Fifth Edition

First Principles of Gastroenterology

The Basis of Disease and an Approach to Management

JANSSEN-ORTHO

A.B.R. Thomson and E.A. Shaffer, editors

15

The Applications of Recombinant DNA Technology in Gastrointestinal Medicine and Hepatology: The Basic Paradigms of Molecular Cell Biology

G.E. Wild, P. Papalia, M.J. Ropeleski, J. Faria and A.B.R. Thomson

1. INTRODUCTION[1-8]

For most gastroenterologists, the principles of cell and molecular biology have not played a major role in day-to-day clinical practice.[1] However, tremendous advances in the discipline of molecular medicine have provided new insights into the cellular and molecular pathologic basis of disease. This ever-increasing expansion of our knowledge base has transformed our understanding and management of an array of diseases. The cumulative research efforts in cell and molecular biology have translated into clinically relevant information in every medical subspecialty. For example, hematologists have defined the molecular basis of the hemoglobinopathies. Endocrinologists have defined the cellular and molecular networks that mediate the action of hormones. Neurologists have identified a host of gene mutations that lead to neurodegenerative disorders. Finally, the identification of the cystic fibrosis transmembrane regulator has facilitated the molecular diagnosis of the disease and as a result, gene therapy protocols are being conducted at several centers.

Many of the recent advances in molecular medicine have been driven by the Human Genome Project. It is apparent that molecular biology has accounted for a dramatic paradigm shift in both the teaching and the practice of medicine. This chapter constitutes a framework for integrating new information into the core knowledge base of concepts related to the pathogenesis of gastrointestinal disorders and liver disease. We hope to provide the reader with a set of tools for understanding some basic concepts of recombinant DNA technology and its role in unraveling the intricate molecular pathophysiology of

[1] A list of selected terms and their abbreviations is found at the end of this chapter.

disease. As well, we wish to give the reader a sense of the impact of molecular medicine in the areas of gastroenterology and hepatology. The goal of this chapter is to review the basic principles of eukaryotic gene expression.

In contrast to prokaryotes (where all genes are transcribed by a single RNA polymerase that binds directly to gene promoter sequences), transcription in eukaryotic cells involves several different RNA polymerases that interact with a variety of transcription factors to initiate transcription. This increased complexity characteristic of eukaryotic transcription facilitates the sophisticated and orderly regulation of gene expression that ultimately determines the activities of the diverse cell types seen in multicellular organisms.

Three distinct nuclear RNA polymerases are found in eukaryotic cells. Genes that encode proteins are transcribed into messenger RNA (mRNA) by RNA polymerase II. Ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are transcribed by RNA polymerase I and III, respectively. Some small nuclear and cytoplasmic RNAs are transcribed by RNA polymerase II and III. Finally, mitochondrial genes are transcribed by a separate group of RNA polymerases. RNA polymerases are composed of 8 to 14 different subunits. Although they recognize distinct promoters and transcribe different classes of genes, these RNA polymerases share many common features, including a clear dependence on other proteins to initiate transcription.

The transcription of DNA into RNA is the primary level at which gene expression is controlled in eukaryotic cells. Only a fraction of the transcribed RNA is translated into polypeptides. This is explained on the basis of the following:

1. Some transcription units code for RNA molecules only, as in the case of rRNAs, tRNAs and a host of small nuclear and cytoplasmic RNA molecules.
2. The initial transcription product of those transcription units that do not encode polypeptides is subject to events known as RNA processing. With RNA processing, much of the initial RNA sequence is trimmed to yield smaller mRNA molecules.
3. Only the central region of mRNA is translated; variable portions of the 5′ and 3′ ends of mRNA remain untranslated.

Transcription is mediated by the enzyme RNA polymerase, using DNA as a template and ATP, CTP, GTP and UTP as RNA nucleoside precursors. RNA is synthesized in the 5′-to-3′ direction as a single-strand molecule. Only one of the two DNA strands serves as a template for transcription. Since the growing RNA molecule is complementary to this template strand, the transcript has the same 5′-to-3′ orientation and base sequence (except that U replaces T) as the opposite, nontemplate strand of the DNA double helix. Thus, the nontemplate strand is called the *sense strand*, and the template strand is called the

antisense strand. Gene sequences listed in various databases show only the sequence of the sense strand. Orientation of sequences relative to a gene sequence is dictated by the sense strand and by the direction of transcription (e.g., the 5′ end of the gene refers to the sequences of the 5′ end of the sense strand, and "upstream" or "downstream" of the gene refers to sequences that clamp the gene at the 5′ or 3′ ends of the sense strands, respectively).

