GENE EXPRESSION AND GENE INTERACTION

Gene is a general term meaning, loosely, the physical entity transmitted from parent to offspring in reproduction that influences hereditary traits. Genes influence human traits such as hair color, eye color, skin color, height, weight, and various aspects of behavior—although most of these traits are also influenced more or less strongly by environment. Genes also determine the makeup of proteins such as hemoglobin, which carries oxygen in the red blood cells, or insulin, which is important in maintaining glucose balance in the blood. Genes can exist in different forms or states. For example, a gene for hemoglobin may exist in a normal form or in any one of a number of alternative forms that result in hemoglobin molecules that are more or less abnormal. These alternative forms of a gene are called **alleles**.

From a biochemical point of view, a gene corresponds to a region along a molecule of DNA (deoxyribonucleic acid). DNA is the genetic material. A molecule of DNA consists of two strands wound around each other in the form of a right-handed helix (the celebrated "double helix"). Each strand is a polymer of constituents called **nucleotides**, of which there are four, conventionally symbolized A, T, G, and C according to the nitrogen-rich base that each contains — either adenine (A), thymine (T), guanine (G), or cytosine (C). The paired strands are held together by weak chemical bonds (hydrogen bonds) that form between A and T at corresponding positions in opposite strands or between G and C at corresponding positions in opposite strands (Figure 1.1). Wherever one strand contains an A, the other across the way contains a T; and wherever one strand contains a G, the other across the way contains a C. Because of the pairing of complementary bases—A with T and G with C—a double-stranded DNA molecule contains an equal number of A and T nucleotides as well as an equal number of G and C nucleotides. DNA molecules can be very long. The DNA molecule in the bacterium *E. coli* is about 4.7 million base pairs, that in the largest chromosome in the fruit fly *Drosophila melanogaster* is about 65 million base pairs, and that in the largest human chromosome is about 230 million base pairs. Physical manipulation of such large molecules is impractical. In order to be studied, they must first be broken into smaller pieces.

Gene Expression

Most genes code for the polypeptide chains that constitute proteins. The code is the sequence of nucleotides along the DNA. In the decoding of the nucleotide sequence in DNA and also in the synthesis of proteins, several

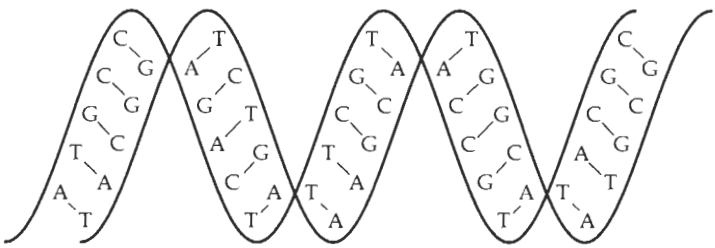

Figure 1.1 Genes are fundamental units of genetic information that correspond chemically to the sequence of nucleotides in a segment of DNA. A molecule of duplex DNA is composed of two intertwined strands, each of which consists of a long sequence of nucleotides. The strands are held together by pairing between the bases A and T in opposite strands and between the bases G and C in opposite strands. The short diagonal lines indicate the paired bases. There are 10 base pairs per turn of the double helix. A typical gene consists of hundreds of thousands of nucleotides, only a few of which are shown here.

types of RNA (ribonucleic acid) are essential. RNA is also a polymer of nucleotides, each of which carries a base. Three of the bases in RNA (A, C, and G) are the same as those in DNA. The fourth [uracil (U)] is different. When an RNA strand pairs with a complementary strand of DNA, U in the RNA pairs with A in the DNA. Hence, the base-pairing role of U in RNA is the same as that of T in DNA.

The essentials of gene expression in the cells of higher organisms (eukaryotes) are outlined in Figure 1.2. The coding regions of the DNA in a

