure 5.2A, the introns of the actin gene family are variable in their site. Organisms like *Saccharomyces pombe,* which is a close relative of brewer's yeast that reproduces by dividing in two, entirely lack introns in their actin genes.

Rates of intron evolution are very different from rates of exon evolution. Exons evolve at a rate of about one substitution per billion years per nucleotide. Introns evolve at a rate about ten times greater than that. This difference in rates of evolution suggests, to a first approximation, that the evolution of intron sequences proceeds at a rate determined either by genetic drift or by some rapid form of natural selection. We will consider this issue in more detail later in this chapter.

The origin of introns has been a source of argument. One theory is that introns are the residue of genetic reorganizations that brought together small exons. Furthermore, it has been proposed that these small exons might represent distinct functional elements, possibly ancient "proto-proteins." These small proto-proteins

FIGURE 5.2A Intron Positions of Actin Genes

may have functioned together with the amino acids encoded by other proto-proteins. Exons could be relics of these cooperating proto-proteins, assembled together later by recombination and selection. This is the *domain* model of intron evolution. This theory proposes that genes code for proteins that are combinations of simpler bits and pieces that come from ancient proteins.

The idea is somewhat as if your car were made of parts that came from different vehicles: the engine from one vehicle, the passenger compartment from a second, and the trunk from still a third. The construction of the car would then involve the assembly of these different elements.

In the case of genes, presumably they evolved when different exons were brought together within a larger gene structure, with the introns representing the relics of the DNA that used to separate the ancient genes long ago.

One of the biggest problems facing this theory is that fairly simple organisms, like bacteria, have very few introns. If ancient organisms were like bacteria in their gene structure, then this type of theory must be wrong. On the other hand, there is the possibility that the genomes of today's bacteria are evolutionary products of selection for removal of introns. ❖

5.3 Transposable elements are mobile genes that make copies of themselves to move about the genome

For decades, geneticists thought that the genes of a species occupied fixed positions on chromosomes, much as ancient astronomers thought that stars were stationary celestial bodies. Geneticists knew that genes changed positions as evolution changed one species into another, but this was regarded almost as a cataclysmic event, unlike normal genetics. One of the best pieces of evidence in favor of such a conservative view of genome structure was that the order of genes along pieces of chromosomes tended to be the same within groups of closely related species, such as species of the fruit-fly genus *Drosophila* or the great ape species—orangutans, gorillas, chimpanzees, and humans.

One of the first anomalies for this view of genes as static compartments of information came from studies of North American maize ("Indian corn") by Barbara McClintock, a pioneering American plant geneticist and Nobel Laureate. The kernels of native American-cultivated maize show extensive **variegation** in color pattern. Each kernel is genetically distinct. Some kernels are yellow, but others, on the same ear of corn, might be dark brown (Figure 5.3A). Further, this pattern was not predictable from one ear of corn to another—unlike normally inherited color patterns in most plants and animals. McClintock proposed that there were *controlling elements* that caused the variegation in the kernel color, elements that we now know are transposable elements: DNA sequences that move about the genome, making new copies of themselves and inserting themselves in sites that they did not previously occupy. In this case, these tranposable elements are inserting into pigment genes, disrupting their function.

Transposable element insertions are known to be selected against when they occur within exons. For example, the *white* locus of *Drosophila* is known to have suffered repeated exon insertion by transposable elements, all deleterious because they impair vision. (A white-eyed fly is shown in Figure 5.3B.) On the other hand, transposable elements that insert in introns appear to have fewer deleterious effects.

FIGURE 5.3A Corn, Showing the Effects of Transposition on Kernel Color

FIGURE 5.3B White-Eyed Male Fruit Fly This type of mutant can be produced by the insertion of a transposable element in the eye color gene.

Typical transposable elements code for **transposase**, a protein that allows the element to make new copies of itself and insert them in the genome at various locations. Sometimes transposable elements code for additional proteins that are also indispensable for their life cycle. Still other transposable elements code for proteins that are unrelated to the replication of the transposable element, such as proteins that help cells resist antibiotics.

Some transposable elements cannot produce transposase. Their transposition depends on the presence of transposable elements that still produce transposase. Because transposition causes frequent mutations, it is common to find that groups of transposable elements include passively transposing mutants that cannot transpose on their own, along with transposable elements that remain intact, as Figure 5.3C shows.

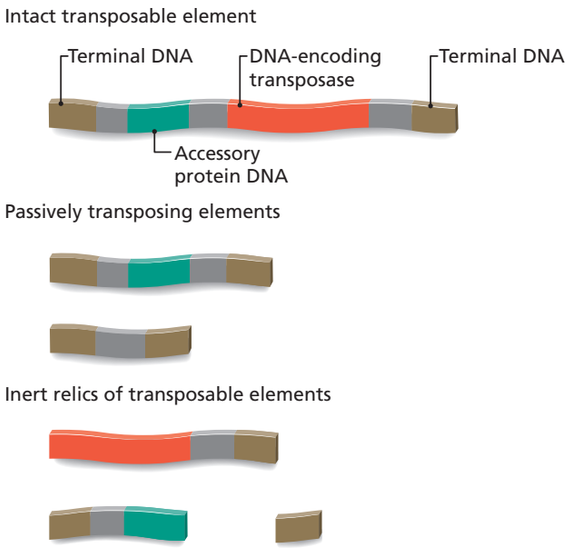


FIGURE 5.3C Polymorphism within a Single Class of Transposable Elements. Transposition requires transposase and both terminal DNA sequences. Elements lacking their own transposase may get it from another transposable element.

There are distinct transposable element life cycles, two of which are displayed in Figure 5.3D. **DNA-based transposition** is shown in part (i) of the figure. Chromosomal DNA is copied by DNA replication to form extrachromosomal DNA. Some of the extrachromosomal transposable element DNA is then inserted at a new site in the genome.

A second class of transposable elements is made up of the **retrotransposons,** shown in part (ii) of Figure 5.3D. These elements, also known as **retroposons**, reside in the genome as DNA. But their replication requires transcription and the formation of an RNA intermediate, as part (ii) of Figure 5.3D shows. This RNA intermediate is then used to guide the synthesis of the corresponding DNA sequence, using the protein **reverse transcriptase.** Reverse transcriptase may be incorporated in the retrotransposon undergoing reverse transcription, or it may come from another transposable element. The DNA produced with the help of the reverse transcriptase is then incorporated in the host genome. Some of these elements consist of little more than a promoter sequence for transcription and flanking sequence information for incorporation of the reverse transcribed DNA back into the genome. An example of this type of element is *Alu I*, which is present in humans in hundreds of thousands of copies in each of our nucleated cells. ❖

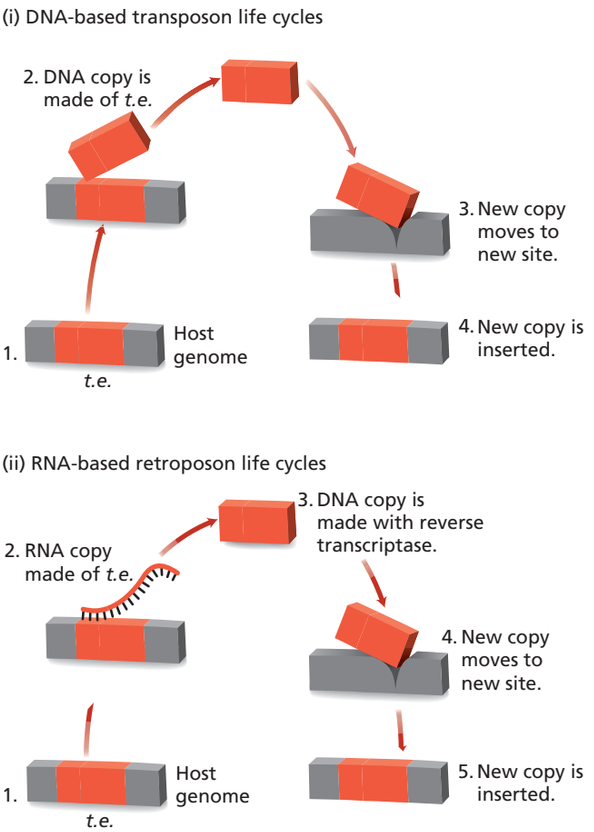


FIGURE 5.3D Life Cycles of Transposable Elements (*t.e.*)

5.4 Tandem arrays of genes increase and decrease gene number by unequal crossing over

Because of its double-helix structure, DNA has an inherent tendency to recombine. However, this tendency to recombine is physically limited by sequence similarity. Dissimilar DNA sequences are unlikely to recombine with each other, because their helices will not pair well. But similar DNA sequences tend to pair. This produces normal recombination during meiosis, in which homologous alleles are swapped between chromosomes due to pairing of helices from different chromosomes.

