Chapter 1

Basics of Molecular Biology

Deborah Ann Payne

Introduction

Molecular biology entails the analysis and study of the chemical organization of the cell. Molecules comprise the smallest chemical component capable of performing all the activities (structural or catalytic) of a substance. One or more atoms constitute each molecule. This chapter describes the physical organization of cells, cellular organelles, and molecules important in cell division, inheritance, and protein synthesis.

Organization of the Cell

The cell is a mass of protoplasm surrounded by a semipermeable membrane.[1] Cells constitute the smallest element of living matter capable of functioning independently; however, within complex organisms, cells may require interaction with other cells. To function independently, cells must produce nucleic acids, proteins, lipids, and energy. In complex organisms, these organic processes form and maintain tissues and the organism as a whole.

Genes consist of discrete regions of nucleic acids that encode proteins, and control the function of the cell. Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) comprise the two types of nucleic acids found in all cells. Chromosomes, made up of double-stranded DNA complexed with proteins, contain all the genes required for the cell to live and function.

Prokaryotic Cells

Prokaryotic cells are simple organisms lacking subcellular compartments, such as bacteria. The majority of prokaryotic nucleic acids form circular strands comprising approximately 1×10^6 base pairs (bp) (Table 1-1). Additional extrachromosomal genetic elements consist of circular plasmids also known as episomes and linear mobile genetic elements called transposons (30–40 bp). Plasmids range in size from 33 bp to 230 bp[2] and first gained notoriety in the 1950s by being associated with antibiotic resistance in bacteria.[3,4] Transposons also may confer antibiotic resistance on the host bacteria. All these genetic elements exist in direct contact with the bacteria's cytoplasm.

Eukaryotic Cells

Cytoplasm

In contrast to prokaryotic cells, eukaryotic cells are complex, highly compartmentalized structures. The cytoplasm contains multiple membrane-bound compartments known as organelles. The cellular membrane separates the cellular cytoplasm from the external environment. The membranes consist of hydrophobic lipid bilayers. The lipid bilayer contains proteins that serve as receptors and channels.

Nucleus and Nucleolus

The nucleus of the cell contains the cell's linear chromosomes and serves as the primary locus of inherited genetic material. Inner- and outer-pore-containing membranes define the nucleus and separate the chromosomes from the surrounding cytoplasm. Further partitioning occurs within the nucleus to generate the nucleolus, which functions as the ribosome-generating factory of the cell. Instead of additional membranes, fibrous protein complexes separate the nucleolus from the rest of the nucleus. In this structure, the nucleolus organizer (a specific part of a chromosome containing the genes that encode ribosomal RNA) interacts with other molecules to form immature large and small ribosomal subunits. Following processing, immature subunits exit the nucleolus and enter the nucleus. Eventually, mature ribosomal subunits and other molecules exit the nucleolus through the nuclear pores and enter the cytoplasm.

Table 1-1. Comparison of DNA Sizes of Various Genetic Elements

Genomic Element	Size in Base Pairs
Human chromosome	1–3 × 10^9
Bacterial chromosome	1–4 × 10^6
Mitochondrial chromosome	16,569
Bacteriophage	39,000
CAM plasmid	230
R388 plasmid	33
Transposons	30–40

Mitochondria

Mitochondria are membrane-bound organelles within the cytoplasm of cells that have several cellular functions. Inheritable genetic material, independent from the nuclear chromosomes, resides in mitochondria. These maternally derived organelles contain their own circular chromosomes (16,569 bp) and replicate independently from the cell and one another. As a result, not all mitochondria in a given cell have the same mitochondrial chromosomal sequence, resulting in genetic diversity of these organelles within and between different cells of the same organism, which is known as heteroplasmy. Mitochondrial genes encode mitochondria-specific transfer RNA molecules (tRNA). In addition, the mitochondrial chromosomes contain genes that encode proteins used in oxidative phosphorylation, including subunits of the cytochrome c oxidase, cytochrome b complex, some of the ATPase complex and various subunits of NAD dehydrogenase. Other components of the oxidative phosphorylation pathway are encoded by nuclear genes. For this reason, not all mitochondrial genetic diseases demonstrate maternal transmission. Analysis of mitochondrial DNA has applications for diagnosis of mitochondrial-inherited genetic diseases as well as for forensic purposes in the identification of severely decomposed bodies.

Other Cellular Organelles

Membranes not only segregate heritable genetic molecules into the nucleus and mitochondria, but also separate various cellular functions into distinct areas of the cell. The compartmentalization of cellular functions, such as molecular synthesis, modification, and catabolism, increases the local concentration of reactive molecules, thus improving the cell's biochemical efficiency. This partitioning also protects inappropriate molecules from becoming substrates for these processes. One example of this segregation is the endoplasmic reticulum (ER), which consists of a complex of membranous compartments where proteins are synthesized. Glycoproteins are synthesized by ribosome-ER complexes known as rough ER (RER), while lipids are produced in the smooth ER. The Golgi apparatus consists of numerous membrane-bound sacs where molecules generated in the ER become modified for transportation out of the cell. In addition, peroxisomes and lysosomes segregate digestive and reactive molecules from the remainder of the cellular contents to prevent damage to the cell's internal molecules and infrastructure.

Biological Molecules

Carbon can covalently bond to several biologically important atoms (i.e., oxygen, hydrogen, and nitrogen) and forms the scaffold for all biomolecules. Basic subunit biomolecules can combine to form more complex molecules such as carbohydrates, nucleic acids, and amino acids.

Carbohydrates

Carbohydrates serve as energy reservoirs and are a component of nucleic acids. In addition, carbohydrates also attach to lipids and proteins. The basic unit of a carbohydrate consists of the simple sugars or monosaccharides. These molecules have carbon, oxygen, and hydroxyl groups that most commonly form ringed structures. The oxygen can react with the hydroxyl group of another simple sugar to form a chain. As a result, the formula for a simple sugar is $(CH_2O)_n$, where n represents various numbers of these linked building block units.

Two pentose sugars, deoxyribose and ribose, comprise the sugar element of DNA and RNA molecules, respectively. As the name indicates, deoxyribose ("de-," a prefix meaning "off" and "oxy," meaning "oxygen") lacks one hydroxyl (OH) group compared to ribose.

Nucleic Acids

Nucleic acids are composed of chains of nucleotides. Each nucleotide is composed of a sugar (either ribose or deoxyribose), a phosphate ($-PO_4$) group, and a purine or pyrimidine base. The nucleotides are joined into a DNA or RNA strand by a sugar-phosphate-linked backbone with the bases attached to and extending from the first carbon of the sugar group. The purine and pyrimidine bases are weakly basic ring molecules, which form N-glycosidic bonds with ribose or deoxyribose sugar. Purines are comprised of two rings, a six-member ring and a five-member ring ($C_5H_4N_4$), while pyrimidines consist of a single six-member ring ($C_4H_2N_2$). Purines (guanine, G, and adenine, A) pair with pyrimidines (cytosine, C, and thymine, T) via hydrogen bonds between two DNA molecules (Figure 1-1). The additional hydrogen bond that forms between G and C base pairing (i.e., three hydrogen bonds) dramatically enhances the strength of this interaction compared to the two hydrogen bonds present between A and T nucleotides. This hydrogen-bonding capacity between G:C and A:T forms a pivotal molecular interaction for all nucleic acids and assures the passage of genetic information during