The processes of DNA replication and transcription occur inside the nucleus. By contrast, protein synthesis takes place in the cytoplasm. Protein synthesis is termed *translation* and is directed by mRNA templates. The translation of mRNA is only the first step in the formation of a functional protein. Importantly, the polypeptide chain must subsequently fold into the appropriate three-dimensional configuration and undergo various processing steps prior to being converted into its active form. In eukaryotic cells, these processing steps are closely related to the sorting and transport of different proteins to their appropriate destinations within the cell.

While the regulation of gene expression occurs primarily at the level of transcription, the expression of many genes can also be controlled at the level of translation. Most proteins can be regulated in response to extracellular signals and, in addition, intracellular protein levels can be controlled by differential rates of protein degradation. Thus, the regulation of both the amounts and activities of intracellular proteins ultimately determines all aspects of cell behavior.

Proteins are synthesized on mRNA templates by a process that is remarkably similar in both prokaryotes and eukaryotes. The mRNAs are translated in the 5′-to-3′ direction, and polypeptide chains are synthesized from the amino to the carboxy terminus. The amino acids incorporated into the polypeptide chains are specified by three bases (A, U, and C or G – i.e., a *codon*) in the mRNA, determined by the genetic code. Translation occurs on ribosomes with tRNAs serving as adapters between the amino acids being incorporated into the nascent protein strand and the mRNA template. Thus, protein synthesis involves interactions between three species of RNA molecules: mRNA templates, rRNAs, tRNAs.

At the 5′ end of the mRNA is the Cap sequence, followed by the 5′ untranslated region (UTR), and then by the AUG codon that signals the initiation of translation. Toward the 3′ end of the mRNA there is a signal for the termination of translation (UAA, UAG or UGA) followed by the 3′-UTR. At the extreme 3′ end of the mRNA is the poly A tail. Protein synthesis starts at the AUG codon and proceeds in the 5′-to-3′ direction until a termination codon is reached, which heralds the end of protein synthesis.

The genetic code comprises 64 codons, each containing three bases (A, U, and either C or G). This accounts for the permutations of the four bases in groups of three. The 64 codons code for 61 amino acids and termination

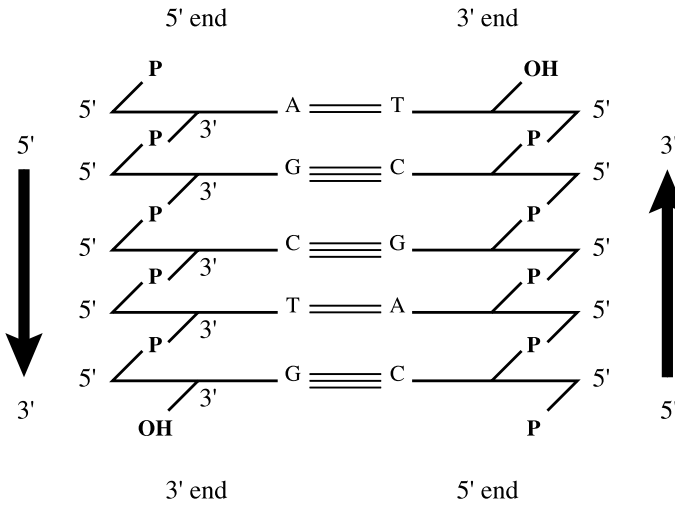

FIGURE 1. Base pairing and the antiparallel orientation of DNA. The two DNA strands in the helix have opposite polarity, with one strand running in a 5′-to-3′ direction and the other running in the 3′-to-5′ direction. Four bases (A, T, C and G) reside on the inside of the helix to allow hydrogen bonding between purine and pyrimidine residues.

signals. The genetic code, with some minor exceptions, is ubiquitous: the same codons always code for the same amino acid. Minor variations do occur in the mitochondria. More than one codon can code for the same amino acid in a species-specific fashion. This is known as "redundancy of the genetic code."