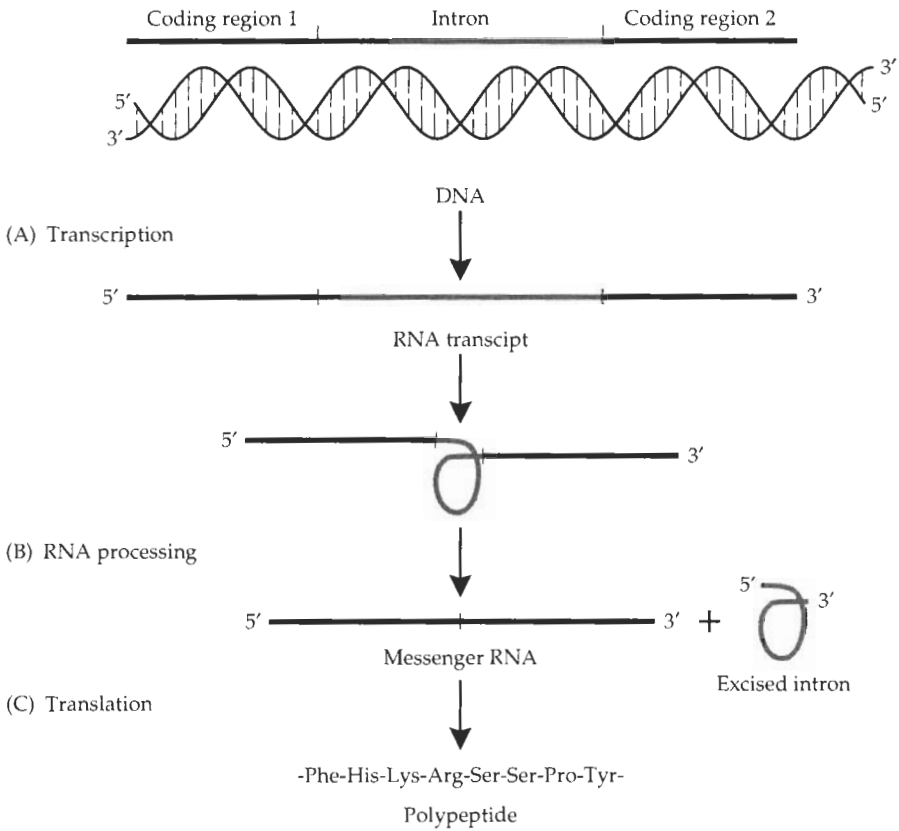

Figure 1.2 Processes in gene expression in eukaryotic cells. (A) DNA regions coding for the amino acids in a single polypeptide can be interrupted by noncoding regions (introns). (B) When the DNA is copied into RNA in transcription, both coding and noncoding regions are transcribed. However, the introns are removed from the transcript by processing. (C) In the messenger RNA, the coding regions are contiguous. The messenger RNA is translated to form the chain of linked amino acids constituting the polypeptide.

gene, which code for amino acids, are often interrupted by one or more noncoding regions known as intervening sequences or **introns**. In the first step in gene expression (transcription), a molecule of RNA is produced that is complementary in base sequence to one of the strands of DNA (Figure 1.2A). Every gene includes a regulatory region (sometimes more than one) that determines when transcription takes place, the types of cells in which it takes place, and the strand that is to be transcribed. Because of the base pairing rules, a DNA sequence—say, 3′–ATCG–5′—results in a complementary RNA sequence—in this example, 5′–UAGC–3′. Note that the DNA and RNA strands each have a **polarity** or directionality. The terms 5′ and 3′ refer to the polarity of the strands. The 5′ end typically terminates with a free phosphate group and the 3′ end typically terminates with a free hydroxyl group (—OH). When two strands of nucleic acid are paired, the polarity of each strand is opposite to that of the other. In the duplex DNA in Figure 1-2, for example, the left-to-right polarity of one strand is 5′-to-3′, whereas the left-to-right polarity of the partner strand is 3′-to-5′. Similarly, in transcription, the template DNA strand has a left-to-right polarity of 3′-to-5′, whereas the RNA transcript has the left-to-right polarity of 5′-to-3′. Because of the complementary base pairing between DNA and RNA nucleotides, the base-sequence code in DNA becomes converted into a base-sequence code in RNA. In transcription, the base sequence present in the introns is also faithfully copied into the base sequence of the RNA transcript.

The second step in gene expression in eukaryotes is RNA processing (Figure 1.2B). The beginning and end of the RNA transcript are chemically modified and the introns are removed by splicing (cutting and rejoining). RNA processing results in a molecule called **messenger RNA (mRNA)**, in which the coding regions have been made contiguous. The regions in the original RNA transcript that are retained in the mature mRNA are called **exons**. The central part of the mRNA contains the spliced exons that code for the amino acid sequence of a polypeptide chain. The mRNA also includes exons upstream and downstream from the protein-coding region. The upstream region is the *5′ untranslated region* and the downstream region is the *3′ untranslated region*.

