CHAPTER

1

Introduction to Molecular Genetics

The Human Genome Project is a bold undertaking to understand, at a fundamental level, all of the genetic information required to build and maintain a human being. The human **genome** is the complete information content of the human cell. This information is encoded in approximately 3.2 billion base pairs of DNA contained on 46 **chromosomes** (22 pairs of **autosomes** plus the two sex chromosomes—see Figure 1.1). The completion, in 2001, of the first draft of the human genome sequence was only the first phase of this project (Venter et al. 2001; Lander et al. 2001).

To use the metaphor of a book, the draft genome sequence gives biology all of the letters, in the correct order on the pages, but without the ability to recognize words, sentences, and punctuation, or even an understanding of the language in which the book is written. The task of making sense of all of this raw biological information falls, at least initially, to **bioinformatics** specialists who make use of computers to find the words and decode the language. The next step is to integrate all of this information into a new form of experimental biology, known as **genomics**, that

Essentials of Medical Genomics, *Second Edition* By Stuart M. Brown

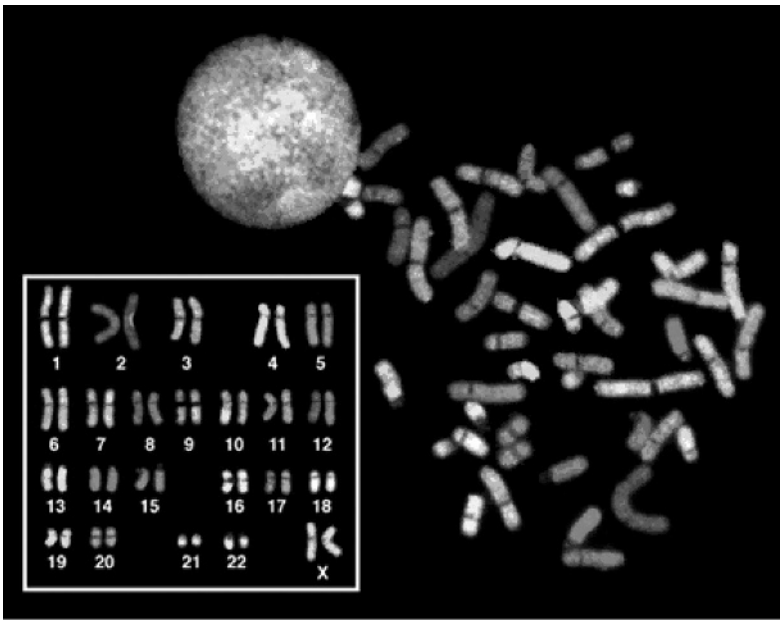

FIGURE 1.1. Human karyotype—SKY image: available at `http://www.accessexcellence.org/AB/GG/sky.gif`; credit to Chroma Technology Inc. (See insert for color representation.)

can ask meaningful questions about what is happening in very complex systems where tens of thousands of different genes and proteins are interacting simultaneously.

The primary justification for the considerable amount of money spent on sequencing the human genome (from governments and private corporations) is that this information will lead to dramatic medical advances. In fact, the first wave of new drugs and medical technologies derived from genome information is currently making its way through clinical trials and into the healthcare system. However, to effectively utilize these new advances, medical professionals need to understand something about genes and genomes. Just as it is important for physicians to understand how to Gram-stain and evaluate a culture of bacteria, even if they never actually perform this test themselves in their

medical practices, it is important to understand how DNA technologies work in order to appreciate their strengths, weaknesses, and peculiarities.

However, before we can discuss whole genomes and genomic technologies, it is necessary to understand the basics of how genes function to control biochemical processes within the cell (molecular biology) and how hereditary information is transmitted from one generation to the next (genetics).

THE PRINCIPLES OF INHERITANCE

The principles of genetics were first described by the monk Gregor Mendel in 1866 in his observations of the inheritance of traits in garden peas ["Versuche über Pflanzen-Hybriden" (Mendel 1866)]. Mendel described "differentiating characters" (*differierende Merkmale*) which may come in several forms. In his monastery garden, he made crosses between strains of garden peas that had different characters, each with two alternate forms that were easily observable, such as purple or white flower color, yellow or green seed color, smooth or wrinkled seed shape, and tall or short plant height. (These alternate forms are now known as alleles.) Then he studied the distribution of these forms in several generations of offspring from his crosses.

Mendel observed the same patterns of inheritance for each of these characters. Each strain, when bred with itself, showed no changes in any of the characters. In a cross between two strains that differ for a single character, such as pink versus white flowers, the first generation of hybrid offspring (the F_1) all resembled one parent—all pink. Mendel called this the **dominant** form of the character. After self-pollinating the F_1 plants, the second-generation plants (the F_2) showed a mixture of the two parental forms (see Figure 1.2). This is known as **segregation**. The **recessive** form that was not seen in the F_1s (white flowers) was found in one-fourth (25%) of the F_2 plants.

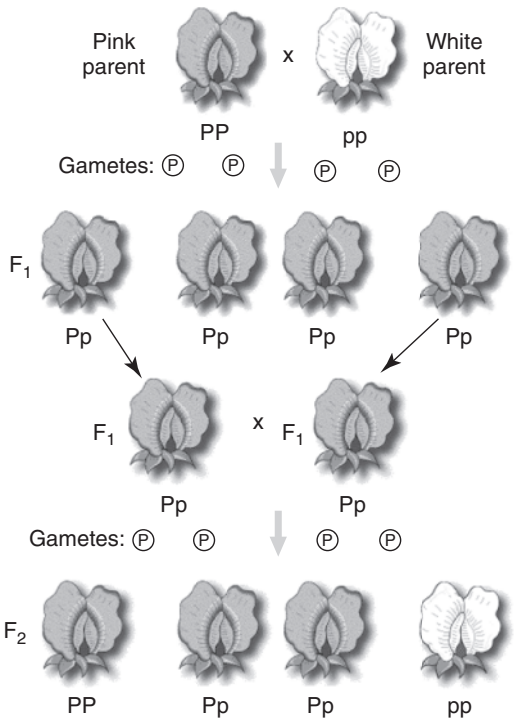

FIGURE 1.2. Mendel observed a single trait segregating over two generations. Pink and white parents have all pink F_1 progeny (heterozygous), but one-fourth of the F_2 generation are white and three-fourths are pink.

Mendel also made crosses between strains of peas that differed for two or more traits. He found that each trait was assorted independently in the progeny—there was no connection between whether an F_2 plant had the dominant or recessive form for one character and which form it carried for another character (see Figure 1.3).