But the situation is different when DNA sequences are arrayed *in tandem*, with repeated sequences following one after another along a chromosome, as shown in Figure 5.4A. Many such **tandem arrays** of DNA sequences are known to occur in nature. These may be simple sequence repeats, or they may be repeats of sequences that code for proteins or ribosomal RNA, as shown in Figure 5.4A.

The excitement comes from the things that DNA can do when sequences occur in tandem arrays. Figure 5.4B shows that DNA sequences in tandem arrays can be misaligned during meiosis, before recombination occurs. With this misalignment and recombination, which is called **unequal recombination** or **unequal crossing over**, one of the two products of the recombination event will contain more copies of the repeated sequence, while the other recombination product will have fewer copies of the repeated sequence.

At this point there are two main possibilities:

1. The repeated sequence is of no functional significance. Then the number of copies of a repeated sequence may rise or fall with accidents of unequal recombination. This is a drift process like that of genetic drift, involving alternative alleles at a normal Mendelian locus, as described in Chapter 3.

2. The repeated sequence is a functional genetic sequence, such as a genetic locus that codes for a protein, or a sequence that is used to transcribe the RNA components of a ribosome. Then natural selection will oppose unequal recombination that leads to a lack of the functional sequence. It may also select against high numbers of the sequence, perhaps because too much production of a protein results. This will give rise to *stabilizing selection*, described in Chapter 4, in which an intermediate copy number would be the state favored by selection.

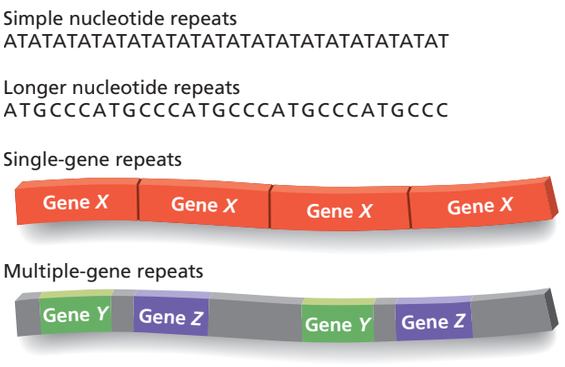


FIGURE 5.4A Tandem Array Structures

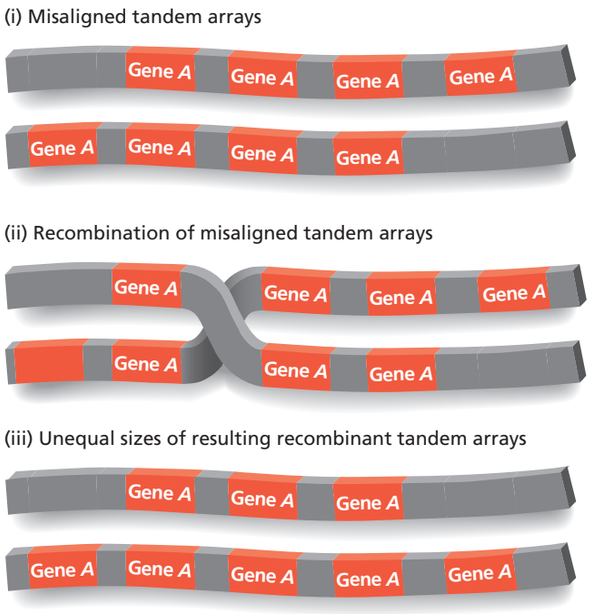
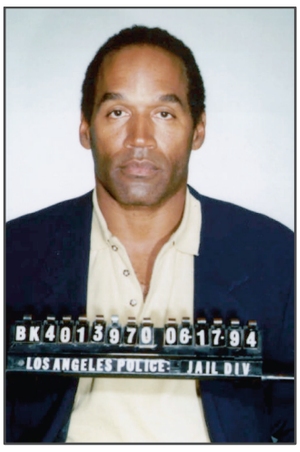


FIGURE 5.4B Unequal Crossing Over in Tandem Arrays



Tandemly repeated DNA sequences have many applications in genetics. They are used as highly variable markers in molecular ecology, particularly to determine who is mating with whom in natural populations. They are also used in *genetic fingerprinting*, to distinguish between suspects in criminal cases, particularly rape, where the perpetrator almost always leaves his genes behind (see Figure 5.4C).

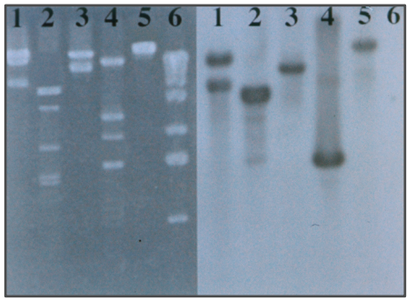


FIGURE 5.4C Forensic science routinely uses tandem array DNA as evidence in criminal cases.

Prokaryotic genomes are concatenations of genes with occasional inserted sequences, whereas eukaryotic genomes have large intergenic regions that play no apparent role in gene replication or function

5.5

In the prokaryotes, especially the bacteria, genes are closely packed together, with little intergene DNA. There are also few introns. This genome can be thought of as maximally efficient in the use of DNA. The **prokaryotic genome** is a compact compendium of genetic loci with the occasional transposable element inserted here and there. In such compact genomes, the evolution of the genome is not that different from the evolution of many individual genes combined. Indeed, such genomes are often largely free of introns, making them an even tidier story. This genome structure is sketched in Figure 5.5A.

Some eukaryotes, such as some yeast species, also have very compact genomes. Like prokaryotes, such unicellular eukaryotes have little DNA between genes. They also tend to have relatively few introns. Again, the genome is very compact.

The existence of compact eukaryotic genomes shows that there is no functional requirement for abundant DNA between genes. Nor is there any apparent requirement for introns. These prominent features of genome structure in most eukaryotes appear to be dispensable, at least for some microbial organisms.

Unlike yeast, most eukaryotes have the the type of genomes shown in Figure 5.5B, with large regions of DNA between genic regions, abundant introns, transposable elements, and so on. Most of the DNA of animals and plants has no protein-coding function. The genomes of such organisms are usually a sprawling affair. For long stretches of DNA, there are no genes at all. One way to understand the difference between these genome structures is to think of bacterial genomes as villages, yeast genomes as small cities, and most eukaryotic genomes as megalopolises like Los Angeles (Figure 5.5C).

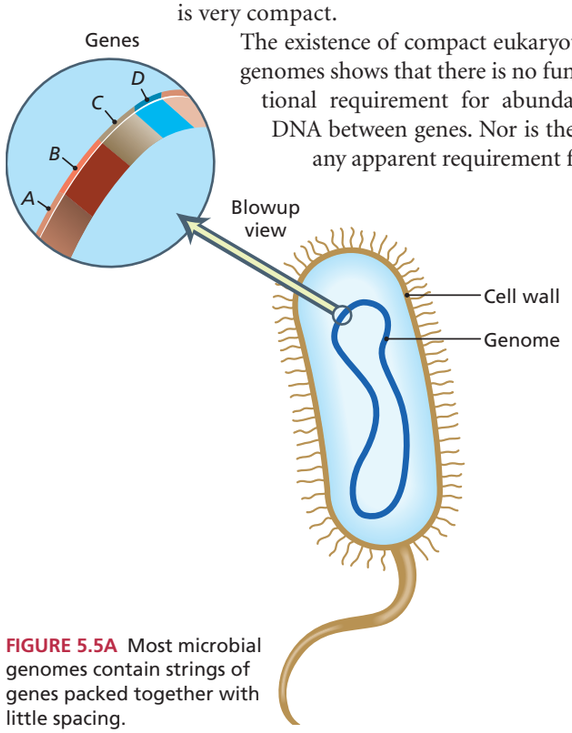


FIGURE 5.5A Most microbial genomes contain strings of genes packed together with little spacing.

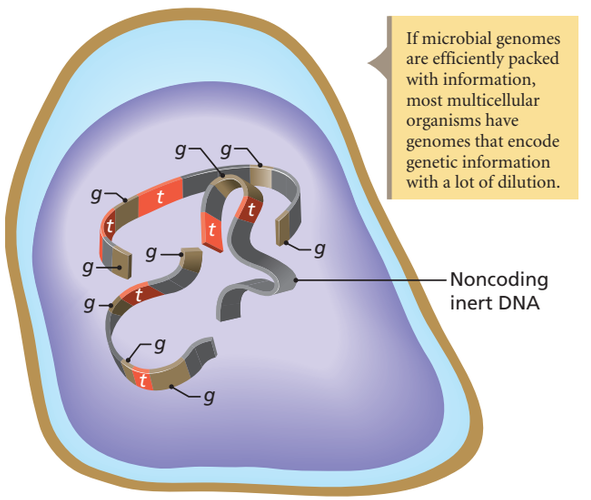


FIGURE 5.5B Big eukaryotic genomes have a lot of noncoding intergenic DNA. There is no clear functional purpose for much of this DNA.