2. Eukaryotic Gene Structure and DNA Replication

2.1 Nucleic Acids and Information Transfer in Cells[1-8]

DNA (deoxyribonucleic acid) is the storage form of genetic information in cells. The structure of DNA was determined by Watson and Crick in 1953, a discovery that has revolutionized the thinking in modern cell biology. All DNA molecules consist of four types of nucleotides joined together by phosphodiester bonds to form polynucleotides. The nitrogenous bases found in DNA consist of purines (i.e., adenine [A] and guanine [G]) and pyrimidines (i.e., cytosine [C] and thymine [T]) (Figure 1). The nucleotides are linked together by covalent phosphodiester bonds that join the 5′ carbon of one deoxyribose to the 3′ carbon of the adjacent deoxyribose to form

polynucleotide genes. The double-stranded DNA helix with its two polynucleotide strands of DNA run in an antiparallel orientation, and the DNA strands are held together by hydrogen bonding between A and T residues and G and C residues. The antiparallel orientation in base pairing is an important concept in nucleic acid biochemistry. One strand runs in a 5′-to-3′ direction, and the complementary strand runs in the 3′-to-5′ direction (Figure 1). Thus, the two strands of the double helix are complementary. For example, the sequence CTGAAGCGCTTA on one strand of DNA will have the complementary sequence GACTTCGCGAAT on the opposite strand of DNA in an antiparallel orientation. The variation of the sequence of nucleotides along the DNA strand determines the function of each section of the DNA molecule, as well as its ability to transmit information to RNA and protein.

RNA (ribonucleic acid) molecules consist of nucleotides linked together by phosphodiester bonds. RNA generally occurs as single-stranded polynucleotides and contains ribose in place of the deoxyribose found in DNA. RNA is made up of the bases A, G and C, but contains uracil (U) in place of T as the fourth base. Since U has the ability to bind with A in the same way that T binds with A, the four bases found in RNA—A, U, G and C—can form complementary pairs with other RNA bases as well as with the bases found in DNA. These biochemical properties highlight the major function of the RNA molecule in the transfer of information from DNA to protein in eukaryotic cells. RNA often contains intramolecular hydrogen bonding that gives rise to secondary structures. Intrastrand base pairing creates structures known as *stem loop structures*, with the base pairing sections forming the stem and noncomplementary bases forming the loop.

Eukaryotic cells contain five classes of RNA: (1) messenger RNA (mRNA), (2) transfer RNA (tRNA), (3) ribosomal RNA (rRNA), (4) heterogeneous nuclear RNA (hnRNA) and (5) small nuclear RNA (snRNA). mRNA makes up a small percentage of the total RNA (1–5%) in eukaryotic cells, has a short half-life and demonstrates a large variation in base sequence from one mRNA molecule to another. mRNA is the chemical messenger that carries information from the DNA helix to the protein-synthesizing machinery in the cytoplasm.

tRNA molecules are polynucleotides ranging from 75 to 95 nucleotides in length that carry specific amino acids to the ribosomes during protein synthesis. There is a unique tRNA that specifically recognizes each of the 20 amino acids. In some instances, there is more than one tRNA species for a single amino acid.

rRNA is the most abundant of the RNA species in eukaryotic cells, and it is found associated with proteins in structures called *ribosomes.* These specific rRNAs of eukaryotic cells are designated by their sedimentation

coefficients (S values). Human ribosomes contain 28S, 18S, 5.8S and 5S rRNA species.

Heterogeneous nuclear RNA and small nuclear RNA species are located in the nucleus of eukaryotic cells. hnRNA is the immediate product of transcription, and is complementary to one strand of the DNA helix. hnRNA is the precursor to mRNA before it undergoes further processing. snRNA is associated with specific proteins that are involved in the processing of the hnRNA to mRNA prior to departure of the mRNA from the nucleus to the cytoplasm. The role of these RNA molecules in transcription and translation will be discussed in detail in subsequent sections.