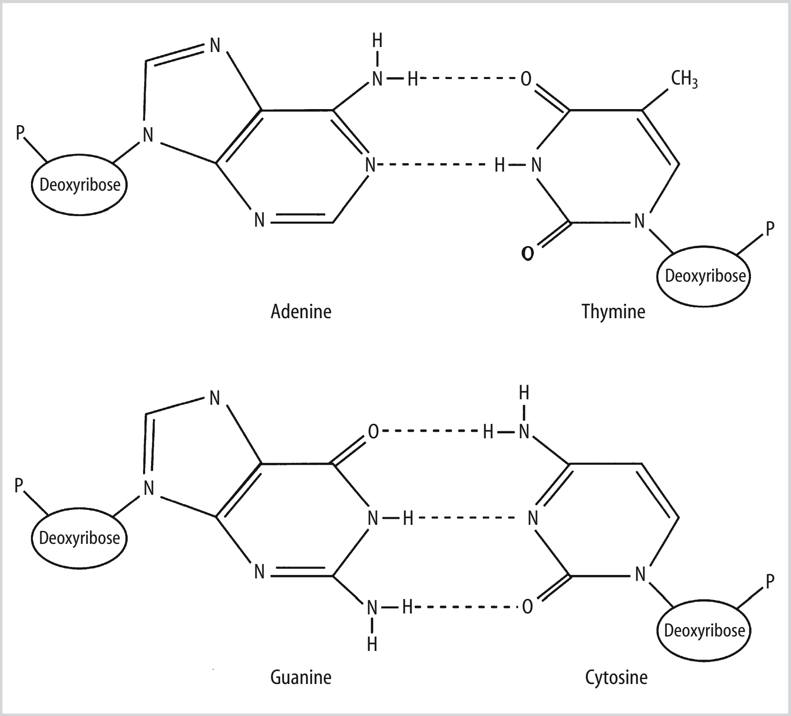

Figure 1-1. DNA base pairing. DNA nucleotides are composed of three moieties (e.g., sugar, base, and phosphate groups). The bases are either purine (adenine and guanine) or pyrimidine (thymine and cytosine). Note the difference in hydrogen bonds between adenine and thymine base pairs, with two hydrogen bonds, compared to cytosine and guanine base pairs, with three hydrogen bonds. (Reprinted from Leonard D. *Diagnostic Molecular Pathology*, copyright 2003, with permission from Elsevier.)

DNA replication, RNA synthesis from DNA (transcription), and the transfer of genetic information from nucleic acids to the amino acids of proteins.

Amino Acids

Amino acids are the building blocks of proteins. Amino acids linked together via peptide bonds form large, complex molecules. Amino acids consist of an amino group (NH_3), a carboxy group (COO–), an R group, and a central carbon atom. The R group can be a simple hydrogen, as found in glycine, or as complex as an imidazole ring, as found in histidine. Twenty different R groups exist; and determine whether an amino acid has a neutral, basic, or acidic charge (Table 1-2). The amino group of the polypeptides is considered the beginning of a protein (N-

Table 1-2. Amino Acids

	Amino Acid Symbols		
Amino Acid	Three Letter	Single Letter	R Group
Alanine	ala	A	CH_3—$CH(NH_2)$—COOH
Arginine	arg	R	HN=$C(NH_2)$—NH—$(CH_2)_3$—$CH(NH_2)$—COOH
Asparagine	asn	N	H_2N—CO—CH_2—$CH(NH_2)$—COOH
Aspartic acid	asp	D	HOOC—CH_2—$CH(NH_2)$—COOH
Cysteine	cys	C	HS—CH_2—$CH(NH_2)$—COOH
Glutamine	glu	Q	H_2N—CO—$(CH_2)_2$—$CH(NH_2)$—COOH
Glutamic acid	gln	E	HOOC—$(CH_2)_2$—$CH(NH_2)$—COOH
Glycine	gly	G	NH_2—CH_2—COOH
Histidine	his	H	NH—CH=N—CH=C—CH_2—$CH(NH_2)$—COOH
Isoleucine	ile	I	CH_3—CH_2—$CH(CH_3)$—$CH(NH_2)$—COOH
Leucine	leu	L	$(CH_3)_2$—CH—CH_2—$CH(NH_2)$—COOH
Lysine	lys	K	H_2N—$(CH_2)_4$—$CH(NH_2)$—COOH
Methionine	met	M	CH_3—S—$(CH_2)_2$—$CH(NH_2)$—COOH
Phenylalanine	phe	F	Ph—CH_2—$CH(NH_2)$—COOH
Proline*	pro	P	NH—$(CH_2)_3$—CH—COOH
Serine	ser	S	HO—CH_2—$CH(NH_2)$—COOH
Threonine	thr	T	CH_3—CH(OH)—$CH(NH_2)$—COOH
Tryptophan	trp	W	Ph—NH—CH=C—CH_2—$CH(NH_2)$—COOH
Tyrosine	tyr	Y	HO—p—Ph—CH_2—$CH(NH_2)$—COOH
Valine	val	V	$(CH_3)_2$—CH—$CH(NH_2)$—COOH

*Proline has a ring shape arising from the covalent bond formed between the amino group and the central carbon.

terminus), while the carboxyl group is at the opposite end, providing directionality to the protein.

Genetic Molecules

Nucleic acids encode genetic information but also participate in additional physiological processes ranging from metabolism to energy transfer. Nucleotides constitute the monomeric units of nucleic acids (Figure 1-1). Nucleosides consist of two components (ribose or deoxyribose in RNA and DNA, respectively, and either a purine or pyrimidine base). A nucleotide is produced from a nucleoside by the addition of one to three phosphate groups through a covalent bond with the hydroxyl group of the 5′ carbon of the nucleoside's sugar ring.

Nucleic acids are formed by chains of nucleotides linked by phosphodiester bonds between the 3′ carbon of the first nucleotide's sugar ring and the 5′ carbon of the adjacent nucleotide's sugar ring. The phosphodiester linkages cause nucleic acids to have a 5′ to 3′ directionality. The alternating sugar-phosphate chain forms a continuous molecule with bases extending from the 1′ carbon of each sugar. For this reason, the sugar-phosphate chain is referred to as the backbone of nucleic acids (Figure 1-2). The phosphate groups give nucleic acids a negative charge that imparts important physiochemical properties to nucleic acids. The negative charge of DNA facilitates the binding of mammalian DNA to various proteins and allows separation of nucleic acid molecules by charge and size during gel or capillary electrophoresis.

Structure

In double-stranded DNA, the two DNA strands are held together by exact A:T and G:C hydrogen bonding between the bases of the two strands, in which case the two strands are said to be complementary. The two strands are oriented in opposite 5′ to 3′ directions, such that one strand is oriented 5′ to 3′ (↓) and the complementary strand is oriented 3′ to 5′ (↑) in an antiparallel fashion (see Figure 1-2). In this case, "anti-" refers to the head (or 5′ end) of one DNA strand being adjacent to the tail (or 3′ end) of the opposite strand.

The molecular curves of the two DNA strands form antiparallel helices known as the DNA double helix. This double helix form (the B form) has ten nucleotide pairs (base pairs) per turn, occupying 3.4 nm. Because the bonds between the sugar and the base are not perfectly symmetrical, the strands curve slightly. The slight curve of the offset glycosidic bonds results in major and minor grooves characteristic of the B form of the double helix.[5] Many molecular diagnostic assays target the minor groove of DNA with sequence-specific probes known as minor groove binding (MGB) probes. Two other forms of DNA exist as the Z and A forms. The Z form acquires a zigzag shape, while the A form has a very shallow and very deep groove.

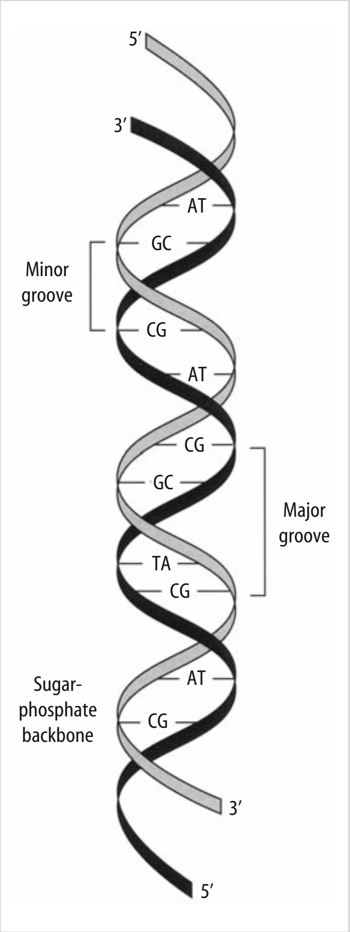

Figure 1-2. Double-stranded DNA. The two DNA strands are oriented in an antiparallel relationship, with asymmetric base pairing of two DNA strands that generates the minor and major grooves of the DNA double helix. (Reprinted from Leonard D. *Diagnostic Molecular Pathology*, copyright 2003, with permission from Elsevier.)