FIGURE 5.5C Several Views of Los Angeles

Neutral Molecular Evolution

5.6 The neutral theory of molecular evolution is based on genetic drift

In the 1960s, Motoo Kimura and a few other evolutionary theorists proposed that the explanation for molecular evolution was not selection, but genetic drift. They proposed that segregating molecular genetic variation is not important for natural selection. This **neutral theory of molecular evolution** was at first discussed in fairly extreme terms. Many argued against it on the grounds that evolution is full of adaptations, which must therefore be products of natural selection. There were two weaknesses to this criticism. The first weakness was that adaptations can be produced by evolutionary processes other than natural selection specifically favoring their evolution, perhaps as a result of natural selection for very different features. For example, there is some evidence that feathers, which are adaptations for flight, were first produced by selection for some other function in reptiles, perhaps temperature regulation.

The second weakness of this criticism is that evolution has multiple levels. Most molecular variation may have no effect on the organism's phenotypes, even if there is other genetic variation that does affect the organism's phenotypes. This second type of genetic variation can then be shaped by natural selection, even if the first type is not. There is no incompatibility between the neutral theory of molecular evolution and the action of natural selection in adaptive phenotypic evolution.

It is difficult to predict whether or not particular proteins will be favored by natural selection. At the level of nucleotides, however, it is somewhat easier to say in advance how natural selection will be structured, because of the considerable difference between nucleotides in their roles within the genome. There are two levels to this problem: one is the individual gene and the other is the entire genome. Let's consider each in turn.

Figure 5.6A shows a gene, with different regions labeled according to their coding and regulatory roles. The sequences in the middle of the introns are usually free of natural selection—unless a particular intron sequence disrupts the excision of the intron during the processing of the initial RNA transcript, in which case that intron sequence would be selected against. In exons, the third position of some codons is free to vary, because some nucleotide changes at this position are synonymous, as shown in Table 5.6A. For example, all RNA triplets (codons) that start with the sequence UC– code for the amino acid serine. The third nucleotide doesn't matter, in this case.

However, in many cases where third position nucleotides code for the same amino acids, they do not occur at uniform frequencies. Instead, there may be a great preponderance of a particular triplet. This is called **codon use bias.** There is no generally agreed explanation for it. In some cases, it may be an unlikely product of genetic drift. In other cases, it may reflect some distortion in the biochemistry of nucleotides.

An additional possibility for neutral genetic variation arises when nucleotide substitutions result in the use of a different amino acid that is effectively equivalent, perhaps because of similar structure, to the original amino acid.

Far from genes, genetic drift can act with impunity. Note, however, that in such regions there may be transposable elements that evolve on their own, subject to their own natural selection for effective spread through the genome. In addition, there may be stretches of simple repeats, such as ATATATATAT, that expand and contract with unequal crossing over. Therefore, even the seemingly lifeless expanses between genes may evolve by processes more complex than genetic drift on its own.

TABLE 5.6A Synonymous Substitutions

The molecular code is degenerate. More than one triplet of RNA nucleotides may code for the same amino acid.

RNA Triplets	Amino Acid
UUU, UUC	Phenylalanine
UUA, UUG, CUU, CUA, CUG	Leucine
AUU, AUC, AUA	Isoleucine
AUG	Methionine
GUU, GUC, GUA, GUG	Valine
UCU, UCC, UCA, UCG, AGU, AGC	Serine
CCU, CCC, CCA, CCG	Proline
ACU, ACC, ACA, ACG	Threonine
GCU, GCC, GCA, GCG	Alanine
UAU, UAC	Tyrosine
UAA, UAG, UGA	STOP
CAU, CAC	Histidine
CAA, CAG	Glutamine
AAU, AAC	Asparagine
AAA, AAG	Lysine
GAU, GAC	Aspartic acid
GAA, GAG	Glutamic acid
UGU, UGC	Cysteine
UGG	Tryptophan
CGU, CGC, CGA, CGG, AGA, AGG	Arginine
GGU, GGC, GGA, GGG	Glycine

DNAs also evolve by genomic processes like duplicative transposition (described earlier) and gene duplication by unequal crossing over (described earlier). Many of these events may also be free of natural selection, or close to free from it, especially when they occur in intergenic regions.

There is now little doubt that the evolution of many DNA sequences is not directly determined by the action of natural selection. Instead, it is widely agreed that mutations to DNA sequences often have no effect on the phenotype of the organism, especially because a great deal of the DNA of eukaryotic organisms has no role in determining either amino acid sequences or the regulation of gene expression. When a new molecular variant of no selective significance arises, it is likely to be lost accidentally almost immediately. If that does not happen, the variant molecule will fluctuate in frequency for a time, producing molecular polymorphism in the population. But this polymorphism will have no selective significance. Finally, some neutral DNA sequence variants may rise to fixation in the population, in an accidental *substitution*. Then the population will regain polymorphism only once a new mutation has occurred.

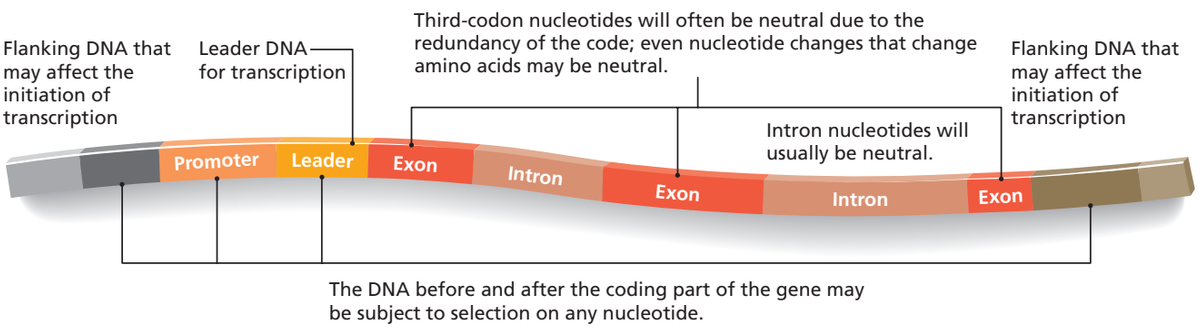


FIGURE 5.6A Gene Structure and Neutrality

5.7 The molecular clock is based on the observation that the rate of molecular evolution is roughly constant

In the 1960s, the first data on the amino acid sequences of proteins were published. These data were collected from several different species, especially mammals. The kinds of protein that were studied included hemoglobins and cytochrome *c*. Having the amino acid sequences for the same protein in different species naturally led scientists to look at the relationship between the time since the species last had a common ancestor, called **divergence time**, and the number of fixed amino acid differences between any two such species, called the **number of substitutions.**

This finding was puzzling because evolutionary processes scale with the number of generations, not elapsed chronological time.

Emil Zuckerkandl and **Linus Pauling** (Figure 5.7A) pointed out that the number of substitutions per unit of time seems to be roughly constant. There appeared to be a **molecular clock,** which recorded the passage of time by substitutions of amino acids. This finding was puzzling because evolutionary processes scale with the number of generations, not elapsed chronological time. Many species have generation times much less than a year, or much more. The species used in an analysis of molecular evolution might have very different generation times: rodents and apes, for example. The time unit of calendar years was used in studies of molecular evolution anyway.

A further anomaly arises with multiple amino acid substitutions. An observed difference at a particular amino acid site might have occurred after a sequence of several amino acid substitutions at that site, though only a single difference would be detected in the comparison of two species at that site. Yet even with these problems of timescale and multiple substitutions, the data for molecules like hemoglobin often follow a linear pattern, with amino acid differences accumulating in a clocklike pattern with time.

For a more evolutionary understanding of the molecular clock, look at Figure 5.7B. It is important to bear in mind that evolutionary divergence is a dual process: Two distinct evolutionary lineages are undergoing genetic substitutions through evolutionary time. Therefore, a correct estimate of the **rate of evolutionary divergence** is *not* the ratio of substitutions (K) over time (T), or K/T. Rather, the correct estimate of the rate of evolutionary divergence is as follows:

$$r = K/(2T)$$

There is twice as much evolutionary time as the total time since the last common ancestor suggests, because both of the descendant species diverge from the evolutionary state of the common ancestor.