2.2 Molecular Anatomy of Eukaryotic Genes [4, 8]

Eukaryotic genomes are larger and more complex than those of primitive prokaryotes (bacteria). For example, the human genome contains approximately 100,000 genes, and much of its complexity arises from the abundance of several different types of noncoding DNA sequences.

A gene can be defined as a segment of DNA that is expressed to yield a functional product that may be either an RNA or a peptide. The structural features common to all eukaryotic genes are illustrated in Figure 2. The sequence of base pairs confers gene specificity and determines the specificity of the product that it encodes. However, not all of the nucleotides present in the gene are expressed in the final product. Eukaryotic genes are often split into (1) *exons* – sequences that remain in the final mature mRNA, and (2) *introns* – sequences that are removed from the primary mRNA transcript early during processing, most which have no known function. In addition to encoding sequence information that ultimately defines the protein product, exons contain other sequences that are essential to the organized functioning of mRNA. Thus, an exon is defined as a sequence in the primary RNA transcript that is conserved during the processing of the transcript into a mature mRNA molecule.

Unique sequences that signal the start of transcription are present in each gene. These sequences represent promoter sequences and they determine the site at which the initiation of transcription begins on the DNA molecule. Transcription is initiated when RNA polymerase and transcription factors bind to the promoter site and catalyze the synthesis of RNA. RNA polymerase transcribes RNA using the sequence of bases from one strand of the DNA double helix, which serves as a template. RNA is synthesized as a single-stranded molecule in the 5′-to-3′ direction.

Further processing of mRNA transcripts to yield a mature RNA product involves a series of steps that include the addition of a Cap structure at the 5′ end of the mRNA and the addition of a poly A tail at the 3′ end. Untranslated

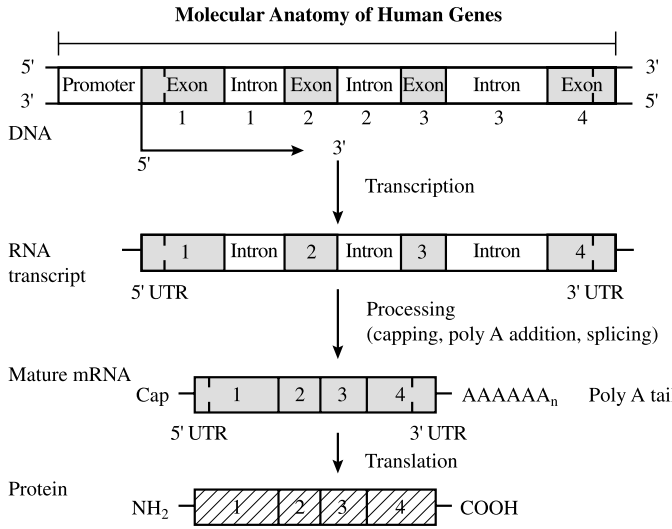

FIGURE 2. A typical human gene contains exon and intron sequences that are transcribed by RNA ploymerase into the primary transcript. This primary transcript is subsequently processed by the addition of a Cap structure at the 5′ end and the addition of a poly A tail at the 3′ end. The intron sequences are removed and the exonic RNA sequences are spliced together. The mature mRNA contains only exonic RNA sequences that have information for protein sequences as well as signals for the initiation and termination of protein synthesis. UTR = untranslated regions.

regions called UTRs are situated at both the 3′ and 5′ ends of the mRNA and represent sequences in the exons that remain in the mRNA but are not translated into proteins. These regions contain signals required for processing of mRNA and its subsequent translation into protein.

2.3 Organization of Eukaryotic Genomes[4, 8]

The average polypeptide is approximately 400 amino acids long. Thus, the average size of the coding sequence of a gene is 1,200 base pairs. Each amino acid is determined by a set of three nucleotides (a codon). In contrast to E. coli and yeasts, the human genome contains large amounts of noncoding DNA. Thus, only a small proportion of the total 3×10^9 base pairs of the human genome is expected to correspond to protein-coding sequences. The average gene spans 10,000 to 20,000 base pairs (including introns) such that the human genome consists of approximately 100,000 genes, representing 3% of the total human DNA.