Thermodynamics of Nucleotide Base Pairing

Thermodynamics plays a major role in the structure and stability of nucleic acid molecules. The core mechanism of nucleic acid thermodynamics centers on the hydrogen-bonding capabilities of the nucleotides. The stability of these interactions not only influences the formation and stability of duplex nucleic acids but also impacts the structure and catalytic characteristics of single-stranded nucleic acids through intramolecular base pairing. In addition to these physiological functions, the phenomenon of complementary base pairing profoundly impacts clinical diagnostic assay development. Prior to the advent of clinical molecular diagnostic testing, many diagnostic tests required obtaining an antibody to identify or detect a target protein. The procedures for generating and validating diagnostic antibodies required extensive time and expense. The application of techniques utilizing the capability of two molecules to form a base pair as the basis for detection and characterization of target nucleic acids has greatly facilitated diagnostic test development. The formation of hydrogen bonding between two pieces of nucleic acid is called hybridization, or annealing, and the disruption of the hydrogen bonds holding two nucleic acid molecules together is called denaturation, or melting. The fact that molecular diagnostic tests use hybridization techniques based on A : T and G : C base pairing underscores the necessity for understanding the thermodynamics of hydrogen base pairing of nucleic acids.

Short pieces of DNA or RNA called probes, or primers, that contain a specific sequence complementary to a disease-related region of DNA or RNA from a clinical specimen are frequently used in the molecular pathology laboratory. To achieve hybridization of a DNA or RNA probe to genomic DNA for a diagnostic test, the two genomic DNA strands must be separated, or denatured, prior to probe hybridization. Increasing the temperature of a DNA molecule is one mechanism for disrupting the hydrogen bonds between the DNA base pairs and denaturing double-stranded DNA into single-stranded form. The temperature at which 50% of the double-stranded DNA molecules separate into single-stranded form constitutes the melting temperature (T_m). The shorter the two complementary DNA molecules are, the easier it is to calculate the T_m. This primarily results from the decreased likelihood of nonspecific intramolecular annealing or base pairing compared to inter- and intramolecular base pairing. The simplest and least accurate formula for determining the T_m for short double-stranded DNA multiplies the sum of the G : C base pairs by 4 and multiplies the sum of the A : T base pairs by 2 and then adds these numbers together.

$$T_m = [4(G:C)] + [2(A:T)]$$

Although this is the least accurate method for calculation of the T_m of a double-stranded DNA molecule, it mathematically illustrates that G : C bonds are roughly twice the strength of A : T bonds. This formula works fairly well for short DNA molecules (i.e., ≤18 bp); however, as the length of the DNA molecule increases to 100 bp, the nearest neighbor T_m calculation for DNA and RNA is more accurate.[6,7]

$$T_m = \frac{\Delta H}{\Delta S + R\ln(Ct)} - 273.15$$

where

ΔH = enthalpy of the nucleic acid fragment
ΔS = entropy of the nucleic acid fragment
$R = 1.987\ \mathrm{cal\ K^{-1}\ mol^{-1}}$
Ct = total strand concentration

For longer sequences (>100 bp), the most accurate formula for calculation of T_m is as follows:[8]

$$T_m = 81.5 + 16.6 \log[NA] + 0.41\ [\%G = \%C\} - 0.65\ (\%\ \text{formamide}) - 675/\text{length} - \%\ \text{mismatch}$$

Table 1-3 demonstrates the effect of increasing the relative amounts of G : C base pairs on the T_m using these formulas.

Intramolecular base pairing also generates complex three-dimensional forms within single-stranded nucleic acid molecules. As a result, the single-stranded nature of eukaryotic RNA molecules affords great structural diversity via intramolecular base pairing. These conformations strain the linear RNA molecule and produce chemically reactive RNA forms. Catalytic RNA molecules play pivotal roles in cellular functions and in gene-targeting therapies.

Table 1-3. Melting-Temperature Calculations for Short Oligomers

Total Length	Number of G:C	Number of A:T	T_m*	%G:C†	A:T + G:C‡
30	30	0	106.2	100.0	100.0
30	25	5	101.2	93.2	100.0
30	20	10	89.5	79.5	90.0
30	10	20	83.4	72.7	80.0
30	0	30	71.6	59.0	60.0
20	20	0	90.4	88.8	80.0
20	10	10	72.7	65.1	60.0
20	0	20	55.9	47.8	40.0

*Nearest neighbor calculation of T_m.[6]
† T_m method for sequences over 100 bases.[8]
‡ $4(G + C) + 2(A + T)$ formula.

Intra- and intermolecular base pairing can negatively affect hybridizations. Dimers, bulge loops, and hairpin loops exemplify some of these interactions. Hairpins inhibit plasmid replication and attenuate bacterial gene expression.[2] These detrimental effects may also include initiation of spurious nonspecific polymerization, steric hindrance of hybridization of short stretches of nucleic acids (i.e., 10 to 30 base pieces of single-stranded nucleic acids, known as oligomers or primers), and depletion of probes or primers away from the specific target by either primer dimerization or other mechanisms. These interactions can result in poor sensitivity or poor specificity for diagnostic molecular tests.

Topology

The DNA and RNA molecules assume various geometric shapes or topologies that are independent of base pair interactions. Eukaryotic nucleic acids take on linear forms, in contrast to the circular forms of mitochondrial and bacterial chromosomal DNA. Viral genomes occur as different forms, ranging from segmented linear to circular. Although the conformation of RNA molecules can be complex via intramolecular base pairing, the topology of messenger RNA (mRNA) molecules is primarily linear. An organism's genomic topology influences the biochemical mechanisms used during replication and the number of replication cycles a given chromosome can undertake. In contrast to circular genomes, linear genomes limit the total number of possible replication cycles due to progressive shortening of the linear chromosome.

Mammalian Chromosomal Organization

The human genome contains approximately 10^9 base pairs of DNA. The total DNA is contained in 46 double-stranded

DNA pieces complexed with proteins to form chromosomes. The diploid human genome, therefore, is contained in 46 chromosomes: two of each of the 22 autosomal chromosomes, plus either two X chromosomes, in females, or one X and one Y chromosome, in males. Since the length of each helical turn of a double-stranded DNA molecule is 3.4 nm, consisting of ten bases, the length of the total genomic DNA in each cell measures approximately 1 m in length. For each cell to contain this long molecule, the double-stranded DNA must be compressed. A complex of eight basic histones (two copies of histone 2 [H2], H3, H4, and H5) package the DNA.[9] The histone complex contains positively charged amino acids that bind to 146 bases of negatively charged DNA. Histones fold the DNA either partially or tightly, resulting in compression of the DNA strand. Tight folding of the DNA condenses the DNA into heterochromatin. Following packaging and condensation, the nucleic acid strand widens from 2 nm to 1400 nm, with extensive overall shortening of the nucleic acid in the metaphase chromosome. Light microscopy easily permits the visualization of condensed metaphase chromosomes.

Less-condensed DNA binds histone 1 (H1) proteins or other sequence-specific DNA-binding molecules. Some of these DNA-binding molecules regulate gene expression (discussed later in this chapter). In contrast, tightly condensed chromosomes lack the "open spaces" for binding of regulatory proteins and prevent gene expression from tightly condensed DNA regions. These proteins may also prevent access to nucleic acid probes or primers for molecular diagnostic tests. As a result, many DNA extraction protocols include a protein-digestion step to liberate the DNA from these DNA-binding proteins. Removal of these proteins facilitates hybridization with short pieces of nucleic acid, such as primers or probes.