An interesting scientific maneuver is to use the rate equation to estimate divergence times. If we assume that the rate of protein evolution is constant, then we can use the total number of substitutions to estimate the evolutionary time separating two species. This is done by rearranging the previous formula to obtain

$$T = K/(2r)$$

In one of the most important scientific applications of the molecular clock concept, in 1967 Vince Sarich and Allan Wilson of the University of California, Berkeley, applied this calculation to the evolution of primates. They arrived at the remarkable estimate of 5–8 million years since the last common ancestor

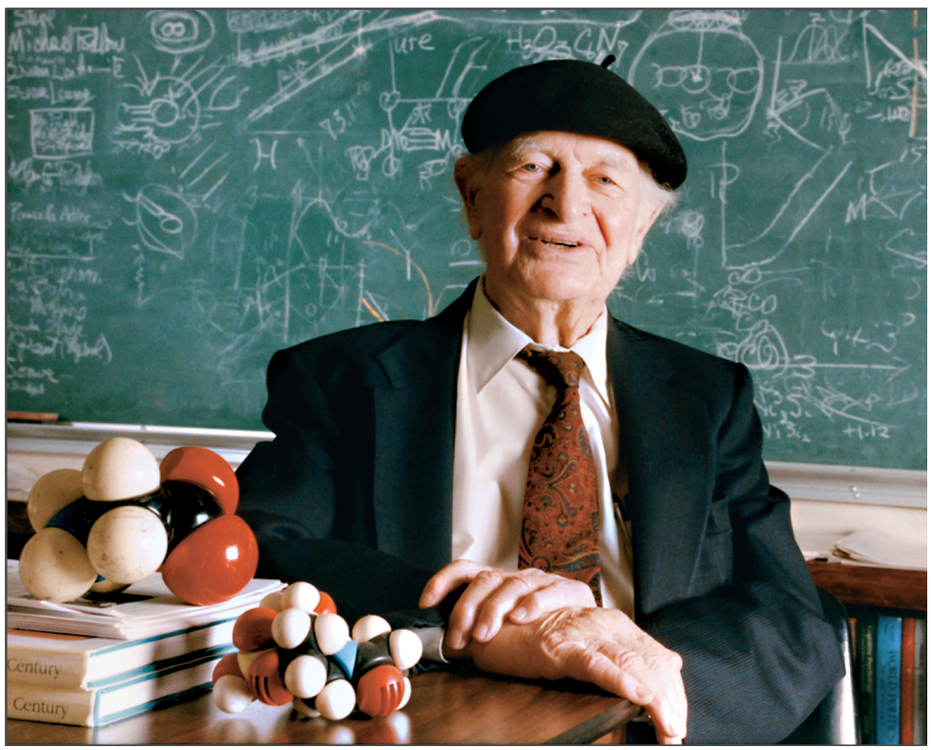


FIGURE 5.7A **Linus Pauling, Nobel Laureate**

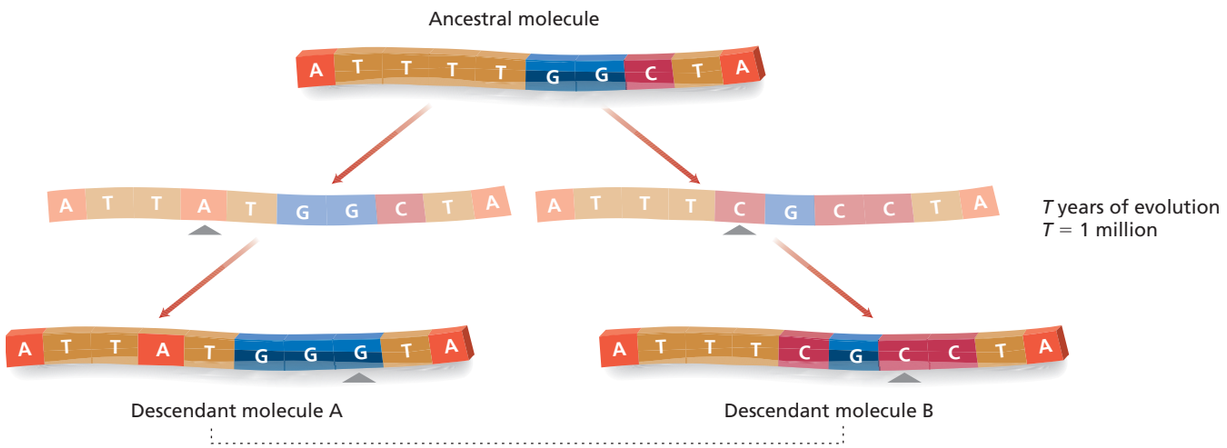


FIGURE 5.7B **The Concept of Molecular Divergence**

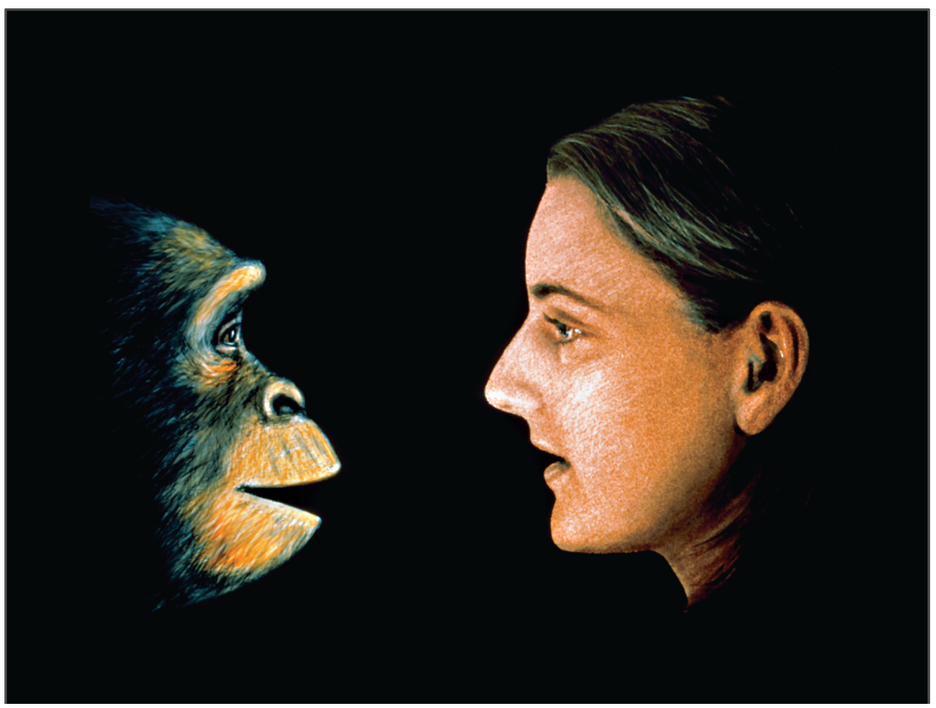


FIGURE 5.7C Chimpanzees are our closest living evolutionary relatives.

between chimpanzee and human (Figure 5.7C). At that time, the consensus in the paleoanthropological community was that the common ancestor of chimp and human lived about 20 million years ago.

For this reason, the paleoanthropologists strongly rejected Sarich and Wilson's argument. However, 30 years of further research have led to considerable reductions in nonmolecular estimates of the time since chimps and humans last had common ancestors. (This is discussed further in Chapter 21.) Nonmolecular and molecular clock estimates of the time of divergence separating chimps and humans are now roughly similar.

5.8 Unlike nonsynonymous substitutions, synonymous substitutions proceed at a fairly uniform rate across a wide range of DNA sequences

To set a uniform standard for the molecular clock, scientists wanted data on molecular evolution that would not be affected by natural selection and other variable evolutionary processes. With that in mind, they obtained data on **synonymous DNA substitutions**, changes in DNA sequences that do not change the amino acid composition of proteins. Such synonymous substitutions arise because the genetic code is *redundant:* more than one triplet of RNA codes for most amino acids. For example, serine is coded for by six different RNA triplets: UCU, UCC, UCA, UCG, AGU, and AGC. The coding for serine is thus highly redundant. Tryptophan, on the other hand, is coded for by a single RNA triplet, UGG. It has no redundancy at all. (Note that uracil replaces thymine in RNA molecules, which accounts for the "U" symbol in Table 5.6A, giving the genetic code.) This redundancy seems to allow the evolution of some DNA nucleotides to proceed without any influence from natural selection. However, this assumption depends on the cell using each of the alternative codons uniformly, without bias. This isn't always true. However, synonymous substitutions are far more likely to be equivalent to each other in their phenotypic effects than are nonsynonymous ones.

To set a uniform standard for the molecular clock, scientists wanted data on molecular evolution that would not be affected by natural selection and other variable evolutionary processes.

The scientific interest is this: If DNA evolution proceeds in a clocklike fashion when there are no effects on protein evolution, we should find that the number of synonymous substitutions is uniform, across evolutionary time and among different proteins. Synonymous substitutions should give us the most clocklike data for the process of molecular evolution. Figure 5.8A shows the rates of synonymous substitution among a group of common vertebrate proteins. To a reasonable extent, these rates are uniform: 3.5 to 6.5 substitutions per billion years. Therefore, if we use the number of synonymous substitutions separating two species for these proteins, and an evolutionary rate of about five substitutions per site per billion years, then we should be able to estimate the evolutionary time of divergence fairly accurately. This is probably the most reliable kind of molecular clock to use.

One way that we can test for natural selection in molecular evolution is to compare **nonsynonymous substitution** rates with synonymous rates. Nonsynonymous substitutions involve changes to DNA sequences that *do* change the amino acid sequences of proteins. If amino acid sequences are subject to natural selection, then we expect to find more heterogeneity among rates of nonsynonymous substitutions, as compared to the clocklike rates of synonymous substitutions. What do we actually observe?