Several types of highly repeated sequences exist in eukaryotic genomes. One class, called *simple-sequence DNA*, contains tandem arrays of thousands of copies of short sequences ranging from 5 to 200 nucleotides. Such repeat-sequence DNA accounts for approximately 10–20% of the DNA in higher eukaryotes and is referred to as *satellite DNA*. Other repetitive DNA sequences are scattered throughout the genome rather than being clustered as tandem repeats. These sequences are classified as either *short (SINEs)* or *long (LINEs) interspersed elements*. The major SINEs in the mammalian genome are *Alu sequences,* which contain a signal site for the restriction endonuclease *AluI.* These Alu sequences (300 base pairs long) are dispersed throughout the genome and account for nearly 10% of the total cellular DNA. The major human LINEs are about 6,000 base pairs in length and repeat approximately 50,000 times in the human genome. In contrast to Alu sequences, LINE sequences are transcribed, and some encode proteins of unknown function.

Eukaryotic DNA is tightly associated with small, basic proteins (i.e., rich in arginine and lysine) called *histones*. The complexes between eukaryotic DNA and proteins consist of *chromatin*, which contains about twice as much protein as DNA. The basic amino acids contained in histones have been identified: H1, H2A, H2B, H3 and H4. In addition, chromatin contains a variety of nonhistone chromosomal proteins that are involved in DNA replication and gene expression. The association of DNA and protein to form chromatin is illustrated in Figure 3.

The basic structural unit of chromatin is called the *nucleosome*, which is composed of repeating 200 base pair units. Nucleosomes contain a core particle that contains 146 base pairs of DNA wrapped 1.75 times around a histone core consisting of two molecules each of H2A, H2B, H3 and H4. The other structural feature of the nucleosome is the *chromatosome,* which contains two full turns of DNA (166 base pairs) held in place by one molecule of H1. The structure (i.e., degree of condensation) of chromatin is closely linked to the control of gene expression in eukaryotes. The extent of chromatin condensation varies during the life cycle of the cell. In nondividing cells, most of the chromatin, called *euchromatin*, is decondensed and distributed throughout the nucleus. Genes are transcribed during this period of the cell cycle and the DNA is replicated in preparation for mitosis. By contrast, about 10% of interphase chromatin is in a very highly condensed state called *heterochromatin*. Heterochromatin is transcriptionally inactive and contains highly repeated DNA sequences.

The human genome is distributed among 24 chromosomes (22 autosomes and the two sex chromosomes), each containing between 5×10^4 and 26×10^4 kilobases of DNA. The chromosomes have three well-defined structures that

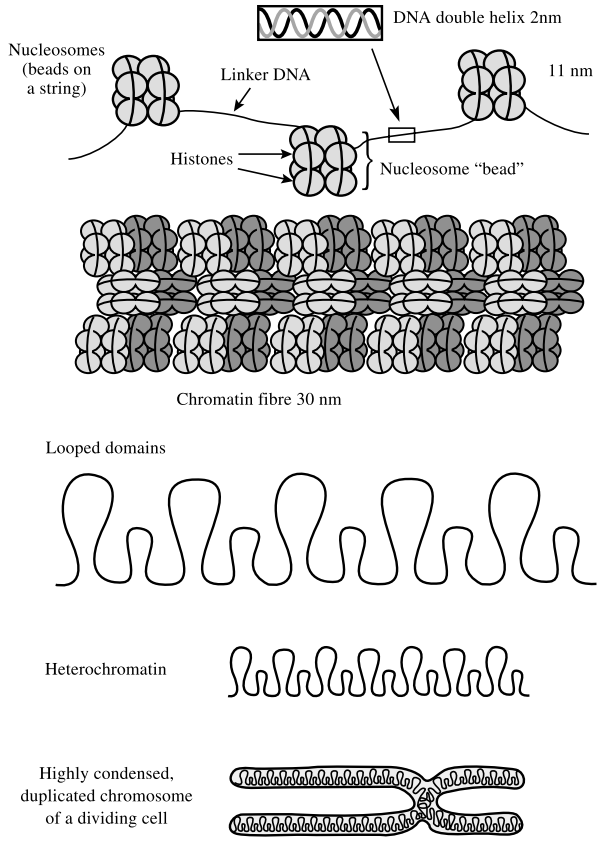

FIGURE 3. The packaging of DNA in the nucleus. A model is depicted for the progressive stages of DNA coiling and folding in the nucleus. The hierarchy of structural features arising from the DNA double helix includes nucleosomes, chromatin fibers and their looped domains, and heterochromatin, which makes up the arms of chromosomes.