Mendel created a theoretical model ("Mendel's laws of genetics") to explain his results. He proposed that each individual has two copies of the hereditary material for each character, which may determine different forms of that character. These two copies separate and are subjected to independent assortment

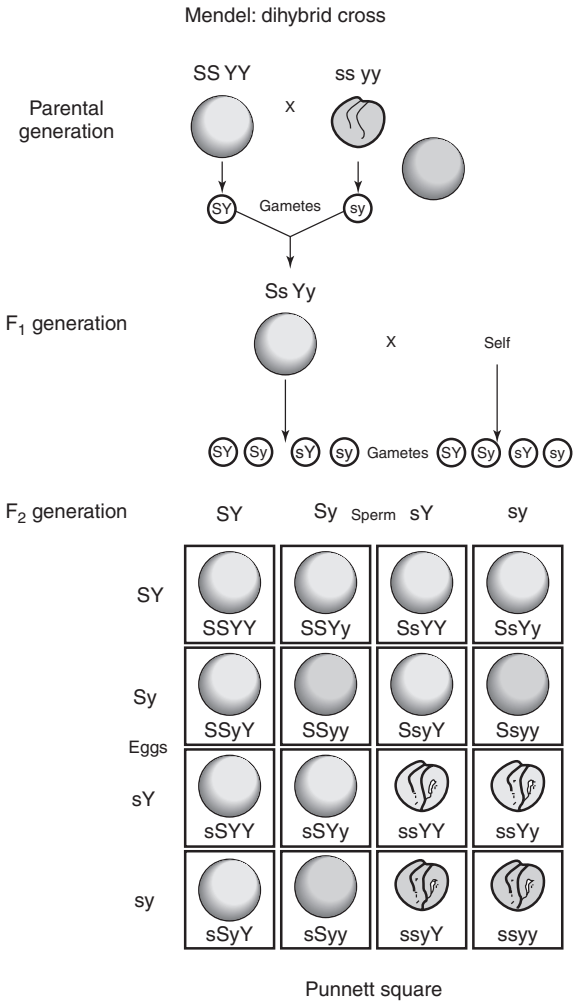

FIGURE 1.3. A cross where two independent traits are segregating (Y = yellow; S = smooth).

during the formation of gametes (sex cells). When a new individual is created by the fusion of two sex cells, the two copies from the two parents combine to produce a visible trait depending on which form is dominant and which is recessive. Mendel did not propose any physical explanation for how these traits were

passed from parent to progeny; his characters were purely abstract units of heredity.

Modern genetics has completely embraced Mendel's model with some additional detail. There may be more than two different alleles for a gene in a given population, but each individual has only two, which may be the same **(homozygous)** or different **(heterozygous)**. In some cases two different alleles combine to produce an intermediate form in heterozygous individuals, so that red and white flower alleles may combine to produce pink or type A and type B blood alleles, which in turn combine to produce the AB blood type.

GENES ARE ON CHROMOSOMES

In 1902, Walter Sutton, a microscopist, proposed that Mendel's heritable characters resided on the chromosomes which he observed inside the cell nucleus (see Figure 1.4). Sutton observed that "the association of paternal and maternal chromosomes in

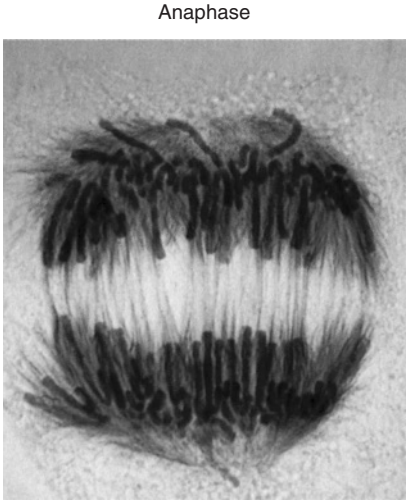

FIGURE 1.4. Anaphase chromosomes in a dividing lily cell. (See insert for color representation.)

pairs and their subsequent separation during cell division . . . may constitute the physical basis of the Mendelian law of heredity" (Sutton 1903).

In 1909, the Danish botanist Wilhelm Johanssen coined the term "gene" to describe Mendel's heritable characters. In 1910, Thomas Hunt Morgan found that a trait for white eye color was located on the X chromosome of the fruitfly and was inherited together with a factor that determines sex (Morgan 1910). A number of subsequent studies by Morgan (1919) and others showed that each gene for a particular trait was located at a specific spot or **locus** on a chromosome in all individuals of a species. The chromosome was perceived as a linear organization of genes, like beads on a string. Throughout the early part of the twentieth century, a gene was considered to be a single, fundamental, indivisible unit of heredity, in much the same way as an atom was considered to be the fundamental unit of matter.

Each individual has two copies of each type of chromosome, having received one copy from each parent. The two copies of each chromosome in the parent are randomly divided into the sex cells (sperm and egg) in a process called segregation. It is possible to observe the segregation of chromosomes during **meiosis** using only a moderately powerful microscope. It is an aesthetically satisfying triumph of biology that this observed segregation of chromosomes in cells exactly corresponds to the segregation of traits that Mendel observed in his peas.

Recombination and Linkage

In the early twentieth century, Mendel's concepts of inherited characters were broadly adopted by practical plant and animal breeders as well as experimental geneticists. It rapidly became clear that Mendel's experiments represented an oversimplified view of inheritance. He must have intentionally chosen characters in his peas that were inherited independently. In the breeding

experiments where many traits differ between parents, it is commonly observed that progeny inherit pairs or groups of traits together from one parent far more frequently than would be expected by chance alone. This observation fits nicely into the chromosome model of inheritance—if two genes are located on the same chromosome, then they will be inherited together when that chromosome segregates into a gamete, and that gamete becomes part of a new individual.

However, it was also observed that "linked" genes do occasionally separate. A theory of **recombination** was developed to explain these events. During the process of meiosis, it was proposed that the homologous chromosome pairs line up and exchange segments in a process called **crossing over**. This theory was supported by microscopic evidence of X-shaped structures called **chiasmata** forming between paired homologous chromosomes in meiotic cells (see Figure 1.5).

If a parent cell contains two different alleles for two different genes, then after the crossover, the chromosomes will contain new combinations of alleles. For example, if one chromosome contains alleles A and B for two genes, and the other chromosome contains alleles a and b, then without crossovers, all progeny must inherit a chromosome from that parent with either an A–B or an a–b allele combination. If a crossover occurs between the two genes, then the resulting chromosomes will contain the A–b and a–B allele combinations (see Figure 1.6).

Morgan, continuing his work with fruitflies, demonstrated that the chance of a crossover occurring between any two linked

FIGURE 1.5. Chiasmata visible in electron micrograph of meiotic chromosome.

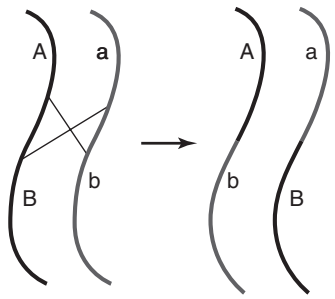

FIGURE 1.6. Schematic diagram of a single crossover between a chromosome with A–B alleles and a chromosome with a–b alleles to form A–b and a–B recombinant chromosomes. (See insert for color representation.)

genes is proportional to the distance between them on the chromosome. Therefore, by counting the frequency of crossovers between alleles of a given pair of genes, it is possible to create genetic maps of chromosomes. Morgan was awarded the 1933 Nobel Prize in Medicine for this work. In fact, it is generally observed that on average there is more than one crossover between every pair of homologous chromosomes in every meiosis, so that two genes located on opposite ends of a chromosome do not appear linked at all. On the other hand, alleles of genes that are located very close together are very rarely separated by recombination (see Figure 1.7).