Figure 5.8B shows some of the heterogeneity for substitution rates in some common vertebrate proteins, the same proteins that were used to estimate synonymous substitution rates. The rates of substitution are far more heterogeneous for nonsynonymous substitutions—that is, for the DNA changes that result in changes in amino acid sequences. Therefore, even if DNA evolution is fairly uniform when it

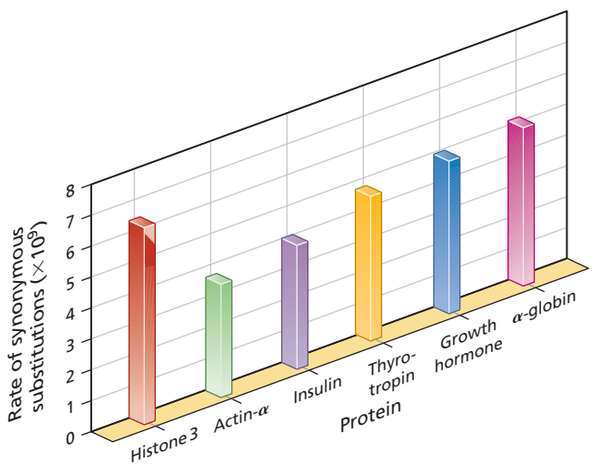


FIGURE 5.8A **The Rate of Synonymous Substitutions for Genes of Different Proteins**

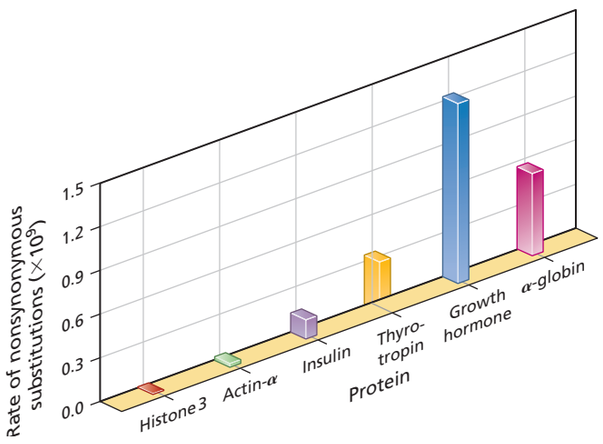


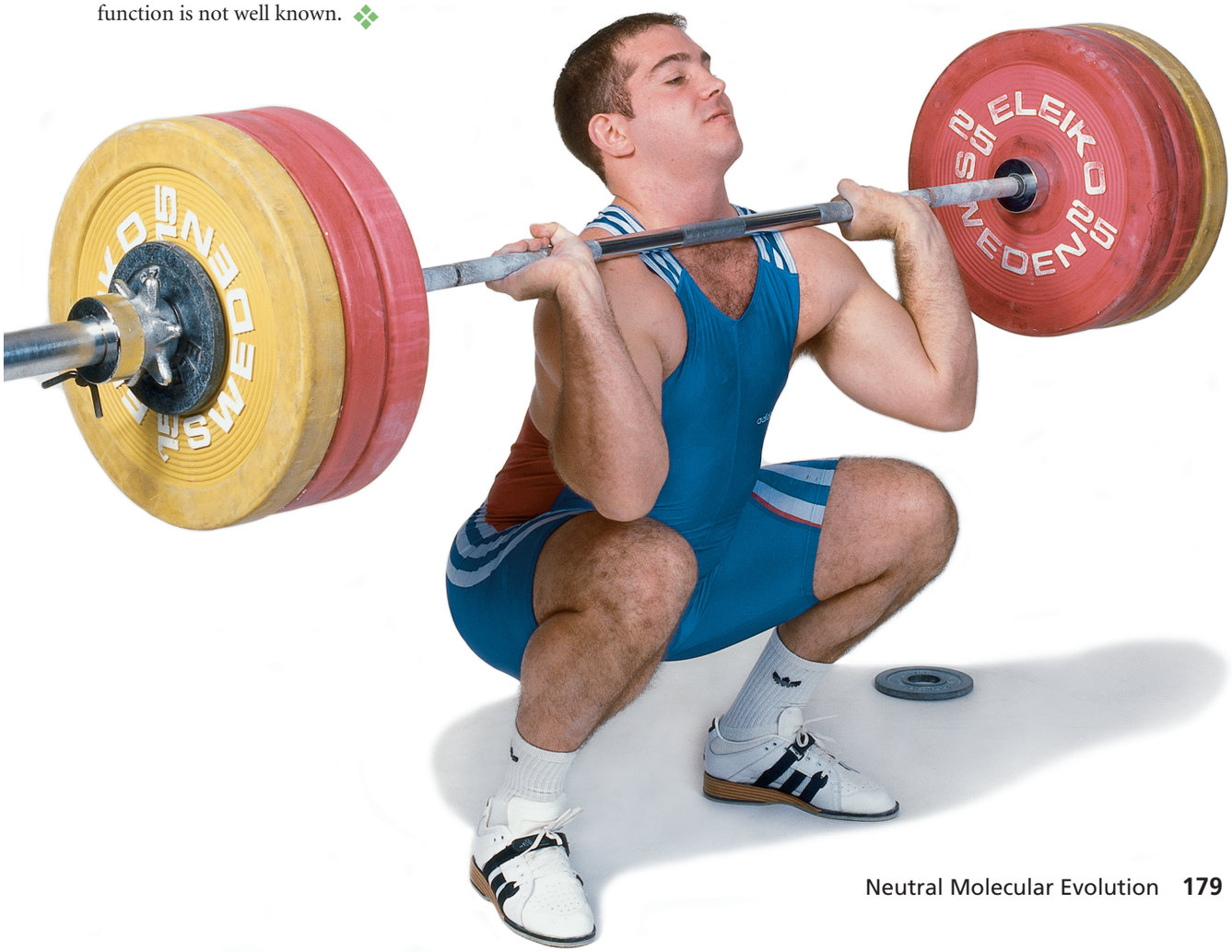
FIGURE 5.8B **The Rate of Nonsynonymous Substitutions for Genes of Different Proteins**

does not make any difference to amino acid sequence (as in the case of synonymous substitutions), DNA evolution is *not* uniform when it changes amino acid sequence. Therefore, the overall pattern of nonsynonymous DNA evolution may involve natural selection.

Is there anything predictable about the rates of evolution of different proteins? One generalization is that the "role" of a protein sometimes suggests what its rate of evolution will be. Consider the evolution of two very different types of protein: structural proteins and the proteins of the immune system. Structural proteins have to fit mechanically with other proteins or cell constituents. For example, **histones** are used to package DNA. **Actin** is a structural part of muscle fibrils. We would expect such structural building blocks to evolve slowly, because they literally have to fit other molecules. As Table 5.8A shows, they do evolve slowly. **Immunoglobulins,** on the other hand, are part of the vertebrate immune system, which generates random variation in antibodies. We would expect some selection for rapid evolution of immunoglobulin amino acid sequence, because this protein must continually evolve in response to the challenge posed by new foreign molecules. And rapid evolution is what we observe, as Table 5.8A shows. In these extreme cases, we can make reasonable guesses about relative rates of protein evolution. However, such guessing is not always so easy, particularly when we are considering the evolution of proteins whose function is not well known. ❖

TABLE 5.8A Nonsynonymous Substitution Rates in Molecules of Different Types

In mammals, times 10^9	
a. Structural molecules that "fit" closely	
Histone 3	0.0
Histone 4	0.0
Actin α	0.01
Actin β	0.03
b. Immunoglobulins	
Ig V_H	1.07
Ig γ1	1.46
Ig k	1.87



Selective Molecular Evolution

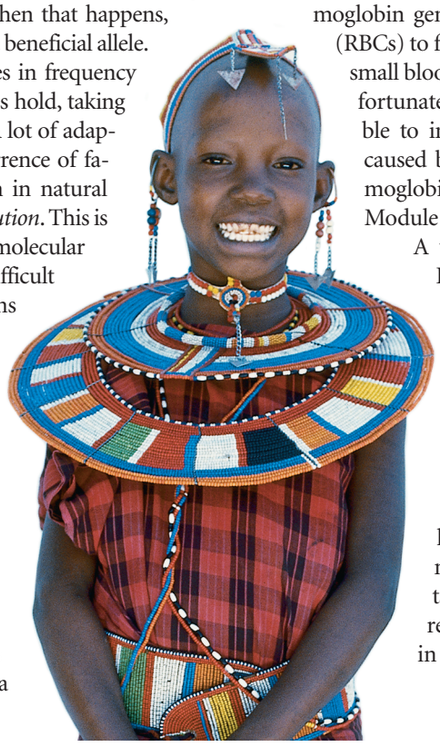
5.9 Natural selection eliminates, substitutes, and maintains specific molecular genetic variants

As described in Chapter 4, natural selection acts on three kinds of genetic variants. The first kind consists of all those genetic variants that are clearly inferior to normal alleles. These inferior alleles undergo purifying selection and are usually eliminated (see Module 4.16). These inferior genetic variants are probably the second most common type of new mutation, after neutral mutations.