DNA Replication

Eukaryotic DNA Replication

The replication of DNA is a complex process requiring specific physiological temperatures and a host of proteins. As mentioned previously, molecular diagnostic techniques rely on the ability to denature or melt a double-stranded DNA template. Using chemical or physical conditions, separation of DNA strands can be accomplished with alkali or high temperatures (i.e., 95°C). Under physiological conditions, dissociation of DNA strands for replication is accomplished by numerous enzymes, such as helicases and topoisomerases. The region of transition from the double-stranded to separated single-stranded DNA is called the replication fork. The replication fork moves along the double-stranded DNA molecule as replication proceeds. At the replication fork, various primases, initiating proteins, and polymerases bind to the original or parental DNA strands and generate new daughter strands. Known collectively as a replisome, these enzymatic activities generate two new nucleic acid strands that are complementary to and base paired with each of the original two template or parent DNA strands. This replication process is known as semiconservative because each resulting double-stranded DNA molecule consists of one new and one old DNA strand (Figure 1-3).

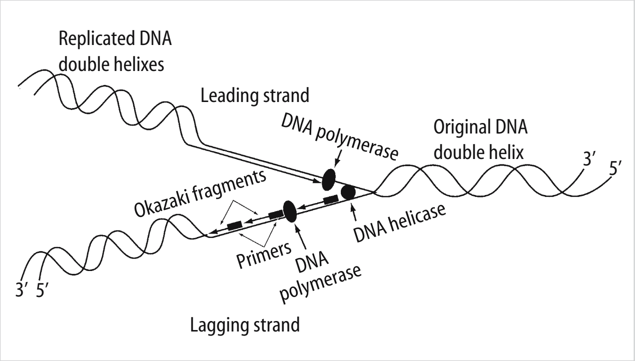

Figure 1-3. DNA replication. Replication fork depicting the leading and lagging strands and the numerous proteins and Okazaki fragments involved with replication. (Reprinted from Leonard D. *Diagnostic Molecular Pathology*, copyright 2003, with permission from Elsevier.)

Polymerases function to synthesize new nucleic acid molecules from nucleotide building blocks. The sequence of the new strand is based on the sequence of an existing nucleic acid molecule, and the polymerase adds nucleotides according to the order of the bases of the parent strand, using G : C and A : T pairing. The new strand is antiparallel to the parent strand and is synthesized in a 5′ to 3′ direction. Of the two parent strands of genomic DNA, one strand (called the leading strand) can be read by the polymerase continuously in a 3′ to 5′ direction, with the new strand generated in a continuous 5′ to 3′ direction. In contrast, the opposite strand (known as the lagging strand) cannot be read continuously by the polymerase. The replication fork moves along the lagging strand in a 5′ to 3′ direction, and polymerases synthesize only by reading the parent strand in a 3′ to 5′ direction while synthesizing the new strand in a 5′ to 3′ direction. Therefore, synthesis cannot proceed continuously along the lagging strand, which must be copied in short stretches primed from RNA primers and forming short DNA fragments known as Okazaki fragments. The new strand complementary to the lagging strand is formed by removal of the RNA primer regions and ligation of the short DNA fragments into a continuous daughter strand complementary to the lagging strand. Discontinuous 3′ to 5′ replication results in the progressive loss of ends of the chromosomes known as telomeres in normal cells. Some malignant cells retain telomerase activity that permits the addition of these terminal telomeric sequences to the chromosomes.

While replication requires many proteins, the polymerase determines the speed and accuracy of new strand synthesis. The rate that the four nucleotides are polymerized into a nucleic acid chain defines the processivity of the enzyme. The processivity of most polymerases approximates 1000 bases per minute.

Table 1-4. Fidelity of Various Polymerases

Polymerase	Error Rate
pol β*	8×10^{-4}
pol α*	1×10^{-4}
pol θ*	$1.7–4 \times 10^{-5}$
Pfu†	1.3×10^{-6}
Deep vent†	2.7×10^{-6}
Vent†	2.8×10^{-6}
Taq†	8×10^{-6}
UITma†	55×10^{-6}
Klenow‡	$1–10 \times 10^{-7}$
HIV reverse transcriptase	$6–30 \times 10^{-4}$

* Reference 37.
† Reference 38.
‡ Reference 39.

The fidelity of the polymerase refers to the accuracy of the enzyme to incorporate the correct complementary bases in the newly synthesized DNA. Incorporation of incorrect bases or other replication errors can result in cell death or oncogenesis. The error rate of polymerases varies widely from 1 in 1200 to 1 in 1,000,000 bases (Table 1-4). To correct the erroneous incorporation of bases or other replication errors, protein complexes proofread and correct synthesis errors. In normal cells, the cell cycle pauses to facilitate error repair in the G2 phase of the cell cycle (Figure 1-4). Malignant cells may not pause to allow for error correction, resulting in the accumulation of damaged or mutated DNA.

The complexity of the biochemical reactions necessary for replicating eukaryotic nuclear DNA demonstrates a high degree of regulation for generating two strands from one replication fork. In addition to these complexities, replication in eukaryotic cells occurs at multiple origins of replication (Ori). These multiple sites grow progressively until the newly generated strands join to form complete chromosomal-length DNA.

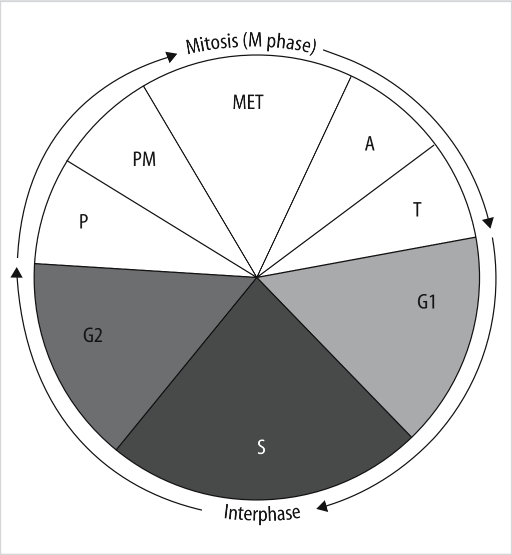

Figure 1-4. Cell cycle. The clear panels are the ordered phases of mitosis (M phase), while the gray and black panels are the ordered stages of interphase. P, prophase; PM, prometaphase; MET, metaphase; A, anaphase; T, telophase; G1, gap 1; S, DNA synthesis; G2, gap 2.

Bacterial and Mitochondrial Replication

The relatively small chromosomes of bacteria (~10^6 base pairs) utilize a simpler mechanism than eukaryotic replication. A single origin of replication initiates the duplication of the bacterial chromosome, and replication occurs simultaneously on both strands in opposite directions from the origin of replication. This efficient replication process depends on the circular topology of the bacterial genome.

Another unique feature of prokaryotic chromosomal replication is the mechanism bacteria have evolved to protect their chromosomes. The lack of a protective nuclear membrane in bacteria makes the chromosome susceptible to attack by viruses (bacteriophages). As a result, many bacteria produce restriction enzymes that degrade foreign nucleic acids. These restriction enzymes recognize specific short sequences and cleave the DNA at those sites. However, methylation of the recognition sequences in the bacterial chromosomal DNA prevents most restriction enzymes from digesting the chromosomal DNA of the bacteria. In this way, methylating enzymes add methyl groups to the replicated bacterial chromosome, preventing chromosomal degradation by its own restriction enzymes. This methylation and restriction process functions as a primitive immune system by destroying foreign bacteriophage DNA before it can usurp the bacteria's replication system. Bacterial restriction enzymes are used to specifically cleave DNA in molecular diagnostic tests and are useful for identifying genetic variations.