are essential for their replication: (1) DNA replication origins, (2) centromeres and (3) telomeres. The DNA replication origins will be considered in detail in Section 2.6. *Centromeres* consist of highly repetitive DNA sequences and are the site where the two sister chromatids (daughter strands of a duplicated chromosome) are attached. The function of the centromere is to ensure the equal distribution of each chromosome to the daughter cells at cell division.

The *telomere* is an important structure associated with the ends of all human chromosomes. Telomeric DNA consists of multiple tandem repeats of the sequence TTAGGG, located at both ends of each chromatid. Telomeres perform a variety of functions in human cells, including the following:

1. Telomeres maintain chromosomal stability and prevent the formation of end-to-end fusions. The presence of telomeric sequences protects chromosomal ends from nuclease degradation.
2. They ensure the proper replication of the ends of chromosomes. DNA ends are not completely replicated during DNA replication and require the presence of the enzyme *telomerase* to add nucleotides to the extreme ends of the DNA molecule. The presence of noncoding telomeric sequences at the chromosomal ends protects the coding sequences of the DNA located near the terminal ends of a chromosome from being lost during each cycle of replication.
3. They serve as markers of chromosomal integrity. In the event that a chromosome is damaged, the cell cycle stops temporarily so that DNA repair mechanisms can repair the damage.

2.4 The Flow of Genetic Information in Eukaryotic Cells[1-8]

The expression of genetic information in all eukaryotic cells is largely a one-way system. DNA directs the synthesis of RNA, and RNA specifies the synthesis of polypeptides that subsequently form proteins. Because of its universality, the *DNA → RNA → protein* flow of genetic information is called the "central dogma of molecular biology." The synthesis of RNA by RNA polymerase using DNA as a template is called *transcription.* Transcription occurs in the nucleus of eukaryotic cells, and to a limited extent in mitochondria. The second step involves polypeptide synthesis and is called *translation.* Translation occurs on ribosomes, which are large RNA protein complexes found in the cytoplasm. The RNA molecules that specify polypeptides are known as messenger RNAs (mRNAs).

Gene expression has been held to follow a *colinearity principle* where the linear sequence of the nucleotides in DNA is decoded to give a linear sequence of nucleotides in RNA. This linear sequence can be decoded in turn to give rise to a linear sequence of amino acids in the polypeptide product. This concept has been challenged by recent findings that eukaryotic cells, including mammalian cells, contain nonviral chromosomal DNA sequences that encode cellular *reverse transcriptases*. Many different classes of viruses have a genome that consists of RNA. Retroviruses such as HIV are a subclass of RNA viruses in which the RNA replicates via a DNA intermediate, using reverse transcriptase, which is an RNA-dependent DNA polymerase. Because