The final step in gene expression is **translation**, in which the mRNA molecule combines with ribosomes and other types of RNA molecules in the cytoplasm to produce the final polypeptide (Figure 1.2C). In the coding region of the mRNA, each adjacent group of three nucleotides constitutes a separate coding group or **codon** that specifies which amino acid is to be incorporated into the polypeptide chain. The ribosome moves along the mRNA in steps of three nucleotides (codon by codon). As each new codon comes into place, the correct amino acid is brought into line and attached to the end of the growing chain of amino acids. New amino acids are added to the growing chain until a codon specifying "stop" is encountered. At this point synthesis of the chain of amino acids is finished and the polypeptide is released from the ribosome.

In prokaryotes, which includes bacteria and other organisms lacking a nucleus, gene expression is essentially identical to that in eukaryotes except for the absence of RNA processing. Genes in prokaryotes do not contain introns and so splicing is unnecessary. In prokaryotes, the original RNA transcript is used immediately as mRNA and translated into a polypeptide. Because there is no separate nucleus, translation in prokaryotes often begins immediately when the 5′ end of an RNA transcript comes off the DNA and even before transcription of the 3′ end of the same molecule has been completed.

The central role of RNA in gene expression is one of the oddities of biology that makes sense in the light of evolution. That gene expression is configured around RNA is a legacy of the earliest forms of life when RNA molecules served both as carriers of genetic information and as catalytic molecules. The role of RNA as carrier of genetic information was gradually replaced by DNA, and the role of RNA as catalytic molecules was gradually replaced by proteins. At every step along the way, as the RNA world evolved into the DNA world, the role of RNA was indispensable in the processes of information transfer and protein synthesis, and so the RNA intermediates became locked in place.

The Genetic Code

The **genetic code** is the list of all codons showing which amino acid each codon specifies. Table 1.1 shows the standard genetic code used in nuclear genes in most organisms. A few organisms and some cellular organelles, such as mitochondria, use slightly altered codes. The codons in Table 1.1 are those found in the mRNA. The amino acids are given by three-letter abbreviations as well as by conventional single-letter abbreviations. Codon AUG is the start codon in polypeptide synthesis; it specifies methionine (Met) at the beginning of the polypeptide as well as at internal positions. Three codons are stops that result in termination of polypeptide synthesis: UAA, UAG, and UGA. The genetic code is redundant in that most amino acids are specified by more than one codon. Most of the redundancy is in the third codon position.

A code for an amino acid is *twofold degenerate* if either of two sequences specifies the same amino acid. Twofold degenerate codes have the pattern ··Y or ··R, where ·· stands for the bases in codon positions 1 and 2. The symbol Y stands for any pyrimidine base (either U or C); the symbol R stands for any purine base (either A or G). For example, CAU and CAC both code for histidine (His), fitting the pattern CAY; and CAA and CAG both code for glutamine (Gln), fitting the pattern CAR. A code for an amino acid is *fourfold degenerate* if any of four sequences specifies the same amino acid; fourfold degenerate codes have the form ··N, where N means any nucleotide (U, C, A, or G). For example, GUU, GUC, GUA, and GUG all code for valine (Val),

TABLE 1.1 THE STANDARD GENETIC CODE

First nucleotide in codon	Second nucleotide in codon: U	C	A	G
U	UUU, UUC: Phe (F)	UCU, UCC, UCA, UCG: Ser (S)	UAU, UAC: Tyr (Y)	UGU, UGC: Cys (C)
	UUA, UUG: Leu (L)		UAA: Stop; UAG: Stop	UGA: Stop; UGG: Trp (W)
C	CUU, CUC, CUA, CUG: Leu (L)	CCU, CCC, CCA, CCG: Pro (P)	CAU, CAC: His (H); CAA, CAG: Gln (Q)	CGU, CGC, CGA, CGG: Arg (R)
A	AUU, AUC, AUA: Ile (L)	ACU, ACC, ACA, ACG: Thr (T)	AAU, AAC: Asn (N); AAA, AAG: Lys (K)	AGU, AGC: Ser (S); AGA, AGG: Arg (R)
	AUG: Met (M*)			
G	GUU, GUC, GUA, GUG: Val (V)	GCU, GCC, GCA, GCG: Ala (A)	GAU, GAC: Asp (D); GAA, GAG: Glu (E)	GGU, GGC, GGA, GGG: Gly (G)

Note: Codons are nonoverlapping three-base sequences present in mRNA, each of which specifies an amino acid in a polypeptide chain or terminates synthesis ("Stop"). The full names of the amino acids are phenylalanine (Phe), leucine (Leu), isoleucine (Ile), methionine (Met), valine (Val), serine (Ser), proline (Pro), threonine (Thr), alanine (Ala), tyrosine (Tyr), histidine (His), glutamine (Gln), asparagine (Asn), lysine (Lys), aspartic acid (Asp), glutamic acid (Glu), cysteine (Cys), tryptophan (Trp), arginine (Arg), and glycine (Gly).

which fits the pattern GUN. Note in Table 1.1 that the code for isoleucine is *threefold degenerate* and those for leucine, arginine, and serine are each *sixfold degenerate*.