Additional types of replication occur in some viruses and bacteria. The rolling-circle mechanism of replication proceeds with an initial single-strand cut or nick in double-stranded circular genomes, followed by replication proceeding from the nick in a 5′ to 3′ direction. The new strand displaces the old strand. RNA viral genomes use the enzyme transcriptase for replication. In the case of retroviruses, a reverse transcriptase generates an intermediate DNA molecule, which integrates into the host chromosome and then is used for generation of progeny RNA molecules. The high error rate of human immunodeficiency virus (HIV) reverse transcriptase produces numerous mutations in the viral genome.[10,11] Some of these mutations confer resistance to antiretroviral therapies and can be identified by clinical molecular tests.

Cell Division and Cell Cycle

In eukaryotic cells, the cell cycle refers to the entire process of generating two daughter cells from one original cell, with chromosomal replication as one of the steps. The two parts of the cell cycle are called interphase and mitosis.

DNA synthesis occurs during interphase and consists of three stages: gap 1 (G1), synthesis (S), and gap 2 (G2) (Figure 1-4). Regulation of cell division depends on specific cell-cycle-dependent proteins known as cyclins and growth factors. Some of these factors cause the cycle to progress while others stop the cycle at certain stages. Checkpoints, or times when the cycle may be paused, exist at the G1/S and G2/mitosis interfaces and allow the cell time to repair any DNA damage that may be present in the cell before and after replication of the DNA, respectively.

Growth factors initiate the G1 phase via cell surface receptors. Several molecular events such as the dephosphorylation of the retinoblastoma protein and cyclin binding to cyclin-dependent kinases (Cdk) transition the cell toward the G1/S checkpoint. The amount of cellular P53 protein determines whether the cell progresses beyond this checkpoint, with higher levels preventing cell cycle progression. Because various DNA-damaging events, such as ultraviolet light, radiation, carcinogens, and double-stranded DNA breaks, induce production of P53 protein, this molecule serves as a sentinel for mutated DNA. The functional failure of P53 removes this sentinel from the cell cycle process and results in the accumulation of genetic errors. Therefore, inactivation of P53 facilitates oncogenesis.

Once DNA repairs have taken place during G1 prior to replication of the DNA, the cell proceeds to S phase. DNA synthesis occurs in the S phase, followed by the G2 phase. Replication errors occurring during the S phase are corrected in the G2 phase, the G2/M checkpoint. This final checkpoint marks the end of interphase.

Mitosis, the process of physical division of the parent cell into two daughter cells, occurs during the mitosis or M phase of the cell cycle. During mitosis, the duplicated chromosomes are physically separated so that each daughter cell receives the correct number of chromosomes. Mitosis consists of five phases: prophase, prometaphase, metaphase, anaphase, and telophase. The duplicated chromosomes condense during prophase. A structural element known as the mitotic spindle originates from two structures called centrioles, which move to opposite poles of the cell and the spindle forms between the centrioles. The nuclear membrance dissipates, proteins form kinetochores on the chromosmes, and microtubules attach to the kinetochores during prometaphase. The duplicated chromosome pairs attach at central points along the spindles. The arrangement of the highly condensed chromosome pairs along an equatorial cell plane denotes metaphase. As previously discussed, highly condensed chromosomes cannot bind proteins necessary for gene expression. As a result, the cell's internal machinery focuses solely on cell division during metaphase. The centriole-derived spindle guidelines pull the duplicate chromosomes apart and drag them toward each centriole during anaphase. With the separation of the daughter chromosomes (chromatids) into opposite poles of the cell and the reformation of nuclear membranes around the two daughter sets of chromosomes, telophase begins. Cytokinesis, or the division of the cytoplasm, is the last step in cell division. During cytokinesis, the mitochondria are randomly and potentially unevenly distributed in the daughter cells. The cell cycle can then be reinitiated by one or both of the daughter cells to generate additional cells. Alternatively, some cells become quiescent in a G0 phase (between telephase and G1) and either have a prolonged delay before initiating replication again or no longer divide.

Cell division to generate gametes is called meiosis and consists of two divisions, meiosis I and meiosis II. Like mitosis, this process begins with the duplication of chromosomes in prophase I. During metaphase I, the maternal and paternal homologous chromosomes pair (i.e., pairing occurs between each of the pairs of the 22 autosomal chromosomes, the two X chromosomes in females, and the X and Y chromosomes in males). Each pair attaches to the spindle apparatus along the equatorial plane of the cell spindle. DNA may be exchanged between the paired chromosomes by either crossing-over or recombination mechanisms during this pairing stage of meiosis I. During anaphase I, homologous chromosomes separate into daughter cells, resulting in 23 duplicated chromosomes in each daughter cell. A second cell-division cycle, meiosis II, separates the duplicated chromosomes, resulting in haploid cells, egg or sperm, containing only one copy of each of the 22 chromosomes plus an X (egg or sperm) or Y (only sperm) chromosome.

From Gene to Protein

The genomic DNA content is the same in all cells of the same person and encodes all the genetic information for cellular function. Encoded in the DNA are the blueprints for all the RNA and protein molecules present in any type of cell. Different parts of the genetic information are used by different types of cells to accomplish each cell's specific function. DNA is used to produce RNA and protein molecules by processes called transcription and translation, respectively. The regions of DNA that encode RNA and protein molecules are called genes.

Replication requires an increase in building materials for the duplicated daughter cells. Highly condensed metaphase chromatin cannot produce gene products because proteins that initiate gene expression cannot bind to the chromosomes at this phase of replication. In contrast, partially condensed or unfolded chromatin permits the binding of specific proteins (e.g., RNA polymerases) that synthesize mRNA and tRNA. Ultimately, these molecules facilitate the production of gene products, specifically proteins.

RNA molecules function as the mediators between DNA and protein. These molecules essentially speak the same language as DNA because, as nucleic acids, they can base pair with complementary DNA sequences. Like transferring spoken language to a written form, this process of copying information from DNA to RNA is referred to as transcription. The transcription complex of proteins must unwind the double-stranded DNA at the specific gene site to be copied, locate the polymerase binding site on one of the DNA strands, and generate a primary (1°) transcript, which is one component of heterogeneous nuclear RNA (hnRNA) by reading the DNA strand in a 3′ to 5′ direction, with RNA synthesis proceeding in a 5′ to 3′ direction. The 1° RNA transcript is processed into mRNA, and finally the DNA in the region of the gene becomes double-stranded again. Numerous DNA sequences bind proteins that regulate and coordinate gene expression. These sequences can be used to identify the locations of genes within the entire human genome sequence. Since the generation of the first draft of the human genome, the interest in understanding gene structure has increased with the goal of identifying disease-associated genes.[12–14]

Gene Structure

Promoting Transcription

Sequences that bind RNA polymerases in combination with transcription factors drive and regulate the production of 1° RNA transcript (Table 1-5). Proteins and transcription factors bind to sequences located 5′, or upstream, of the gene to be expressed and are collectively called the promoter region of a gene. Negative numbering denotes the location of these sequences upstream of the first protein-encoding base. The promoter sequence initiates (or promotes) transcription of the downstream gene and harbors conserved sequences that are recognized by the transcription complex of enzymes.

The complexity and organization of the transcription regulatory sequences of genes differ between prokaryotic and eukaryotic cells. Prokaryotes contain a simple gene structure with sequences for polymerase binding occurring at –35 and –10 for each gene. The –10 sequence contains a consensus sequence of TATAAT, while the –35 region consists of TTGACA. Variations of these sequences as well as the sequences located adjacent to the gene determine the strength of the promoter's transcriptional activity. For example, small differences such as having a TATATA sequence rather than the consensus sequence at the –10 position will decrease the promoter's ability to bind to the RNA polymerase and result in decreased production of mRNA for that gene. In bacteria, operons regulate expression of multiple genes with related functions from the same promoter.