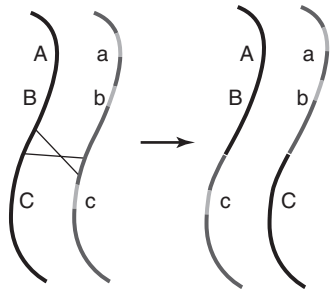

FIGURE 1.7. Genes A and B are tightly linked so that they are not separated by recombination, but gene C is farther away. After recombination occurs in some meiotic cells, gametes are produced with allele combinations ABC, abc, ABc, and abC. (See insert for color representation.)

The relationship between the frequency of recombination between alleles and the distance between genes on a chromosome has been used to construct genetic maps for many different organisms, including humans. It has been a fundamental assumption of genetics for almost a hundred years that recombinations occur randomly along the chromosome at any location, even within genes. However, more recent data from DNA sequencing of genes in human populations suggest that there are recombination hotspots and regions where recombination almost never occurs. This creates groups of alleles from neighboring genes on a chromosome, known as **haplotypes**, that remain linked together across hundreds of generations.

GENES ENCODE PROTEINS

Beadle and Tatum (1941) showed that a single mutation, caused by exposing the fungus *Neurospora crassa* to X rays, destroyed the function of a single enzyme, which interrupted a biochemical pathway at a specific step due to the loss of function of a particular enzyme. This mutation segregated among the progeny exactly as Mendel's traits did in peas. The X-ray-induced damage to a specific region of one chromosome destroyed the instructions for the synthesis of a specific enzyme. Thus a gene is a spot on a chromosome that codes for a single enzyme. In subsequent years, a number of other researchers broadened this concept by showing that genes code for all types of proteins, not just enzymes, leading to the **one gene–one protein** model, which is the core of modern molecular biology. Beadle and Tatum shared the 1958 Nobel Prize in Medicine.

GENES ARE MADE OF DNA

The next step in understanding the nature of the gene was to dissect the chemical structure of the chromosome. Crude

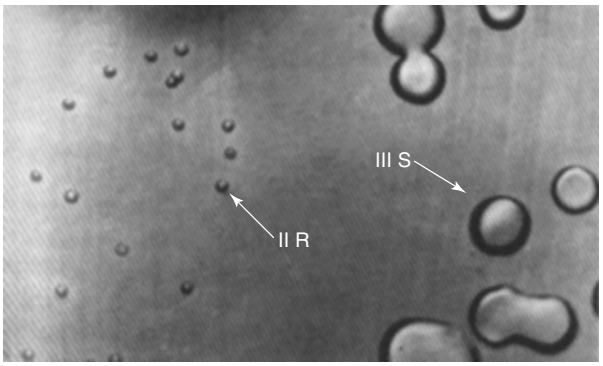

FIGURE 1.8. Transforming experiment: rough (II R) and smooth (III S) *Streptococcus pneumoniae* cells. (From Avery et al., 1944.)

biochemical purification had shown that chromosomes are composed of both protein and DNA. Avery et al. (1944) conducted the classic experiment on the "transforming principle." They found that DNA purified from a lethal S (smooth) form of *Streptococcus pneumoniae* could transform a harmless R (rough) strain into the S form (see Figure 1.8). Treatment of the DNA with protease to destroy all of the protein had no effect, but treatment with DNA-degrading enzymes blocked the transformation. Therefore, the information that transforms the bacteria from R to S must be contained in the DNA (McCarty 1985).

Hershey and Chase (1952) confirmed the role of DNA with their classic "blender experiment" on bacteriophage viruses. The phage were radioactively labeled with either ^{35}S in their proteins or ^{32}P in their DNA. They used a blender to interrupt the process of infection of *Escherichia coli* bacteria by the phage. Then they separated the phage from the infected bacteria by centrifugation and collected the phage and the bacteria separately. They observed that the ^{35}S-labeled protein remained with the phage while the ^{32}P-labeled DNA was found inside the infected bacteria (see Figure 1.9). This proved that it is the DNA portion of the virus that enters the bacteria and contains the genetic instructions

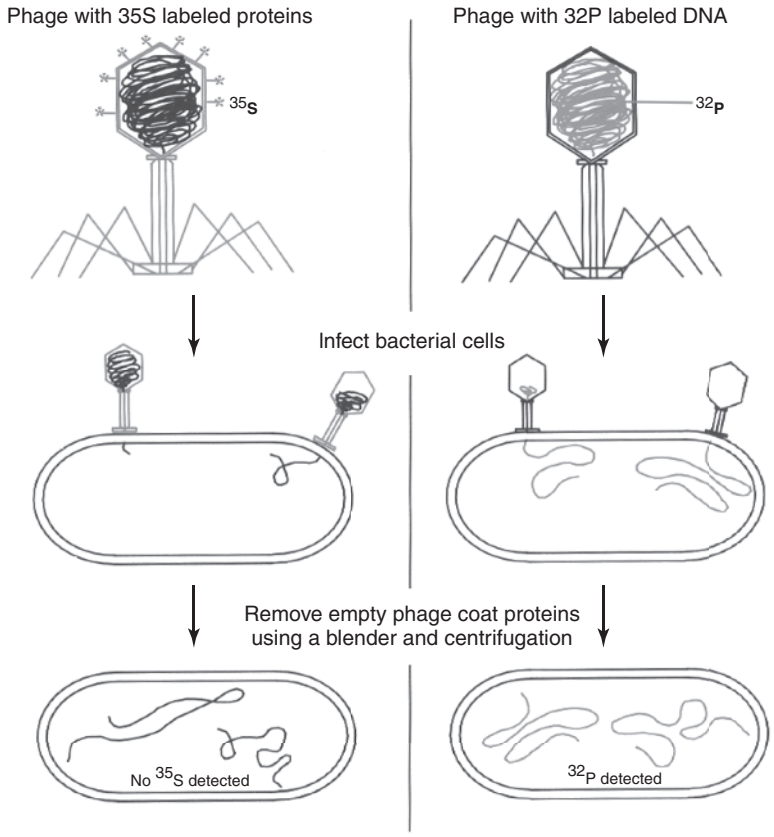

FIGURE 1.9. Hershey–Chase blender experiment. *Escherichia coli* bacteria are infected with phage with ^{35}S-labeled proteins or ^{32}P-labeled DNA. After removing the phage with a blender, the ^{32}P-labeled DNA but not the ^{35}S-labeled protein, is found inside the bacteria. (From Micklos and Freyer, *DNA Science*, Cold Spring Harbor Press, 1990.)

for producing new phage, not the proteins, which remain outside. Hershey was awarded the 1969 Nobel Prize for this work.

DNA Structure

Now it was clear that genes are made of DNA, but how does this chemically simple molecule contain so much information?

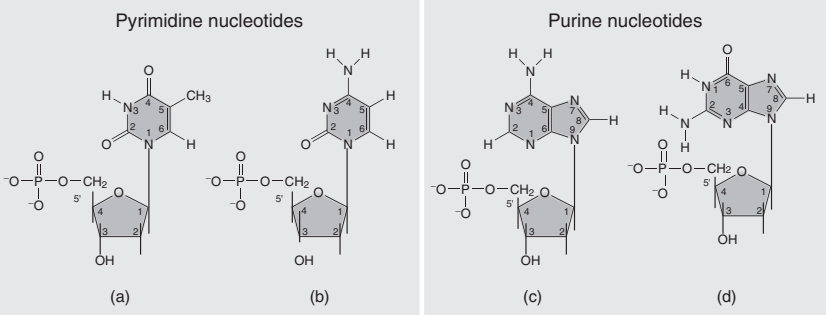

FIGURE 1.10. Chemical structures of the four DNA bases: (a) deoxythymidine monophosphate (dTMP); (b) deoxycytidine monophosphate (dCMP); (c) deoxyadenosine monophosphate (dAMP); (d) deoxyguanosine monophosphate (dGMP).