The codons for amino acids are not used randomly in proteins. There are preferred codons for amino acids that differ from one gene to the next and from one organism to another. Codon preferences exist even within redundancy classes. In *Drosophila*, for example, among codons for histidine, CAC is used more than CAU in a ratio of about 2 : 1. Similarly, among codons for glutamine, CAG is used more than CAA in a ratio of about 3 : 1. Another example of nonrandom codon usage is the AUA codon for isoleucine, which tends to be avoided in most proteins in most organisms. In *Drosophila*, AUU and AUC are used more than AUA in a ratio of about 10 : 1. One evolutionary

hypothesis that explains the avoidance of AUA is that, because of the degeneracy of the genetic code, the AUA codon might sometimes be translated as AUG, which codes for methionine. Because methionine is likely to change protein structure radically, the mistranslation would be a costly mistake. Through evolutionary time, one by one, the AUA codons in a messenger RNA become replaced with AUU or AUC, minimizing this type of misincorporation error. This misincorporation hypothesis for AUA codon avoidance has not been tested, but it is testable.

Alleles

Alternative alleles of a gene differ in their sequence of nucleotides (Figure 1.3). For example, where one allele has a T-A base pair in the DNA, another may have a C-G base pair at the same position. Because of redundancy in the code, not all nucleotide substitutions result in a replacement of one amino acid for another. In Figure 1.3B, for example, if a mutation at the third position in the second codon (asterisk) changes one pyrimidine into the other, the new codon still codes for histidine. On the other hand, some nucleotide substitutions at the third position do result in amino acid replacements. For example, in Figure 1.3C, if the third position in the second codon changes from a pyrimidine to a purine, the codon changes from one for histidine to one for glutamine. Most nucleotide substitutions at codon positions one and two result in amino acid replacements (Figure 1.2D).

Not all alleles differ by a mere nucleotide substitution. Relative to the typical or **wildtype** allele, some alleles may have a deletion of a number of nucleotide pairs or an insertion into the DNA molecule. The number of nucleotides deleted or inserted may be small (as few as one nucleotide pair) or large. Some insertions are thousands of nucleotide pairs in size. Many large insertions result from the activity of **transposable elements**, which are specialized sequences of DNA able to replicate and insert at novel positions virtually anywhere in the DNA of the organism in which they are present. Alleles also may differ in the number of copies of short sequences present in tandem arrays in the DNA. For example, near many genes in human beings are tandem copies of dinucleotides, such as 5′–CACACACA . . . –3′. Such a repeating sequence is symbolized as (5′–CA–3′)n. The number of copies (n) of the dinucleotide repeat often range from fewer than ten to hundreds, and the number of copies may differ dramatically from one allele to the next. Some alleles even differ from wildtype in having an inversion of the nucleotide sequence in a region of DNA.

Genotype and Phenotype

Within a living cell, genes are arranged in linear order along microscopic threadlike bodies called **chromosomes**. A typical chromosome may contain