Table 1-5. Examples of Nucleic Acid Motifs

Name	Sequence
AP1 binding site	TGASTCAG
AP2 binding site	CCCCAGGC
AP3 binding site	GGGTGGGAAAG
AP4 binding site	YCAGCTGYGG
C/EBP	TGTGGAAAG
CCAAT box	CCAAT
CP1 binding site	YN(6)RRCCAATCA
CP2 binding site	YAGYN(3)RRCCAATC
CREB	TGACGTCA
CTF/NF1 binding sites	GCCAAT
GCN4 target site	ATGASTCAT
Glucocorticoid receptor	GGTACAN(3)TGTTCT
Homeobox protein-binding site	TCAATTAAAT
HSTF	CNNGAANNTTCNNG
INF-stimulated response	RGGAANNGAAACT
Lariat consensus sequence	YNYTRAY
MALT box	GGAKGGA
NF-1	TTGGMN(5)GCCAAT
Octamer sequence	ATTTGCAT
Poly A signal	AATAAA
Splice acceptor	Y(11)NYAGG
Splice donor	MAGGTRAGT
TATA box	TATA
Translational initiation sequence	RNNMTGG

R = A/G; Y = C/T; M = C/A; W = A/T; N = A/T/C/G.

In eukaryotic genes, various promoter sequences bind multiple proteins, which catalytically modify and activate other bound proteins. Enhancer sequences increase the production of mRNA but are far removed from the gene. One of the pivotal proximally located sequences comprises a TATA box (TATAAA) located at –25 (Figure 1-5). These bases initiate binding of a TATA-binding protein (TBP) within the transcription factor D complex. Following this binding, transcription factors B, H, and E bind to and open the DNA strands downstream from the promoter. Finally, transcription factor F and RNA polymerase II bind to the transcription complex. The close proximity of these proteins to RNA polymerase II permits phosphorylation of the polymerase and initiation of transcription. In eukaryotic cells, variations in the recognition sequences alter the efficiency of transcription. These variations may be base pair changes or base modifications. For example, promoter sequences that are highly methylated do not bind well to the transcription factors or polymerase. As a result, a gene

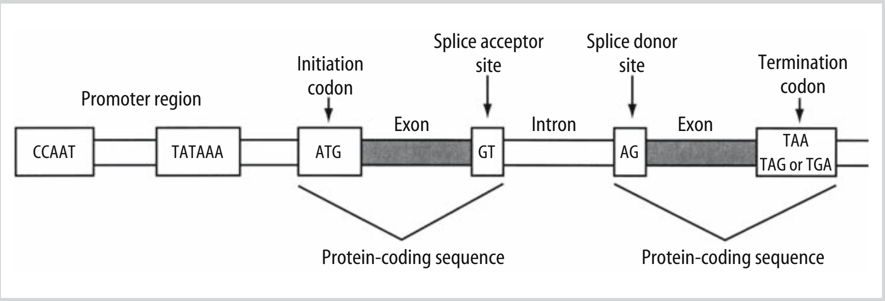

Figure 1-5. Gene structure. Gene structure depicting coding and noncoding regions of the eukaryotic gene. (Reprinted from Leonard D. *Diagnostic Molecular Pathology*, copyright 2003, with permission from Elsevier.)

may appear to be unaltered or intact but may be transcriptionally silent due to methylated bases in the promoter region.

Elongation and Termination of the mRNA

Once the RNA polymerase binds to the promoter, transcription begins at position +1. The polymerase reads the DNA in a 3′ to 5′ direction, while synthesizing the 1° RNA transcript in a 5′ to 3′ direction. In bacteria, the complete transcript serves as the template for translation. Transcription ends with termination. The mRNA must be terminated in bacteria; termination of the transcript can result from attenuation or the formation of hairpin structures. Termination occurs at several sites beyond the polyadenylation signal in eukaryotic cells and is dependent on bases near the stop codon.[15] Because the eukaryotic cell transcripts are polyadenylated, a termination of the transcription process by a process similar to attenuation is not necessary to regulate gene expression. Specifically, transcripts produced after the polyadenylation signal lack a 5′ cap, resulting in rapid degradation.[16,17]

In eukaryotic cells, once the 1° RNA transcript has been produced in the nucleus, this transcript is processed to form an mRNA by splicing to remove the non-protein-coding introns (intervening sequences) and join the protein-coding exons. Introns are located between sequences called exons, which encode the protein sequence and are translated during protein synthesis. Splicing involves a complex of ribonucleoproteins known as a spliceosome, which recognizes consensus sequences at the 5′ and 3′ ends of the intron. Genetic changes to these splice donor (A/C AG G U A/G AGU) and splice acceptor ($[U/C]_{11}$ N C/U AG G/A) consensus sequences may prevent the spliceosome from recognizing and catalyzing the splicing event.[18,19] Autoantibodies directed to or alterations in the steady-state level of the spliceosome may play a role in some diseases.[20–22] Alternate splicing may generate multiple distinct transcripts from a single gene. That is, some exons may be spliced out in one mRNA molecule but retained in another. As a result, alternate splicing generates different proteins from the same gene and 1° RNA transcript.[23–24]

An additional mechanism of generating diversity from 1° RNA transcripts entails trans-splicing (initially identified in *Drosophila* cells). Essentially, two separate, unrelated transcripts form a hybrid molecule by using the splice donor from the first transcripts and the splice acceptor from the second transcripts. Complementary intronic sequences in both transcripts facilitate the generation of the chimeric mRNA. This process has not been demonstrated in other eukaryotic cells. However, when the process is used for gene therapy applications, normal gene function has been restored from defective genes using trans-splicing.[25–26] Other therapeutic applications for catalytic RNA molecules involve innovative treatments for HIV-infected patients. In this application, synthetic ribozymes cleave drug-resistant variants of HIV.[25,27–29]

Additional modifications of the 1° RNA eukaryotic transcript enhance the stability and transport of the mRNA. One such modification occurs immediately on the generation of the 1° transcript and involves addition of a 7-methyl guanosine linked in an unusual 5′ to 5′ linkage to the triphosphate at the 5′ end of the transcript, also known as the 5′ cap. This cap protects the transcript from degradation. Another 1° transcript modification is cleavage at a polyadenylation signal (AAUAA) near the 3′ end of the transcript, followed by the addition of 100 to 200 adenosine residues (poly-A tail) by polyadenylate polymerase. The poly-A tail facilitates transportation of the mature mRNA into the cytoplasm and protection of the transcript from degradation by exonucleases. A given gene may have several polyadenylation signals, providing another level of variation for a single gene.[30–32]

Translation

Translation is the next step in using information from the DNA gene to produce a functional protein. This process changes the genetic information from a nucleic-acid-based language into an amino-acid-based language of polypeptides and proteins. For these reasons, the term "translation" describes this complex cascade of events.

Following transportation of the mRNA into the cytoplasm, translation begins with the mRNA binding to a ribosome and requires additional nucleic acids, specifically protein-associated RNA molecules (Figure 1-6). A ribosome is a complex of about 50 different proteins associated

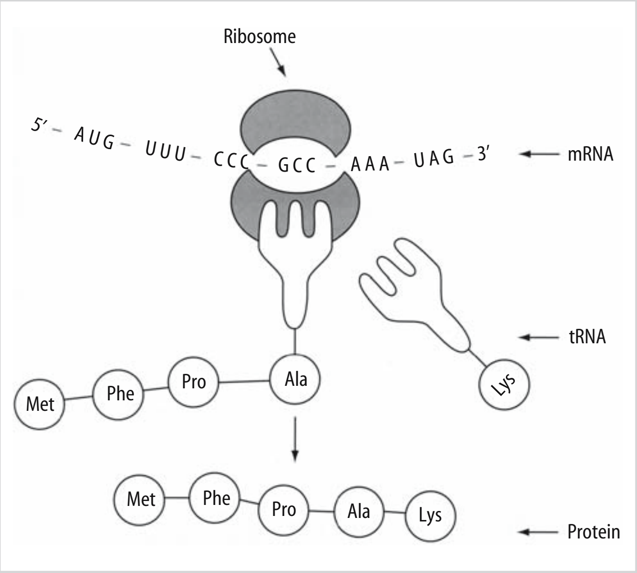

Figure 1-6. RNA translation. RNA is translated through binding events between the mRNA, a ribosome, tRNA, and amino acids, resulting in the production of a protein polypeptide chain. (Reprinted from Leonard D. *Diagnostic Molecular Pathology*, copyright 2003, with permission from Elsevier.)