DNA is a long polymer molecule that contains a mixture of four different chemical subunits: adenine, cytosine, guanosine, and thymine (abbreviated as A, C, G, and T). These subunits, known as **nucleotide bases**, have similar two-part chemical structures that contain a deoxyribose sugar and a nitrogen ring (see Figure 1.10), hence the name deoxyribose nucleic acid. The real challenge is to understand how the nucleotides fit together in a way that can contain a lot of information.

Chargaff (1950) discovered that there was a consistent one-to-one ratio of adenine to thymine and guanine to cytosine in any sample of DNA from any organism. In 1951, Linus Pauling and R. B. Corey described the α-helical structure of a protein (Pauling and Corey 1951). Shortly thereafter, Rosalind Franklin (Sayre 1975) provided X-ray crystallographic images of DNA to James Watson and Francis Crick (see Figure 1.11); this form of DNA was very similar to the α-helix described by Pauling. Watson and Crick's crucial insight (1953) was to realize that DNA formed a double helix with complementary bonds between adenine–thymine and guanine–cytosine pairs.

The Wastson–Crick model of the DNA structure resembles a twisted ladder. The two sides of the ladder are formed by strong

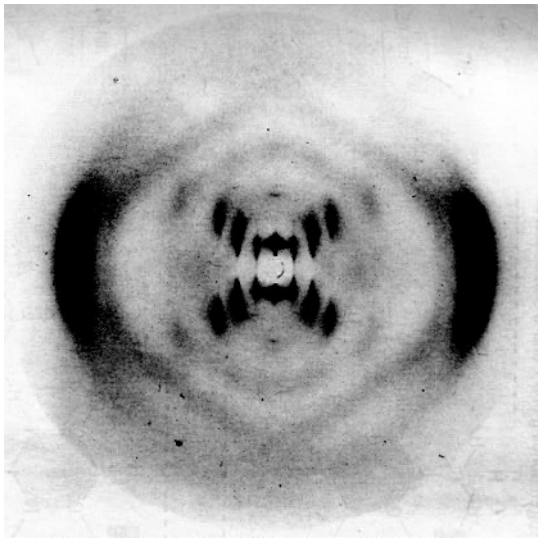

FIGURE 1.11. Rosalind Franklin's X-ray diffraction image of DNA.

covalent bonds between the phosphate on the 5′ carbon of one deoxyribose sugar and the methyl side groups of the 3′ carbon of the next (a phosphodiester bond). Thus, the deoxyribose sugar part of each nucleotide is bonded to the one above and below it, forming a chain that forms the backbone of the DNA molecule (see Figure 1.12). The phosphate-to-methyl linkage of the deoxyribose sugars give the DNA chain a direction or polarity, generally referred to as **5′ to 3′**. Each DNA molecule contains two parallel chains that run in opposite directions forming the sides of the ladder.

The rungs of the ladder are formed by weaker hydrogen bonds between the nitrogen ring parts of pairs of nucleotide bases. There are only two types of base pair bonds: adenine bonds with thymine, and guanine bonds with cytosine. The order of nucleotide bases on both sides of the ladder always reflects this complementary base pairing—so that wherever there is an A on one side, there is always a T on the other side, and vice versa.

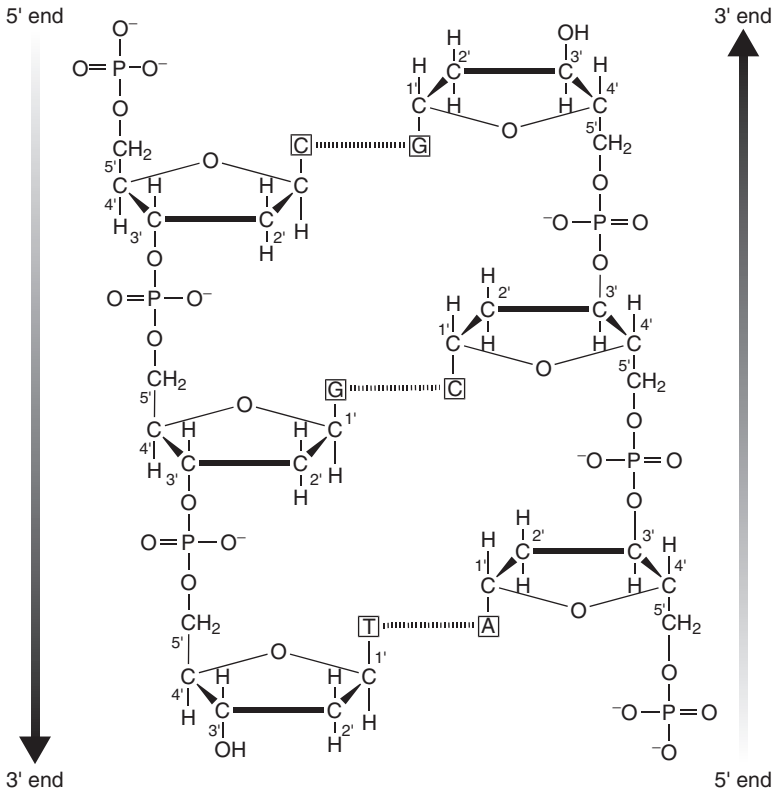

FIGURE 1.12. DNA phosphate bonds.

Since the A–T and G–C units always occur together, they are often referred to as **base pairs**. The G–C base pair has three hydrogen bonds, while the A–T pair only has two (see Figure 1.13), so the bonds between G–C bases are more stable at high temperatures than are A–T bonds. The nucleotide bases are strung together on the polydeoxyribose backbone-like beads on a string. It is the particular order of the four different bases as they occur along the string that contains all of the biological information.

Watson and Crick realized that this model of DNA structure contains many implications (see Figure 1.14). First, the two strands of the double helix are complementary, not identical.

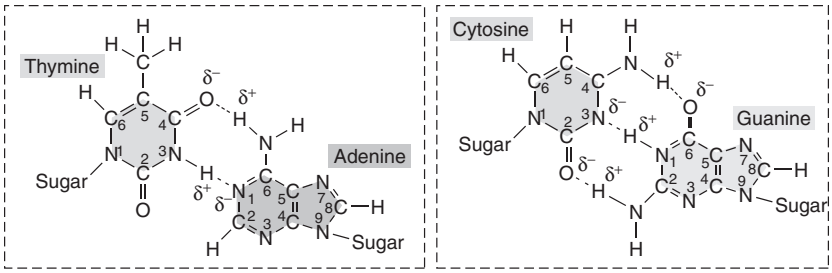

FIGURE 1.13. DNA hydrogen bonds in (a) A–T and (b) G–C base pairs.