The second kind of genetic variant is the class of favored alleles. These variants may be lost due to accidents of sampling, as shown in Figure 5.9A. (Even if an organism has the best genotype, it may still die accidentally.) When that happens, natural selection has failed to recruit a beneficial allele. Otherwise, the favored allele increases in frequency enough so that natural selection seizes hold, taking the favored allele to virtual fixation. A lot of adaptive evolution has involved the occurrence of favorable mutations and their fixation in natural populations, which is a type of *substitution*. This is how many adaptations evolve, even molecular adaptations. Nonetheless, it is often difficult to know which nucleotide substitutions have been selectively favored.

The third kind of genetic variant is made up of those alleles that are not always favored in all genetic combinations. Instead, these alleles are beneficial only in special genotypic combinations. One example of this pattern occurs when genotypes containing two different molecular variants have a fitness advantage over genotypes that have only one of these two variants. This might occur, for example, when a molecule composing the vertebrate immune system leads to more diverse antibodies when it is coded for by two distinct genetic variants, from the same locus. This is *overdominance*, introduced as *heterozygote advantage* in Chapter 4. At the molecular level, an interesting effect of overdominance is that it will foster molecular genetic variation—in principle, at least—as shown in Figure 5.9B. But in practice, convincing examples of overdominance have been very hard to find.

An example of overdominant selection that has already been described is the evolution of the hemoglobin molecule. The hemoglobin genes of northern Europe allow red blood cells (RBCs) to form without sickling. Such RBCs pass through small blood vessels, such as the capillaries, with ease. Unfortunately, these blood cells also leave people vulnerable to infection with malaria, a blood-borne disease caused by a parasitic trypanosome, *Plasmodium*. Hemoglobin evolution is discussed in more detail in Module 4.25.



A variant of the hemoglobin gene causes the RBC to deform. The RBCs take on a sickled shape when two copies of this gene are present, in homozygous combination. This shape makes it difficult for these RBCs to pass through small blood vessels, causing circulatory problems and eventually death.

The heterozygote that combines the alleles for the two kinds of hemoglobin has occasional sickling, but it does not usually cause health problems. The single sickling gene makes it harder for the malaria parasite to establish itself in the circulatory system. For this reason, the heterozygote has the greatest fitness in regions of the world afflicted with malaria.

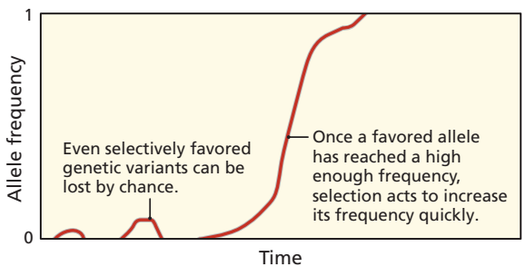


FIGURE 5.9A Evolution of Selectively Favored Genetic Variants

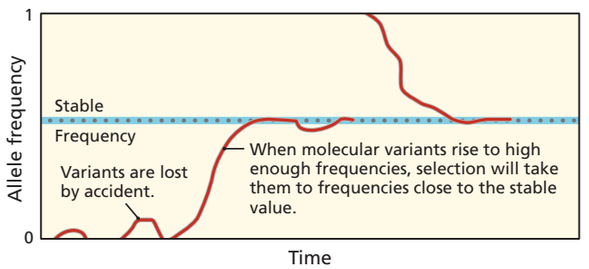


FIGURE 5.9B Evolution with Selectively Balanced Polymorphism

It is uncertain how much nucleotide evolution is due to selection, but there is some evidence for selection on particular nucleotides 5.10

In the 1960s a technique was developed to study variation in the amino acid composition of proteins: **protein electrophoresis.** Electrophoresis involves running a current through a gel, then adding proteins to one end of the gel and letting them migrate through the gel for a fixed period of time. The proteins are then stained using a chemical reaction specific to each type of protein. Usually, proteins that have different amino acid compositions migrate to different points in the gel, as Figure 5.10A shows. This allows geneticists to identify at least some of the variant proteins produced by different organisms from a population. Protein electrophoresis was the first relatively unbiased technique that population geneticists had to study genetic variation in natural populations.

When they finally came, the first data on molecular genetic variation were a shock to almost all evolutionary biologists, even the ones who collected the data themselves. There was a vast amount of genetic variation in the organisms first studied—humans and fruit flies of the genus *Drosophila*. Later work, with other organisms, confirmed this pattern. A few species had little genetic polymorphism. Among this monomorphic group, inbred species like self-fertilizing nematodes and cereals were common. But most outbreeding species had large amounts of genetic variation.

How often is selection involved in maintaining molecular genetic polymorphism? One of the best-studied examples is amino acid polymorphism at the alcohol dehydrogenase (*adh*) locus of *Drosophila melanogaster*. In humans this locus is responsible for metabolizing alcohol, so that we aren't too drunk. There is some evidence that it plays a similar role in the metabolism of alcohol in fruit flies, but it may do other things as well. We don't know.

It has been known for some time that protein electrophoresis detects a protein polymorphism involving two common variants at this genetic locus. One of these variants codes for threonine in exon 4 of the gene, as indicated in Figure 5.10B, while the other variant has lysine in the corresponding site in the protein. This polymorphism is found among *D. melanogaster* populations throughout the world. That it is likely to be subject to selection is also suggested by a north-south gradient in allele frequencies. It is notable that this gradient reverses direction in the Earth's Southern Hemisphere, compared to the Northern Hemisphere. Furthermore, it is known that fruit flies disperse rapidly up and down these gradients. These gradients are therefore unlikely to be ancient relics of migration patterns. Some type of selection must be involved.

The problem is that we do not know what the focus of selection is on *adh*. However, there is every sign that the locus is undergoing selection, selection that maintains genetic variation. This case is an interesting challenge for the next generation of evolutionary biologists.

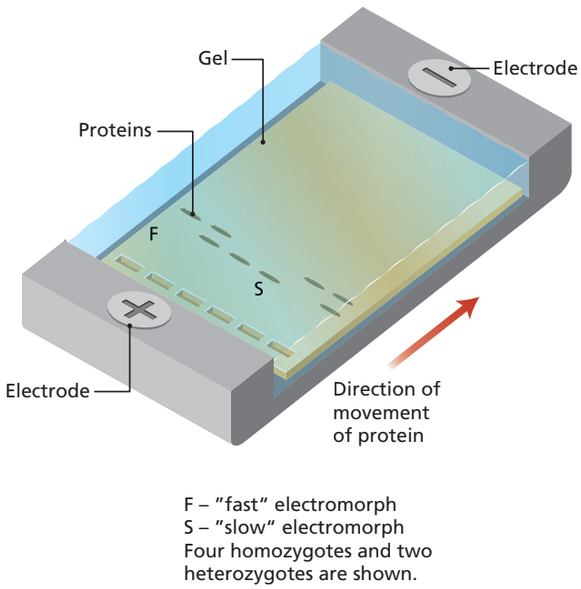


FIGURE 5.10A **Protein Electrophoresis**

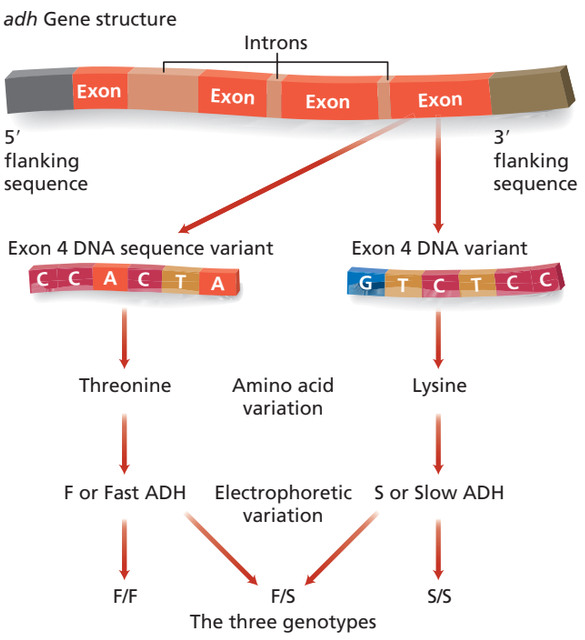


FIGURE 5.10B **Genetics of Alcohol Dehydrogenase Polymorphism in *Drosophila melanogaster***

5.11 Genes that have been duplicated by reverse transcription may degenerate or evolve new functions

Retrotransposons have a remarkable impact on genome evolution over and above their proliferation within genomes. Because they supply genomes with strong promoters for transcription, as well as reverse transcriptase, they can cause the duplication of genes.