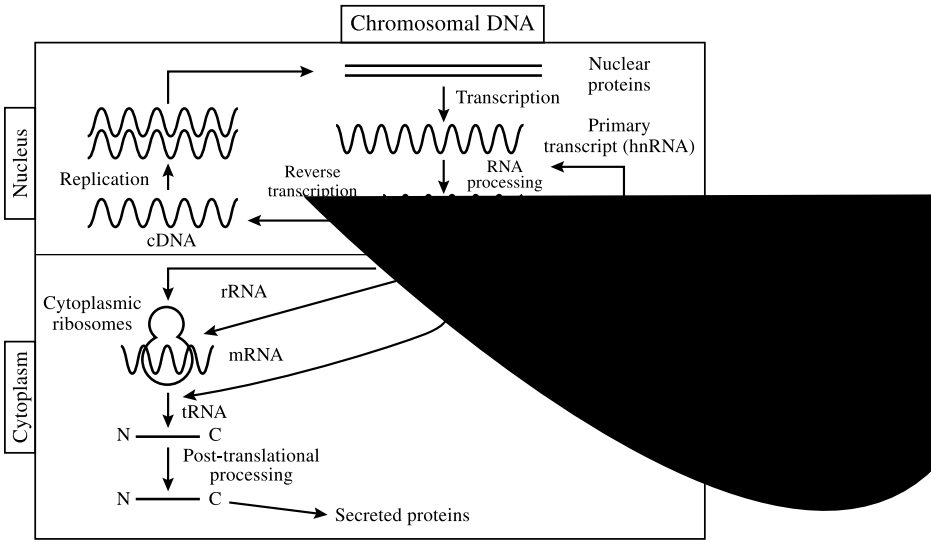

FIGURE 4. Gene expression in the eukaryotic cell. The expression of genetic information in eukaryotic cells is very largely a one-way system. DNA specifies the synthesis of RNA, and RNA specifies the synthesis of polypeptides, which subsequently form proteins. A small proportion of nuclear RNA molecules can be converted to cDNA by reverse transcriptases and subsequently integrate into chromosomal DNA.

some nonviral RNA sequences in eukaryotic cells are known to act as templates for cellular DNA synthesis, the principle of unidirectional flow of genetic information is no longer strictly valid. The overall flow of genetic information and gene expression in eukaryotic cells is illustrated in Figure 4.

2.5 The Cell Cycle[9-13]

The cellular processes that determine DNA replication and mitosis are the keys to normal cell growth and development. These processes occur in a well-regulated and orderly progression through the mammalian cell cycle (Figure 5). The regulation of the cell cycle ultimately determines how a cell progresses between growth, differentiation and division phases. Cell cycle control is a key determinant of either cell differentiation or the decision to halt the cycle. The loss of control of the cell cycle leads to abnormal cell growth, which results in tumorigenesis, developmental defects or premature programmed cell death (i.e., apoptosis).

The mammalian cell cycle comprises four distinct phases: G1 (G = gap), S (synthesis), G2 and M (mitosis). The period between one M phase and the

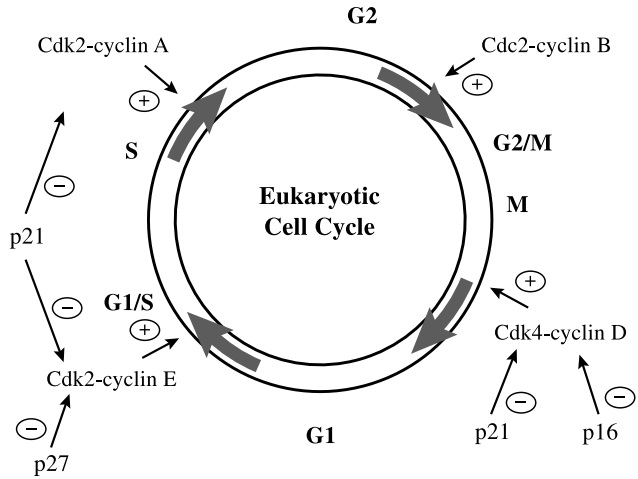

FIGURE 5. Eukaryotic cell cycle. Cyclin-dependent kinases (Cdks), cyclins and Cdk inhibitors (CKIs) interact during the cell cycle. Progression during the cell cycle is regulated by interaction of positive and negative regulatory factors. The positive progression is directed by multiple cyclin-Cdk complexes, which act by phosphorylating various proteins at the different stages of the cycle. Negative regulatory factors include CKIs such as p16, p21 and p27, which inhibit phosphorylation of proteins by kinase and stop the cell cycle.

next – consisting of the three remaining phases G1, S and G2 – is called *interphase*. The G1 phase is the interval between the completion of M phase and the onset of S phase. The G2 phase is the interval between the end of S phase and the beginning of M phase. DNA is replicated during the S phase and is distributed equally to two daughter cells during the M phase. The cells prepare for the S phase during G1, and for the M phase during G2, the interval when proteins are synthesized in preparation for mitosis. Cells that do not undergo division, such as neurons, exit the cell cycle and enter a phase called G0. If cells in G0 are stimulated to grow, they move from G0 into the G1 phase. Progression through the cell cycle is mediated by multiple cyclin-dependent kinases (Cdks) that are sequentially activated by the binding of cyclins. The activated Cdk-cyclin protein complex phosphorylates specific proteins that are required for the reactions unique to each distinct phase of the cell cycle. Cyclins vary dramatically during the cell cycle. For example, cyclin B levels increase during interphase and subsequently decline during M phase. The changes in the level of cyclin B are correlated with the activity of a specific Cdk called Cdc2, which is active when cyclin B levels peak and becomes