Thus one strand can serve as a template for the synthesis of a new copy of the other strand—a T is added to the new strand wherever there is an A, a G for each C, and so on—perfectly retaining the information in the original double strand. In 1953, in a single-page paper in the journal Nature, they said, with a mastery of understatement: "It has not escaped our attention that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" (Watson and Crick 1953).

So, in one tidy theory, the chemical structure of DNA explains how genetic information is stored on the chromosome and how it is passed on when cells divide. That is why Watson and Crick won the 1962 Nobel Prize (shared with Maurice Wilkins).

If the two complementary strands of a DNA molecule are separated in the laboratory by boiling (known as **denaturing** the DNA), then they can find each other and again pair up, by re-forming the complementary A–T and C–G hydrogen bonds (**annealing**). Bits of single-stranded DNA from different genes do not have perfectly complementary sequences, so they will not pair up in solution. This process of separating and rematching complementary pieces of DNA, known as **DNA hybridization**, is a fundamental principle behind many different molecular biology technologies.

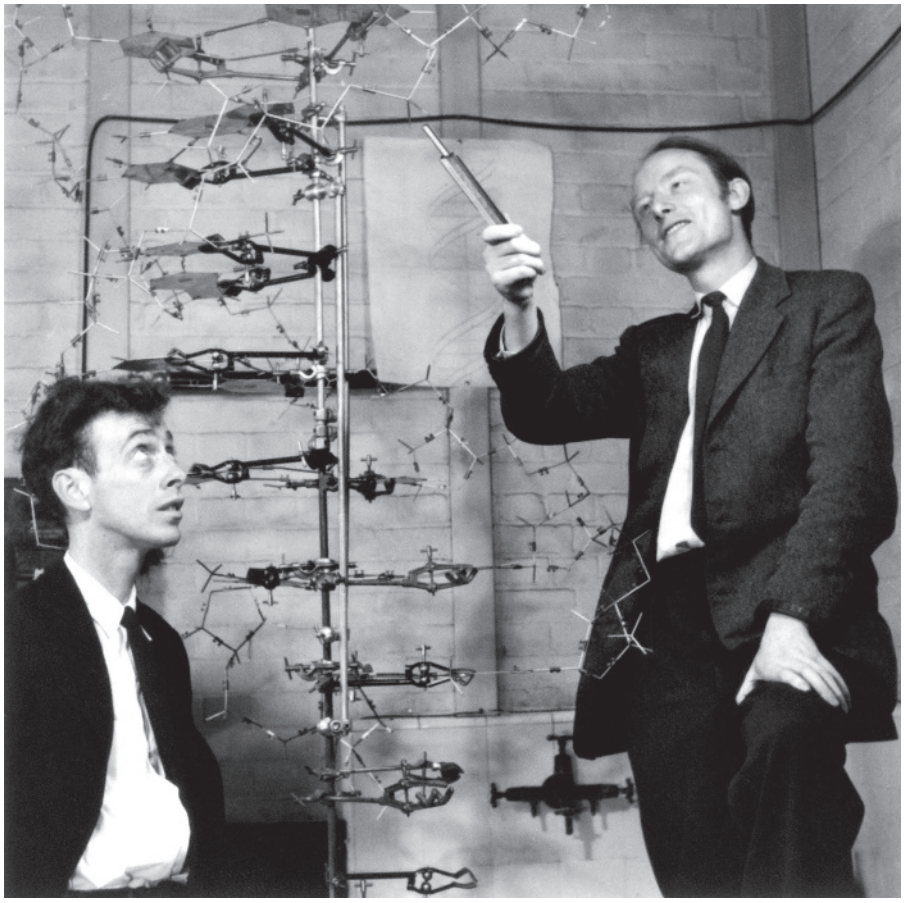

FIGURE 1.14. James Watson (left) and Francis Crick demonstrate their model of the DNA double helix. (From Watson J. 1968. *The Double Helix*, p 125. Atheneum, New York. Courtesy of Cold Spring Harbor Laboratory Archives.)

FIGURE 1.15. The central dogma of molecular biology (as described by Crick in 1957): DNA is transcribed into RNA, which is translated into protein.

THE CENTRAL DOGMA

Crick followed up in 1957 with a theoretical framework for the flow of genetic information in biological systems (Crick 1957). His theory, which has come to be known as the "Central Dogma" of molecular biology, is that DNA codes for genes in a strictly linear fashion—a series of DNA bases corresponding to a series of amino acids in a protein. DNA is copied into RNA, which serves as a template for protein synthesis. This leads to a nice, neat conceptual diagram of the flow of genetic information within a cell: DNA is copied to more DNA in a process known as **replication**, and DNA is **transcribed** into RNA, which is then **translated** into protein (see Figure 1.15).

DNA REPLICATION

Every ordinary cell (**somatic cell**) in an organism has a complete copy of that organism's genome. In mammals and other **diploid** organisms, that genome contains two copies of every chromosome, one from each parent. As an organism grows, cells divide by a process known as **mitosis**. Before a cell can divide, it must make a complete copy of its genome so that each daughter cell will receive a full set of chromosomes. All of the DNA is **replicated** by a process that makes use of the complementary nature of the base pairs in the double helix.

In DNA replication, the complementary base pairs of the two strands of the DNA helix partially separate and new copies of both strands are made simultaneously. A **DNA polymerase** enzyme attaches to the single-stranded DNA and synthesizes new strands by joining free DNA nucleotides into a growing chain that is exactly complementary to the template strand (see Figure 1.16). In addition to a template strand and free nucleotides,

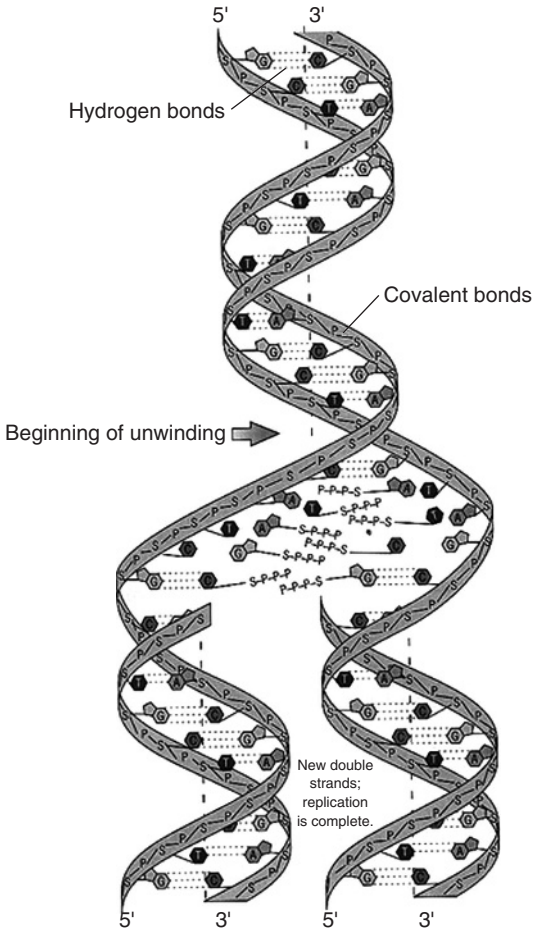

FIGURE 1.16. Diagram of DNA replication showing synthesis of two complementary strands at a replication fork.

the DNA polymerase also requires a primer—a short piece of DNA that is complementary to the template. The primer binds to its complementary spot on the template to form the start of the new strand, which is then extended by the polymerase, adding one complementary base at a time, moving in the 5′→3′ direction. In natural DNA replication, the primer binds to specific spots on the chromosome known as the **origin of replication**.