There are several steps in this process, shown in Figure 5.11A. The first is the over-transcription of genes that are located just after retrotransposons. The transcriptional machinery may continue creating mRNA for some distance after the retrotransposons, if the element lacks sequence that would stop transcription. This may cause an overabundance of RNA transcripts from this region of the genome, including complete transcripts of the downstream genes. This can happen because retrotransposons are not normal functional genes, and therefore will not be selected to regulate the transcription that they stimulate.

The transcriptional machinery may continue creating mRNA for some distance after the retrotransposons, if the element lacks sequence that would stop transcription.

The next step is that the transcribed RNA is edited, removing introns and other extraneous sequences. At this point, the retrotransposon and the downstream gene(s) may be separated. However, let us assume that they are not. The downstream gene(s) are now physically linked with a retrotransposon structure.

The third step is that the retrotransposon and the downstream gene(s) are reverse transcribed back into DNA. Again, the retroposon and the gene may become physically separated at this point.

The fourth step is the reincorporation of the reverse-transcribed DNA, both retrotransposon and regular gene(s). By this point in the process, you can see that the genome size has been increased. Both the retroposon and the downstream gene(s) have made new copies of themselves in the genome. This is like any transposable element, which as a class have the capacity to make many copies of themselves.

But there is a further consequence. There is now a new gene, or genes, in the genome. Its evolution will proceed in one of two directions. The first occurs when there is no useful promoter of transcription located before the gene. As an untranscribed genetic element, which cannot transpose on its own, the new gene is irrelevant where natural selection is concerned. Mutations to its DNA sequence can accumulate, including mutations that stop transcription or translation. Such genes are dead genes, or **pseudogenes.** These genes are detectable by several diagnostic features: close similarity to the exons of another gene, absence of introns due to the processing of the gene as mRNA, and the accumulation of codons that interrupt transcription or translation. Figure 5.11B contrasts a normal gene with a processed pseudogene. Animal and plant genomes are littered with these dead or dying genes.

A new gene takes the second evolutionary direction much less frequently. With considerable rarity, reverse-transcribed genes may reinsert in the genome near an active promoter for transcription—possibly a promoter contributed by a retrotransposon. In this case, the reverse-transcribed gene may still produce a protein, and it may be a target of natural selection.

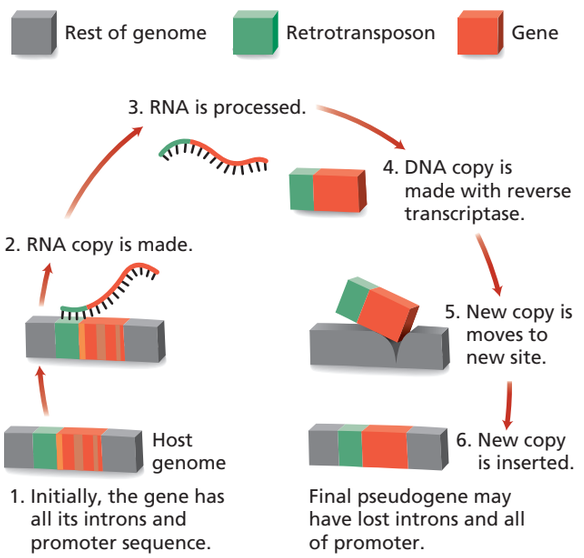


FIGURE 5.11A **How Pseudogenes Are Made**

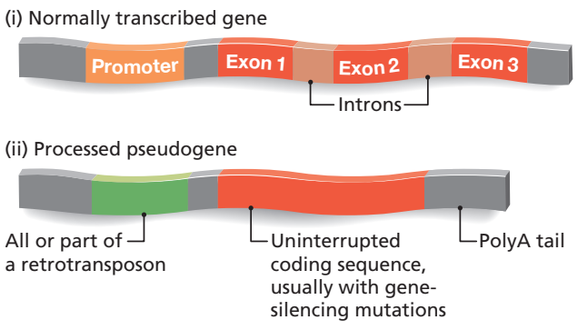


FIGURE 5.11B **Gene Structures of the Quick and the Dead**

Such a gene may be called a **retrogene.** This initial transcription does not guarantee the continued "survival" of the gene. The additional protein produced as a result of the transcription and translation of the processed gene may reduce fitness. If so, natural selection will favor mutations that silence the processed gene. Or the additional protein may acquire a new biological function.

An example of this evolutionary process is the autosomal *PGK* gene in mammals, which is homologous to an X-linked *PGK* gene. The autosomal *PGK* gene has no introns, indicating that its DNA came from an RNA intermediate in which introns were excised. Autosomal *PGK* is expressed almost exclusively in the testes, a novel tissue specificity. The maintenance of gene activity by the autosomal *PGK* gene may have happened because the gene located on the X chromosome is normally shut down during spermatogenesis within the testes. In effect, the accidental creation of the retrogene may have allowed mammalian evolution to correct a problem that had limited spermatogenesis before the gene duplication. In such cases, retrotransposons may actively foster adaptive evolution, an ironic side effect of their lives as genomic parasites.

5.12 Genome size is highly variable, perhaps due to the proliferation of useless elements

The total amount of DNA per haploid cell is known as the **C-value,** where "C" stands for *characteristic*. Some aspects of C-value evolution are easy to understand. Bacteria usually have much smaller C-values, about 500 to 13,000 kilobases of DNA. Eukaryotic animals, on the other hand, have C-values of 50 to 140,000 megabases, much greater in size. Figure 5.12A contrasts bacterial and animal genome sizes. This difference in size makes intuitive sense, because animals have many differentiated cell types, so they should have correspondingly more genetic information.

But there are anomalies in the C-value data. Humans have just 3200 megabases of DNA; some lungfish have 140,000 megabases. Why should lungfish need so much more genetic information than humans? Some ferns have 160,000 megabases of DNA. Even a unicellular ameba (*Amoeba dubia*) has 670,000 megabases of DNA, about 200 times more than humans have. Part (ii) of Figure 5.12A shows the relative magnitudes of three of these different genomes. The amoeba genome, however, is too big to fit in the figure and still see the human genome, because the ameba genome is about 200 times larger.

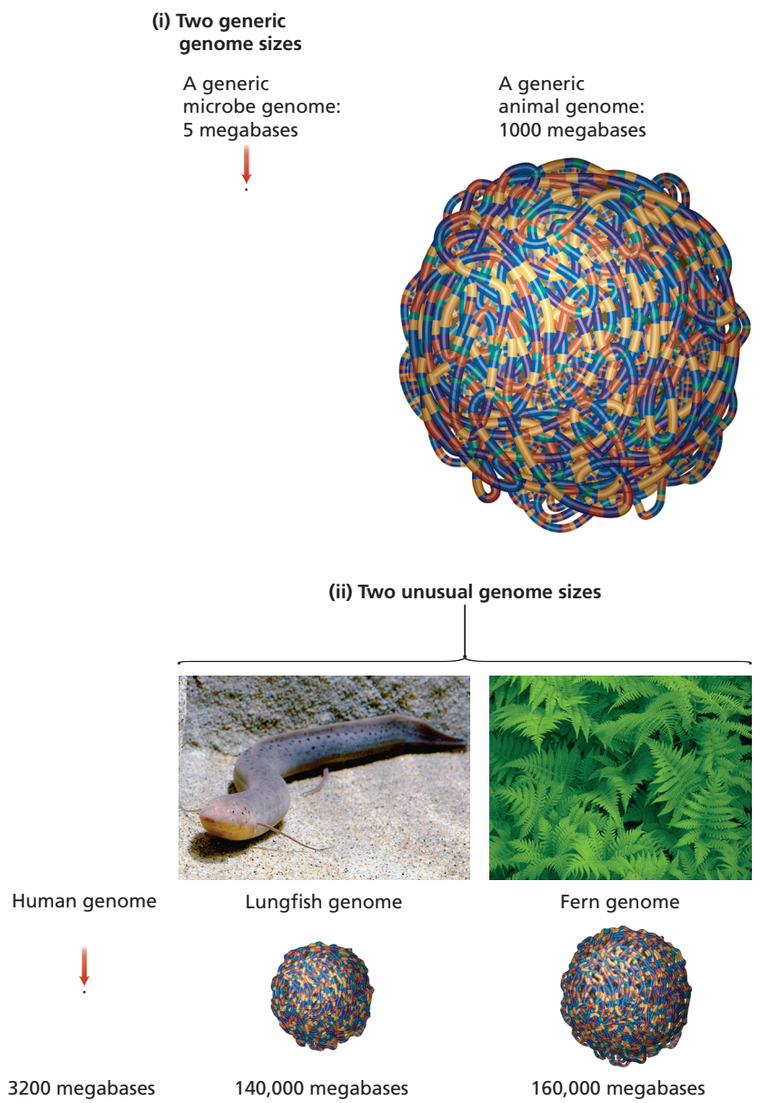


FIGURE 5.12A DNA Content of Microbes Compared with Animals

There are no obvious functional patterns in these DNA data. The amount of DNA that a eukaryotic organism has bears no relation to its morphological or physiological complexity. How are we to make sense of this?