Table 1-6. The Human Genetic Code

		SECOND BASE OF CODON			
		U	**C**	**A**	**G**
FIRST BASE OF CODON	**U**	**UUU** Phenylalanine (Phe/F)	**UCU** Serine (Ser/S)	**UAU** Tyrosine (Tyr/Y)	**UGU** Cysteine (Cys/C)
		UUA Phenylalanine (Phe/F)	**UCA** Serine (Ser/S)	**UAA** Tyrosine (Tyr/Y)	**UGA** Cysteine (Cys/C)
		UUC Leucine (Leu/L)	**UCC** Serine (Ser/S)	**UAC** **STOP**	**UGC** **STOP**
		UUG Leucine (Leu/L)	**UCG** Serine (Ser/S)	**UAG** **STOP**	**UGG** Tryptophan (Trp/W)
	C	**CUU** Leucine (Leu/L)	**CCU** Proline (Pro/P)	**CAU** Histidine (His/H)	**CGU** Arginine (Arg/R)
		CUA Leucine (Leu/L)	**CCA** Proline (Pro/P)	**CAA** Histidine (His/H)	**CGA** Arginine (Arg/R)
		CUC Leucine (Leu/L)	**CCC** Proline (Pro/P)	**CAC** Glutamine (Gln/Q)	**CGC** Arginine (Arg/R)
		CUG Leucine (Leu/L)	**CCG** Proline (Pro/P)	**CAG** Glutamine (Gln/Q)	**CGG** Arginine (Arg/R)
	A	**AUU** Isoleucine (Ile/I)	**ACU** Threonine (Thr/T)	**AAU** Asparagine (Asn/N)	**AGU** Serine (Ser/S)
		AUA Isoleucine (Ile/I)	**ACA** Threonine (Thr/T)	**AAA** Asparagine (Asn/N)	**AGA** Serine(Ser/S)
		AUC Isoleucine (Ile/I)	**ACC** Threonine (Thr/T)	**AAC** Lysine (Lys/K)	**AGC** Arginine (Arg/R)
		AUG START Methionine (Met/M)	**ACG** Threonine (Thr/T)	**AAG** Lysine (Lys/K)	**AGG** Arginine (Arg/R)
	G	**GUU** Valine (Val/V)	**GCU** Alanine (Ala/A)	**GAU** Aspartic Acid (Asp/D)	**GGU** Glycine (Gly/G)
		GUA Valine (Val/V)	**GCA** Alanine (Ala/A)	**GAA** Aspartic Acid (Asp/D)	**GGA** Glycine (Gly/G)
		GUC Valine (Val/V)	**GCC** Alanine (Ala/A)	**GAC** Glutamic Acid (Glu/E)	**GGC** Glycine (Gly/G)
		GUG Valine (Val/V)	**GCG** Alanine (Ala/A)	**GAG** Glutamic Acid (Glu/E)	**GGG** Glycine (Gly/G)

with several ribosomal RNA (rRNA) molecules. Prokaryotic ribosomes consist of 30S and 50S subunits. Svedberg (S) units are the sedimentation rate of a particle. In eukaryotes, rRNA molecules associate with proteins in the nucleolus to form 40S and 60S subunits. Recognition of the 5′ cap of the eukaryotic mRNA by a ribosome initiates the process of translation.[33]

Each amino acid is encoded by one or more 3-nucleotide sequences, which are collectively known as the genetic code (Table 1-6). Each set of 3 nucleotides of an mRNA that encodes an amino acid is called a codon. As is seen in Table 1-6, the first and second nucleotide positions largely determine which amino acid is encoded by the mRNA codon, while the third base has less effect on which amino acid will be incorporated. In addition to encoding amino acids, certain mRNA codons are used to initiate (START) or terminate (STOP) translation. The genetic code differs slightly between organisms and between mitochon-drial DNA and eukaryotic DNA (Table 1-7). Thus, while one mRNA encodes only one protein sequence, a protein sequence can be encoded by several different mRNA sequences. This is referred to as the degeneracy of the genetic code.

Synthesis of the encoded protein begins at the initiation codon of the mRNA, the first AUG codon after the promoter and encodes a methionine amino acid. This methionine codon establishes the reading frame of the mRNA. The next step in the translation process uses RNA molecules to bridge the information from the sequential mRNA codons to the encoded amino acid in the growing polypeptide chain of the protein. Another set of RNA molecules, tRNA, contain a sequence complementary to each mRNA codon known as the anticodon. The 3′ end of each type of tRNA binds the specific amino acid corresponding to its anticodon sequence. Base pairing of codons with complementary anticodons permits sequential alignment of new amino acids of the polypeptide chain and occurs in the

Table 1-7. Exceptions to the Universal Code in Mammals

Codon	Nuclear Code	Mitochondrial Code
UGA	Stop	Trp
AUA	Ile	Met
AGA	Arg	Stop
AGG	Arg	Stop

small subunit of the ribosome. The large subunit of the ribosome catalyzes the covalent bonds linking each sequential amino acid to the growing polypeptide chain.

Translation ceases when the ribosome encounters a stop codon (UAA, UAG, or UGA). Release factors bound to the stop codon catalyze the addition of a water molecule rather than an amino acid, thus resulting in a COOH terminus to the completed polypeptide chain.[34] Some factors bound to the 3′ untranslated portion of the gene also affect termination.

Structure of Proteins

Just as nucleic acids form various structures via intra- and intermolecular base pairing, proteins also assume various structures depending on the types and locations of amino acids. The primary structure of a protein is the sequence of amino acids from amino terminus (NH) to carboxy terminus (COOH) of the protein. The secondary structure refers to how amino acid groups interact with neighboring amino acids to form structure called an alpha helix or beta sheet. The tertiary structure of a protein is created by amino acids sequentially distant from one another creating intramolecular interactions. The quaternary structure of a protein defines the three-dimensional and functional conformation of the protein. The shape that is ultimately assumed by the protein depends on the arrangement of the different charged, uncharged, polar, and nonpolar amino acids.

Posttranslational Modifications

After generation of the polypeptide chain of amino acids, additional enzymatic changes may diversify its function. These changes are termed posttranslational modifications and can include proteolytic cleavage, glycosylation, phosphorylation, acylation, sulfation, prenylation, and vitamin C– and vitamin K–mediated modifications. In addition, selenium may be added to form selenocysteine. The selencysteinyl-tRNA recognizes the UGA stop codon and adds this unusual amino acid.

Mutations: Genotype Versus Phenotype

Genetic information exists in the form of nucleic acids known as the genotype. In contrast, the encoded proteins function to create a phenotype, an outwardly observable characteristic. Genotypic alterations may or may not cause phenotypic alterations. For instance, missense mutations refer to genetic changes that result in the incorporation of a different amino acid at a specific codon location. These changes may not dramatically alter the protein if the replacement amino acid is similar to the original amino acid (for example, a hydrophobic amino acid replaces another hydrophobic amino acid). However, replacement of an amino acid with a different type of amino acid may significantly change the conformation of the protein and thus change its function. For example, in sickle cell anemia, a valine replaces a glutamic acid at a single position and permits the polymerization of the beta globin molecules to cause stiffening and sickling of the red blood under low oxygen conditions. Different forms of proteins (known as conformers) provide the mechanism for diseases ranging from Creutzfeldt-Jacob disease to Huntington disease. Nonsense mutations describe base changes that replace an amino-acid-encoding codon with a stop codon, which causes premature termination of translation and results in a truncated protein.[35] Truncation may result from the addition or deletion of one or two nucleotide bases, resulting in a shift in the reading frame. Frameshifts often result in premature termination when stop codons are formed downstream from the mutation. Alterations in splice donor or acceptor sites may either erroneously generate or prevent appropriate splicing of the 1° transcript, resulting in a frameshift mutation.[36] Genetic changes in the untranslated portions of the gene affecting the promoter, enhancer, or polyadenylation signals may affect the expression of the gene product and result in a phenotypic change. Not all genotypic changes affect the phenotype. Genetic changes affecting the third base of the codon rarely alter the gene code and would therefore be less likely to cause incorporation of a different amino acid.