This semiconservative replication process was demonstrated quite eloquently by the famous 1958 experiment of Meselson and Stahl. They grew bacteria in a solution that contained free DNA nucleotides that contained heavy ^{15}N atoms. After many generations, the bacterial DNA contained heavy atoms throughout. Then the bacteria were transferred to a growth medium that contained normal nucleotides. After one generation, all bacterial cells had DNA with half heavy and half light nitrogen atoms. After two generations, half of the bacteria had DNA with normal nitrogen and the other half had one heavy and one light DNA strand (Meselson and Stahl 1958). After every cell division, the two daughter cells both have chromosomes made up of DNA molecules that have one strand from the parent cell and the other strand that has been newly synthesized. This method of semiconservative DNA replication is common to all forms of life on earth from bacteria to humans.

This mechanism of DNA replication has been exploited in modern DNA sequencing biochemistry, which often uses DNA polymerase from bacteria or other organisms to copy human (or any other) DNA. Key aspects of the replication process to keep in mind are that the DNA is copied linearly one base at a time from a specific starting point (origin), which is matched by a short primer of complementary sequence. The primer is extended by the reaction as new nucleotides are added, so that the primer becomes part of the newly synthesized complementary strand.

TRANSCRIPTION

The DNA in the chromosomes contains genes that are instructions for the manufacture of proteins, which in turn control all of the metabolic activities of the cell. In order for the cell to use these instructions, the genetic information must be moved from the chromosomes inside the nucleus out to the cytoplasm where proteins are manufactured. This information transfer is done using messenger RNA (**mRNA**) as an intermediary molecule. RNA (ribose nucleic acid) is a polymer of nucleotides, chemically very similar to DNA, but with three distinct differences: (1) RNA is a single-stranded molecule, so it does not form a double helix; (2) RNA nucleotides contain ribose rather than deoxyribose sugars; and (3) RNA uses uracil in place of thymine, so the common abbreviations for RNA bases are A, U, G, and C. As a result of these chemical differences, RNA is much less stable in the cell. In fact, the average RNA molecule has a lifespan that can be measured in minutes while DNA can be recovered from biological materials that are many thousands of years old.

The transcription of DNA into mRNA is similar to DNA replication. A single strand of DNA is copied one base at a time into a complementary strand of RNA. The enzyme RNA polymerase catalyzes the incorporation of free RNA nucleotides into the growing chain (see Figure 1.17). However, not all of the DNA is copied into RNA—only those portions that encode genes. In eukaryotic cells, only a small fraction of the total DNA is actually used to encode genes. Furthermore, not all genes are transcribed into mRNA in equal amounts in all cells. The process of transcription is tightly regulated so that only those mRNAs are manufactured that encode the proteins that are currently needed by each cell. This overall process is known as **gene expression**. Understanding the process of gene expression and how it differs in different types of cells or under different conditions is one of the fundamental questions driving the technologies of genomics.

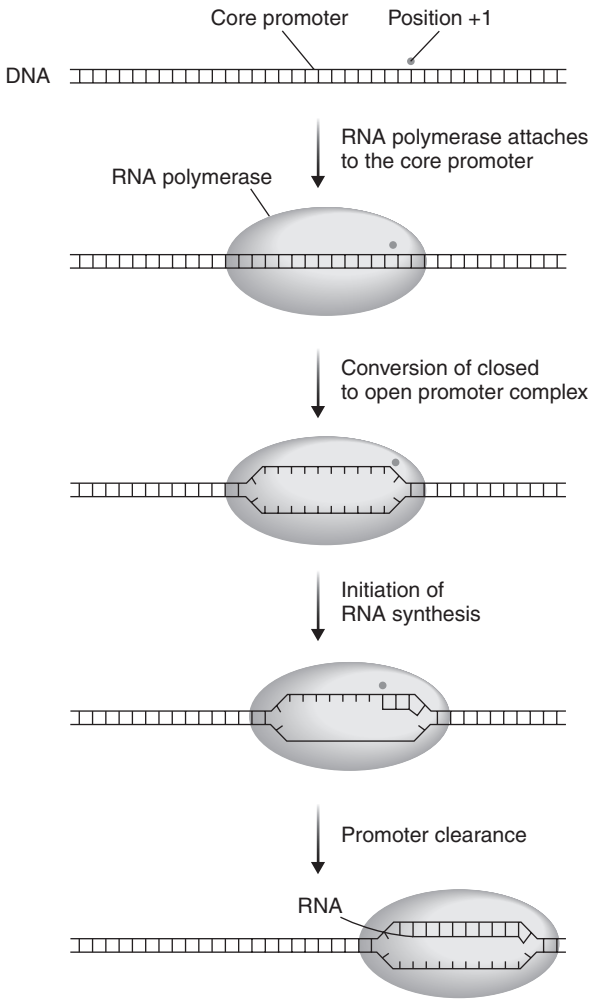

FIGURE 1.17. RNA polymerase II attaches to the promoter and begins transcription.

The primary control of transcription takes place in a region of DNA known as the **promoter**, which occupies a position "upstream" (in the 5′ direction) from the part of a gene that will be transcribed into RNA (the **protein-coding region** of the gene). A huge variety of different proteins recognize specific DNA

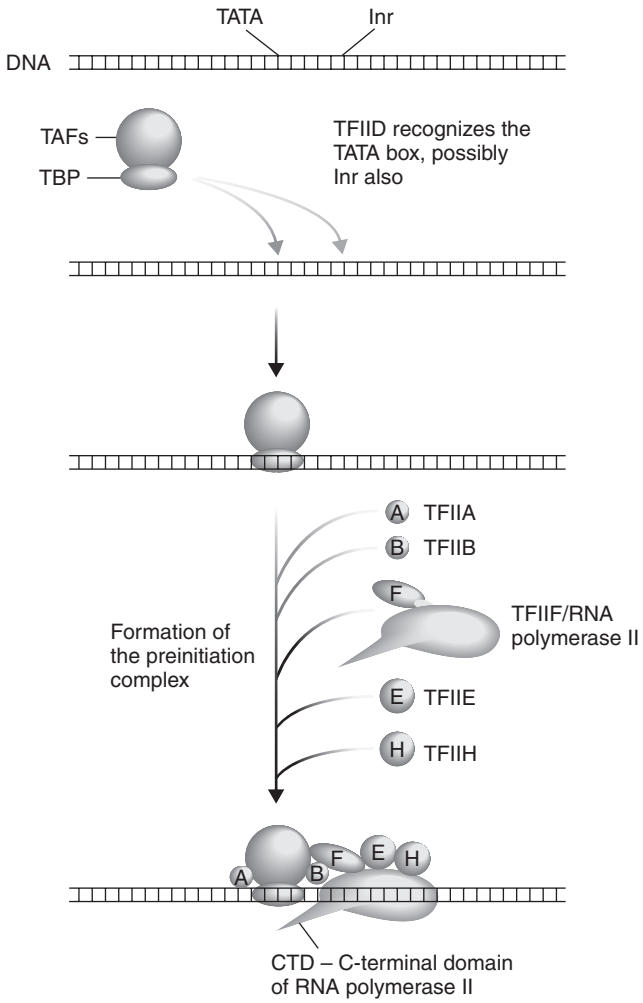

FIGURE 1.18. RNA polymerase II is actually a complex structure composed of many individual proteins.

sequences in this promoter region and bind to the DNA and either assist or block the binding of the RNA polymerase enzyme (see Figure 1.18). These DNA binding proteins work in concert to provide very fine-grained control of the expression of each gene depending on the type of cell, where it is located in the body, its

current metabolic condition, and its responses to external signals from the environment or from other cells.