One way to make sense of the wild variation in DNA content is to view it as a by-product of molecular processes like transposition, processes that do not have a simple correlation with the function of the organism, unless they are merely deleterious. That is, instead of viewing genome structure or size as a character comparable to the size or shape of a bone, we might view the genome as a sort of ecosystem, in which genome properties arise via a kind of evolutionary ecology of DNA sequences that copy themselves and proliferate throughout the genome.

In this DNA ecology, some processes eliminate DNA and other processes cause DNA to proliferate, as shown in Figure 5.12B. These processes may spread particular DNA sequences, as well as eliminate specific DNA sequences. In addition to such random mechanisms of loss, it is possible that selection acts against genomes that are overly large. The best evidence for this is the compactness of the genomes of microbes, for which DNA replication is a major part of their overall metabolism. However, this inference is indirect. Duplicative transposition and amplification of tandem arrays are the two obvious molecular mechanisms by which genome sizes can increase. Recent evidence associates an abundance of transposing elements with larger genomes, but this evidence is still indirect. The evolution of genome size is one of the most interesting topics in the study of molecular evolution. ❖

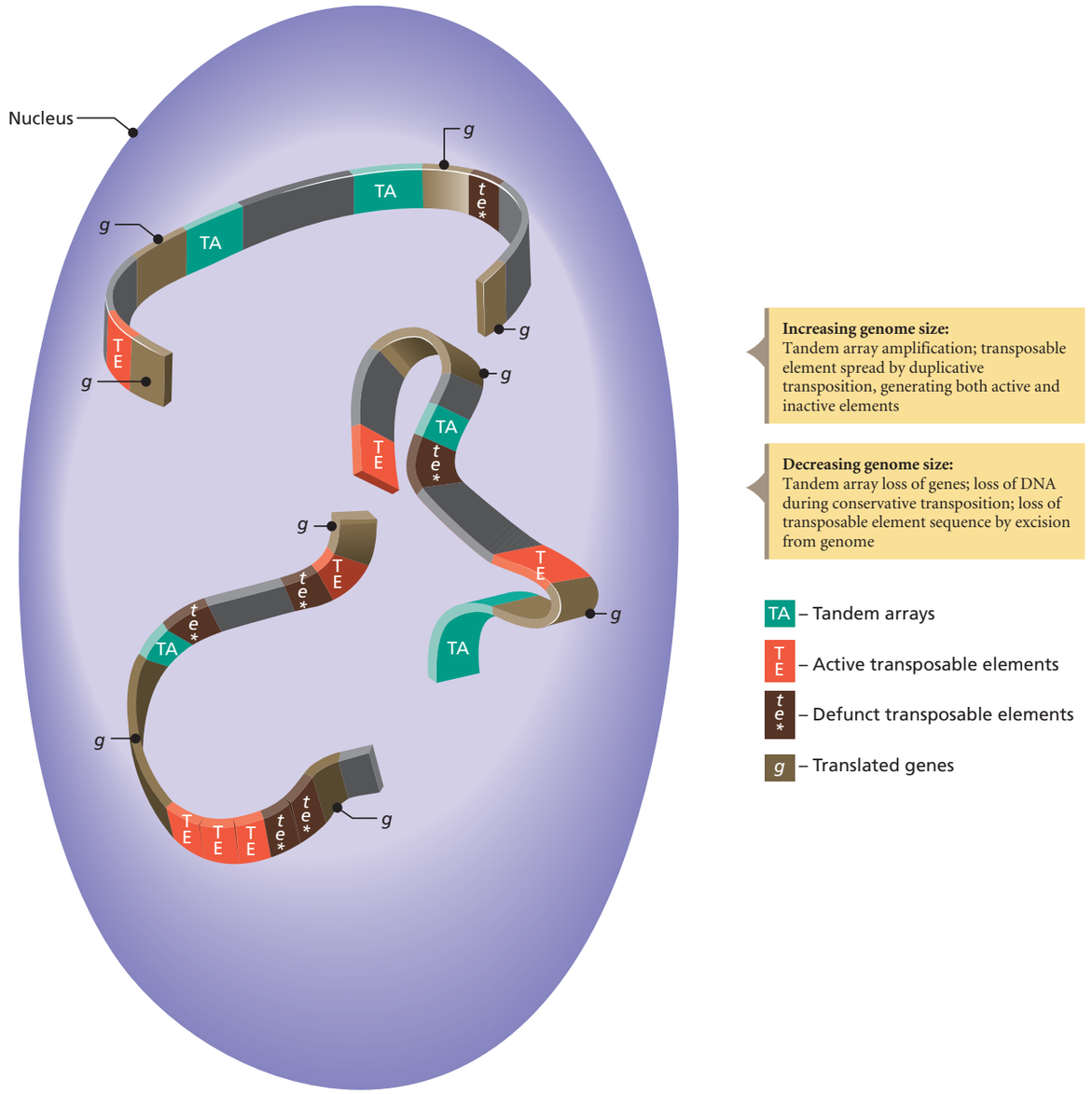


FIGURE 5.12B Mechanisms That Change Genome Size

SUMMARY

1. Biologists have now sequenced the entire human genome, as well as the genomes of other species, from bacteria to mouse. Whole-genome sequencing has clearly revealed patterns of molecular evolution that biologists have been piecing together over the last fifty years.

2. It was once thought that the genome is an organized library packed with information that specified the functioning of millions of genes. We now know that genomes are made up of about 2000–50,000 genes. Prokaryotes have smaller genomes that are reasonably well organized. But eukaryote genomes are generally a muddle. Their genes frequently have useless sequence information inserted at random within them, the introns. There are also large gene deserts between genes, regions that appear to have no function. Some DNA sequences move around genomes, generating mutations and chromosome rearrangements. Large amounts of repetitive DNA may evolve from the repeated unequal crossing over of tandem arrays of repeated DNA sequences.

3. It has been proposed that much of the evolution of DNA sequences within the genome is neutral. That is, it proceeds without control by natural selection, subject primarily to molecular-level processes, like transposition, mutation, and unequal crossing over. Several features of molecular evolution support this model. One is the rough constancy of nucleotide substitutions, called the molecular clock. Another is the relative uniformity of the rate of evolution of DNA sequences that do not affect the amino acid coding of genes, called the synonymous substitution rate.

4. Despite the apparent success of the neutral model of molecular evolution, there must be cases where natural selection intervenes in molecular evolution. Hemoglobin polymorphism in human populations exposed to malaria supplies one case that indicates the action of natural selection. Another example is the polymorphism of alleles at the *adh* locus of *Drosophila melanogaster.* The genome churning of transposable elements also generates new genes, which can be seized on by natural selection to create new genetic functions.

5. Much of molecular evolution is probably irrelevant to the evolution of the visible characters of organisms. But some of it plays a critical role in functional evolution, giving rise to new adaptations at the molecular level.

REVIEW QUESTIONS

1. Transposable elements normally act in what kind of adaptation?
2. Do humans have the largest genome size?
3. Pseudogenes come from what source?
4. Is the genetic code redundant?
5. Are all molecular genetic variants subject to natural selection?
6. When does the molecular clock keep better time?
7. Why does the molecular clock allow us to estimate the times of evolutionary divergence?
8. Offer some explanations for molecular genetic polymorphism.
9. Why is there so much DNA between genes in some eukaryotes?

KEY TERMS

actin
codon
codon use bias
Crick, Francis
C-value
divergence time
DNA-based transposition
eukaryotic gene
exon
genetic transcription
genetic translation
genome
histone
immunoglobulin
intron
Kimura, Motoo
McClintock, Barbara
messenger RNA (mRNA)
molecular clock
neutral theory of molecular evolution
Pauling, Linus
prokaryotic genome
promoter
protein electrophoresis
pseudogene
rate of evolutionary divergence
retrogene
retrotransposon (retroposon)
reverse transcriptase
Sarich, Vince
substitution, nonsynonymous
substitution, number of
substitution, synonymous DNA
tandem array
transposable element
transposase
unequal crossing over
unequal recombination
variegation
Watson, James
Wilson, Allan
Zuckerkandl, Emil

FURTHER READINGS

Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. 2002. *Molecular Biology of the Cell*, 4th ed. New York: Garland Publishing.

Gillespie, John H. 1991. *The Causes of Molecular Evolution.* New York: Oxford University Press.

Lewontin, Richard C. 1974. *The Genetic Basis of Evolutionary Change.* New York: Columbia University Press.

Kimura, Motoo. 1983. *The Neutral Theory of Molecular Evolution.* London: Cambridge University Press.

Li, Wen-Hsiung, and Daniel Graur. 2000. *Fundamentals of Molecular Evolution,* 2nd ed. Sunderland, NJ: Sinauer.

Selander, Robert K., Andrew G. Clark, and Thomas S. Whittam, eds. 1991. *Evolution at the Molecular Level.* Sunderland, NJ: Sinauer.