With the sequencing of the human genome, numerous single nucleotide polymorphisms have been identified, demonstrating the individual nature of human beings. Numerous studies currently target correlating genotype variations to disease phenotypes. These efforts, in combination with improved understanding of gene structure and function, hold the promise of improved diagnosis, treatment, and patient outcomes in the future.

Acknowledgment

I would like to acknowledge Laurie Sower, PhD, for her helpful advice and editorial assistance and our numerous discussions that greatly facilitated the preparation of this chapter.

References

1. Passarge E. *Color Atlas of Genetics*. 2nd ed. Stuttgart: Thieme; 2001.
2. Willets N. Plasmids. In Scaife DLJ, Galizzi A, eds. *Genetics of Bacteria*. London: Academic Press; 1985:165–195.
3. Hewitt WL. Penicillin-historical impact on infection control. *Ann N Y Acad Sci.* 1967;145:212–215.
4. Livermore DM. Antibiotic resistance in staphylococci. *Int J Antimicrob Agents.* 2000;16(suppl 1):S3–S10.
5. Crick JWJF. A structure for deoxyribonucleic acid. *Nature.* 1953;171:737.
6. Breslauer KJ, Frank R, Blocker H, Marky LA. Predicting DNA duplex stability from the base sequence. *Proc Natl Acad Sci U S A.* 1986;83:3746–3750.

7. Freier SM, Kierzek R, Jaeger JA, et al. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci U S A*. 1986;83:9373–9377.
8. Lewis ME, Arentzen R, Baldino F Jr. Rapid, high-resolution in situ hybridization histochemistry with radioiodinated synthetic oligonucleotides. *J Neurosci Res*. 1986;16:117–124.
9. Felsenfeld G, Groudine M. Controlling the double helix. *Nature*. 2003;421:448–453.
10. Preston BD, Poiesz BJ, Loeb LA. Fidelity of HIV-1 reverse transcriptase. *Science*. 1988;242:1168–1171.
11. Roberts JD, Bebenek K, Kunkel TA. The accuracy of reverse transcriptase from HIV-1. *Science*. 1988;242:1171–1173.
12. Kochetov AV, Ischenko IV, Vorobiev DG, et al. Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett*. 1998;440:351–355.
13. Gotoh O. Homology-based gene structure prediction: simplified matching algorithm using a translated codon (tron) and improved accuracy by allowing for long gaps. *Bioinformatics*. 2000;16:190–202.
14. Olivier M, Aggarwal A, Allen J, et al. A high-resolution radiation hybrid map of the human genome draft sequence. *Science*. 2001;291:1298–1302.
15. McCaughan KK, Brown CM, Dalphin ME, Berry MJ, Tate WP. Translational termination efficiency in mammals is influenced by the base following the stop codon. *Proc Natl Acad Sci U S A*. 1995;92: 5431–5435.
16. Frischmeyer PA, Dietz HC. Nonsense-mediated mRNA decay in health and disease. *Hum Mol Genet*. 1999;8:1893–1900.
17. Wilkinson MF, Shyu AB. Multifunctional regulatory proteins that control gene expression in both the nucleus and the cytoplasm. *Bioessays*. 2001;23:775–787.
18. Bruno C, Lofberg M, Tamburino L, et al. Molecular characterization of McArdle's disease in two large Finnish families. *J Neurol Sci*. 1999;165:121–125.
19. Hou VC, Conboy JG. Regulation of alternative pre-mRNA splicing during erythroid differentiation. *Curr Opin Hematol*. 2001;8:74–79.
20. Lipes J, Skamene E, Newkirk MM. The genotype of mice influences the autoimmune response to spliceosome proteins induced by cytomegalovirus gB immunization. *Clin Exp Immunol*. 2002;129: 19–26.
21. Seidl R, Labudova O, Krapfenbauer K, et al. Deficient brain snRNP70K in patients with Down syndrome. *Electrophoresis*. 2001; 22:43–48.
22. Wehner KA, Ayala L, Kim Y, et al. Survival motor neuron protein in the nucleolus of mammalian neurons. *Brain Res*. 2002;945:160–173.
23. Garzon D, Yu G, Fahnestock M. A new brain-derived neurotrophic factor transcript and decrease in brain-derived neurotrophic factor transcripts 1, 2 and 3 in Alzheimer's disease parietal cortex. *J Neurochem*. 2002;82:1058–1064.
24. Sakata N, Yamazaki K, Kogure T, Mukai T. Alternative splicing of Rh blood group polypeptide mRNA produces a novel transcript containing a short nucleotide insertion on human erythroleukemia K562 cells. *Cell Biol Int*. 2001;25:697–703.
25. Liu X, et al. Partial correction of endogenous DeltaF508 CFTR in human cystic fibrosis airway epithelia by spliceosome-mediated RNA trans-splicing. *Nat Biotechnol*. 2002;20:47–52.
26. Phylactou LA, Darrah C, Wood MJ. Ribozyme-mediated trans-splicing of a trinucleotide repeat. *Nat Genet*. 1998;18:378–381.
27. Phylactou LA, Kilpatrick MW, Wood MJ. Ribozymes as therapeutic tools for genetic disease. *Hum Mol Genet*. 1998;7:1649–1653.
28. Lan N, Howrey RP, Lee SW, Smith CA, Sullenger BA. Ribozyme-mediated repair of sickle beta-globin mRNAs in erythrocyte precursors. *Science*. 1998;280:1593–1596.
29. Mansfield SG, Kole J, Puttaraju M, et al. Repair of CFTR mRNA by spliceosome-mediated RNA trans-splicing. *Gene Ther*. 2000;7:1885–1895.
30. Urano Y, Watanabe K, Sakai M, Tamaoki T. The human albumin gene. Characterization of the 5′ and 3′ flanking regions and the polymorphic gene transcripts. *J Biol Chem*. 1986;261:3244–3251.
31. Lin B, Rommens JM, Graham RK, et al. Differential 3′ polyadenylation of the Huntington disease gene results in two mRNA species with variable tissue expression. *Hum Mol Genet*. 1993;2:1541–1545.
32. Boyd CD, Mariani TJ, Kim Y, Csiszar K. The size heterogeneity of human lysyl oxidase mRNA is due to alternate polyadenylation site and not alternate exon usage. *Mol Biol Rep*. 1995;21:95–103.
33. Gallie DR. Protein-protein interactions required during translation. *Plant Mol Biol*. 2002;50:949–970.
34. Chavatte L, Frolova L, Kisselev L, Favre A. The polypeptide chain release factor eRF1 specifically contacts the s(4)UGA stop codon located in the A site of eukaryotic ribosomes. *Eur J Biochem*. 2001;268:2896–2904.
35. Stratakis CA. Mutations of the gene encoding the protein kinase A type I-alpha regulatory subunit (PRKAR1A) in patients with the "complex of spotty skin pigmentation, myxomas, endocrine overactivity, and schwannomas" (Carney complex). *Ann N Y Acad Sci*. 2002;968:3–21.
36. Valentine CR. The association of nonsense codons with exon skipping. *Mutat Res*. 1998;411:87–117.
37. Cline J, Braman JC, Hogrefe HH. PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res*. 1996; 24:3546–3551.
38. Maga G, Shevelev I, Ramadan K, Spadari S, Hubscher U. DNA polymerase theta purified from human cells is a high-fidelity enzyme. *J Mol Biol*. 2002;319:359–369.
39. Kuchta RD, Cowart M, Allen D, Benkovic SJ. Kinetic and structural investigations of the replicative fidelity of the Klenow fragment. *Biochem Soc Trans*. 1988;16:947–949.