In fact, the factors governing the assembly of the set of proteins involved in regulating DNA transcription is much more complicated than the sum of a set of DNA sequences neatly located in a promoter region 5′ to the coding sequence of a gene. In addition to the double helix, DNA has tertiary structures that involve twists and supercoils as well as winding around histone proteins. These three-dimensional (3D) structures can bring distant regions of a DNA molecule into close proximity, so that proteins bound to these sites may interact with the proteins bound to the promoter region. These distant sites on the DNA that may effect transcription are known as **enhancers**. The total set of DNA binding proteins that interact with promoters and enhancers are known as **transcription factors**, and the specific DNA sequences to which they bind are called **transcription factor binding sites**.

RNA Processing

Once a gene is transcribed into RNA, the RNA molecule undergoes a number of processing steps before it is translated into protein. First a 5′ cap is added, then a polyadenine tail is added at the 3′ end. In addition, eukaryotic genes are broken up into protein coding **exon** regions separated by non-protein coding **introns**, which are spliced out. This splicing is sequence-specific and highly precise, so that the final product contains the exact mRNA sequence that codes for a specific protein with not a single base added or lost (see Figure 1.19).

Each of these posttranscriptional processes may serve as a point of regulation for gene expression. Capping, polyadenelation, and/or splicing may be blocked, or incorrect splicing may be promoted under specific metabolic or developmental conditions. In addition, splicing may be altered in order to produce different mRNA molecules.

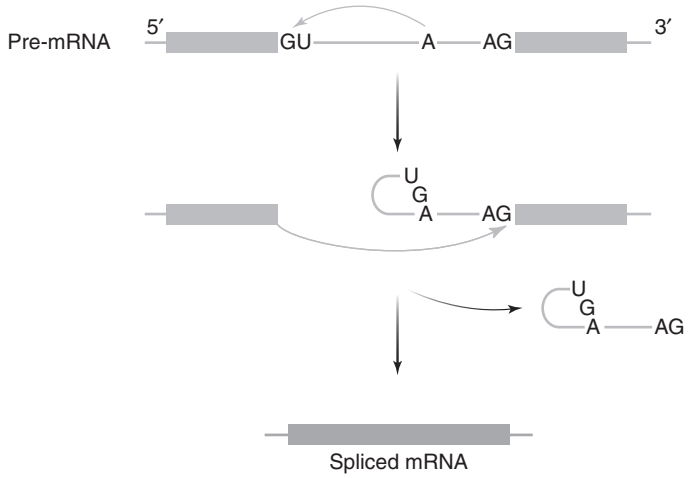

FIGURE 1.19. Model of intron splicing to form a mature mRNA from a pre-mRNA transcript.

ALTERNATIVE SPLICING

Each gene does not encode a single protein, as was originally suggested by the studies of *Neurospora* enzymes by Beadle and Tatum (1941). In many cases, there are several alternate forms of final spliced mRNA that can be produced from a single pre-mRNA transcript—potentially leading to proteins with different biological activities. In fact, current estimates suggest that most genes have multiple alternate splice forms. Alternate splicing my involve the failure to recognize a splice site, causing an intron to be left in, or an exon to be left out. Alternate splice sites may occur anywhere, either inside exons or introns, so that the alternate forms of the final mRNAs may be longer or shorter, contain more or fewer exons, or portions of exons (see Figure 1.20). Thus, each different splice form produced from a gene is a unique type of mRNA, which has the potential to produce a protein with different biochemical properties.

It is not clear how alternative splicing is controlled. The signals that govern RNA splicing may not be perfectly effective, or

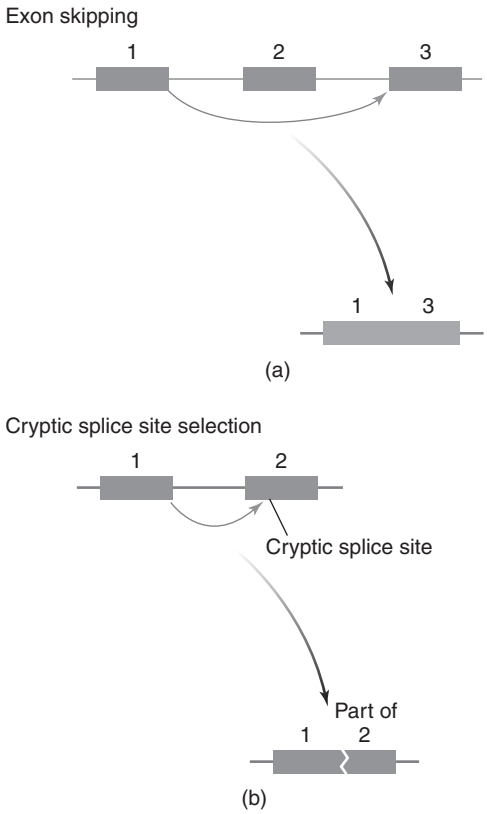

FIGURE 1.20. Two forms of alternative splicing: (a) exon skipping; (b) cryptic splice site selection.

RNA splicing may be actively used as a form of gene regulation. It is entirely possible for the products of other genes to interact with RNA splicing factors—perhaps in conjunction with external signals—to alter RNA splicing patterns for specific genes. The net result will be many different forms of mRNA, some produced only under specific circumstances of development, tissue specificity, or environmental stimuli. Thus, under some conditions a different protein with an added (or removed) functional domain will be produced from a gene, resulting in different protein function.

"Alternative splicing increases protein diversity by allowing multiple, sometimes functionally distinct proteins to be encoded by the same gene" (Sorek and Amitai 2001). The totality of all of these different mRNAs is being called the "transcriptome," which is certainly many times more complex than the genome. The relative levels of alternate splice forms for a single gene may have substantial medical significance. For example, there are 60 kinase enzymes that have alternate splice forms that do not include their catalytic domains, creating proteins that may function as competitive inhibitors of the full-length proteins (Sorek and Amitai 2001).