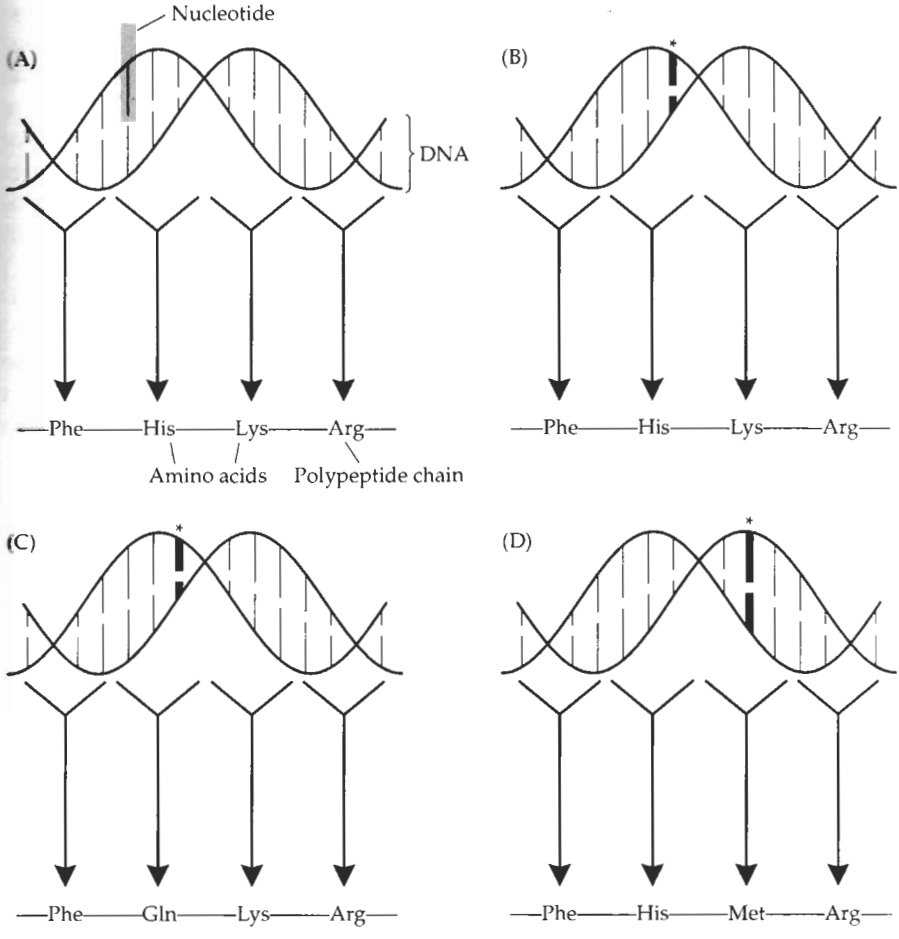

Figure 1.3 Alleles are alternative forms of a gene. (A) The arrows show how the genetic information in a portion of the nucleotide sequence of DNA specifies the amino acid sequence in a portion of a polypeptide. Each group of three adjacent nucleotides corresponds to one amino acid in the polypeptide. (B, C, D) Substitution of one nucleotide for another in the DNA (indicated by the asterisks and heavy lines) can result in the replacement of one amino acid for another in the polypeptide.

several thousand genes. The position of a gene along a chromosome is called the **locus** of the gene. In most higher organisms, each cell contains two copies of each type of chromosome. Such organisms, in which the chromosomes are present in pairs, are said to be **diploid**. In each pair of chromosomes, one

member is inherited from the mother through the egg and the other is inherited from the father through the sperm. At every locus, therefore, diploid organisms contain two alleles, one each at corresponding positions in the maternal and paternal chromosomes. If the two alleles at a locus are chemically identical (in the sense of having the same nucleotide sequence along the DNA), the organism is said to be **homozygous** at the locus under consideration; if the two alleles at a locus are chemically different, the organism is said to be **heterozygous** at the locus. The term *gene* is a general term usually used in the sense of *locus*.

Geneticists make a fundamental distinction between the genetic constitution of an organism and the physical or biochemical attributes of the organism. The genetic constitution of an organism is called the **genotype**; genotype thus refers to the particular alleles present in an organism at all loci that affect the trait in question. For example, if a trait is influenced by two genes, each with two alleles, then there are nine possible genotypes, as follows:

$$\begin{array}{lll} AA\,;BB & AA\,;Bb & AA\,;bb \\ Aa\,;BB & Aa\,;Bb & Aa\,;bb \\ aa\,;BB & aa\,;Bb & aa\,;bb \end{array}$$

where A and a refer to the alleles of the first gene and B and b refer to the alleles of the second gene. In some cases when the genes are **linked** (located in the same chromosome), it is sometimes necessary to distinguish between the genotypes AB/ab and Ab/aB, in which case there are ten possible genotypes.

In contrast to genotype, the physical expression of a genotype is called the **phenotype**. Examples of phenotypes include hair color, eye color, height, weight, number of kernels on an ear of corn, number of eggs laid by a hen, and round versus wrinkled pea seeds. The distinction between the genetic constitution of an organism (genotype) and the physical or biochemical attributes of the organism (phenotype) is particularly important in cases in which the environment can affect the trait; in such cases, two organisms with the same genotype can nevertheless have different phenotypes because of differences in the environment. Conversely, two organisms with the same phenotype can have different genotypes.