TRANSLATION

In order for a gene to be expressed, the mRNA must be translated into protein. This theory behind this process was encapsulated quite neatly in 1957 by Crick's diagram of the Central Dogma, but the details of the information flow from DNA to mRNA to protein took another decade to work out. It was immediately clear that the cell must solve several different problems of information storage and transmission. Huge amounts of information must be stored in the simple 4-letter code of DNA, it must be translated into the quite different 20-letter code of amino acids, and a great deal of punctuation and regulatory information must also be accounted for. The problem of encoding 20 different amino acids in the 4-letter DNA/RNA alphabet intrigued information scientists, and physicists as well as biologists and many ingenious incorrect answers were proposed. The actual solution to this problem was worked out with brute-force biochemistry by Har Gobind Khorana (Soll et al. 1965) and Marshall W. Nirenberg (Nirenberg 1965) by creating an *in vitro* (test tube) system where pure pieces of RNA would be translated into protein. They then fed the system with RNA molecules of very simple sequence and analyzed the proteins that were produced. With several years of

Universal Genetic Code

TTT phe F	TCT ser S	TAT tyr Y	TGT cys C
TTC phe F	TCC ser S	TAC tyr Y	TGC cys C
TTA leu L	TCA ser S	TAA OCH Z	TGA OPA Z
TTG leu L	TCG ser S	TAG AMB Z	TGG trp W
CTT leu L	CCT pro P	CAT his H	CGT arg R
CTC leu L	CCC pro P	CAC his H	CGC arg R
CTA leu L	CCA pro P	CAA gln Q	CGA arg R
CTG leu L	CCG pro P	CAG gln Q	CGG arg R
ATT ile I	ACT thr T	AAT asn N	AGT ser S
ATC ile I	ACC thr T	AAC asn N	AGC ser S
ATA ile I	ACA thr T	AAA lys K	AGA arg R
ATG met M	ACG thr T	AAG lys K	AGG arg R
GTT val V	GCT ala A	GAT asp D	GGT gly G
GTC val V	GCC ala A	GAC asp D	GGC gly G
GTA val V	GCA ala A	GAA glu E	GGA gly G
GTG val V	GCG ala A	GAG glu E	GGG gly G

FIGURE 1.21. Translation table for the eukaryotic nuclear genetic code.

effort (1961–1965), they defined a code of 64 three-letter RNA **codons** that corresponded to the 20 amino acids (with redundant codons for most of the amino acids) and 3 "stop" codons that caused the end of protein synthesis (see Figure 1.21). Also in 1965, Robert W. Holley established the exact chemical structure of **tRNA** (**transfer RNA**), the adapter molecules that carried each amino acid to its corresponding 3-base codon on the mRNA (Holley 1965). There is one specific type of **tRNA** that binds each type of amino acid, but each tRNA has an **anti-codon** which can bond to several different mRNA codons. Holley, Khorana, and Nirenberg shared the 1968 Nobel Prize in Physiology or Medicine for this work.

The translation process is catalyzed by a complex molecular machine called a **ribosome**, which is composed of both protein and **rRNA** (ribosomal RNA) elements. Proteins are assembled from free amino acids in the cytoplasm that are carried to the site of protein synthesis on the ribosome by the tRNAs. The tRNAs contain an anticodon region that matches the three-nucleotide codons on the mRNA. As each tRNA attaches to the anticodon,

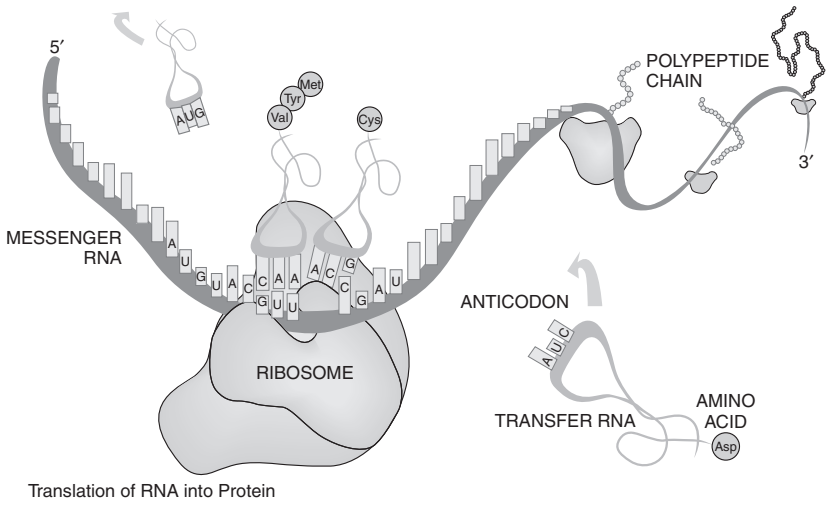

FIGURE 1.22. A diagram of the ribosome interacting with tRNAs as it translates an mRNA into a polypeptide chain.

the amino acid that it carries forms a bond with the growing polypeptide chain; then the tRNA is released and the ribosome moves down the mRNA to the next codon. When the ribosome reaches a stop codon, the chain of amino acids is released as a complete polypeptide (see Figure 1.22).

REFERENCES

Avery OT, MacLeod CM, McCarty M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J Exp Med* **79**:137–158.

Beadle GW, Tatum EL. 1941. Genetic control of biochemical reactions in Neurospora. *Proc Natl Acad Sci USA* **27**:499–506.

Chargaff E. 1950. Chemical specificity of nucleic acids and mechanisms of their enzymatic degradation. *Experientia* **6**:201–209.

Crick FHC. 1957. Nucleic acids. *Sci Am* **197**:188–200.

Hershey AD, Chase M. 1952. Independent functions of viral proteins and nucleic acid in growth of bacteriophage. *J. Gen Physiology* **36**:39–56.

Holley RW. 1965. Structure of an alanine transfer ribonucleic acid. *JAMA* **194**:868–871.

Lander ES et. al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**:860–921.

McCarty, M. 1985. *The Transforming Principle: Discovering that Genes Are Made of DNA*. Norton, New York.

Mendel, G. 1866. Versuche über Pflanzen-Hybriden. *Verhandlungen des naturforschenden Vereines, Abhandlungen, Brünn* **4**:3–47.

Meselson M, Stahl FW. 1958. The replication of DNA in Escherichia coli. *Proc Natl Acad Sci USA* **44**:671–682.

Morgan TH. 1910. Sex-limited inheritance in Drosophila. *Science* **32**:120–122.

Morgan TH. 1919. *The Physical Basis of Heredity*. Lippincott, Philadelphia.

Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottmann F, O'Neal C. 1965. RNA codewords and protein synthesis. VII. On the general nature of the RNA code. *Proc Natl Acad Sci USA* **53**:1161–1168.

Pauling L, Corey R. 1951. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci USA* **37**:235–240.

Sayre A. 1975. *Rosalind Franklin and DNA*. Norton, New York.

Soll D, Ohtsuka E, Iones DS, Lohrmann R, Hayatsu H, Nishimura S, Khorana HG. 1965. Studies on polynucleottides. XLIX. Stimulation of the binding of aminoacyl-sRNAs to ribosomes by ribotrinucleotides and a survey of codon assignments for 20 amino acids. *Proc Natl Acad Sci USA* **54**:1378–1385.

Sorek R, Amitai M. 2001. Piecing together the significance of splicing. *Nat Biotechnol* **19**:196.

Sutton W. 1903. The chromosomes in heredity. *Biol Bull* **4**:231–251.

Venter JC et al. 2001. The sequence of the human genome. *Science* **291**:1304–1351.

Watson JD, Crick FHC. 1953. A structure for deoxyribose nucleic acid. *Nature* **171**:737.