TABLE 1. Cyclin-dependent kinases (Cdks), cyclins and cyclin-dependent kinase inhibitors (CKIs) at different stages of the cell cycle

Cell cycle phase	*Cdk*	*Cyclin*	*CKI*	
			KIP[a]	*INK*[b]
G1	Cdk4	Cyclin D	p21, p27	P15, P16
G1/S	Cdk2	Cyclin E	p21, p27	
S	Cdk2	Cyclin A	p21	
G2/M	Cdc2	Cyclin B	p21	
M	Cdc2	Cyclin B, cyclin A		

[a]KIP proteins (p21, p27) bind multiple cyclin-Cdk complexes that prevent activation or inhibit kinase activity.
[b]INK proteins (p15, p16) are specific for Cdk4/6 and cyclin D. They bind Cdk and inhibit the binding of cyclin D.

inactive as cyclin B levels decline. Thus, the phosphorylating activity of Cdc2 is modulated during the cell cycle by the availability of cyclin B. The activation of Cdc2 also depends on phosphorylation of a specific threonine residue, thus adding a second layer to the control of the kinase activity. A variety of cell-cycle "checkpoints" monitor progression through the cell cycle. Deviation from the normal cell cycle impedes progression beyond the checkpoint and the cell cycle is halted until the defect is corrected. Thus, the orderly progression through the cell cycle depends on both positive factors that drive the cell cycle forward and negative factors that halt the cycle at a particular stage. Cdks and specific cyclins are the main positive factors that function at each stage of the cell cycle. Negative factors that block the activity of the specific Cdks are called cyclin-dependent kinase inhibitors (CKIs) (Table 1).

Several mechanisms may be responsible for the inactivation of an active Cdk-cyclin complex:

1. The cyclin molecule can be degraded through the ubiquitin protein-degrading system.
2. The critical phosphate required for activation of the kinase activity can be removed from the protein by a specific phosphatase.
3. CKI molecules interact with Cdks or Cdk-cyclin complexes and inhibit the kinase activity. Two classes of CKIs have been described, the INK (inhibitor of Cdk) class and the KIP (kinase inhibitory protein) class (Table 1).

Thus the interplay between the activation and deactivation of the Cdk activities at various stages of the cell cycle is the key determinant of the normal progression and regulation of the cell cycle.

2.5.1 THE G1 PHASE

The G1 phase heralds the onset of the cell cycle. Resting cells (G0 phase) that are stimulated to divide enter the G1 phase. Once the cell passes this point it is committed to entering the S phase and subsequently divides. The key positive regulators of the G1 phase are Cdk4 and cyclins of the D family, which form a complex capable of phosphorylating a host of proteins required for cell function in the G1 phase. The retinoblastoma protein (pRb) is a key protein phosphorylated by the Cdk4–cyclin D in G1. pRb exists in a nonphosphorylated form during the first two-thirds of the G1 phase and becomes phosphorylated just prior to the transition from G1 to S phase. Nonphosphorylated pRb restricts cell growth, whereas phosphorylated pRb is associated with a loss of growth inhibitory function and allows the cell to proceed through the cell cycle. Thus, pRb functions as a regulator that represses or activates specific promoters through interaction with and modification of the activities of transcription factors that bind to DNA and regulate the expression of cell-cycle genes. The phosphorylation of pRb by the Cdk4–cyclin D complex allows previously repressed genes to be transcribed and allows the cell to progress from G1 to S phase.

The Cdk inhibitor p27 is a second important control that regulates the progression of a cell from G1 to S phase. This protein binds to the Cdk2–cyclin E complex and inactivates it. The cells are unable to proceed to the S phase and remain arrested in G1. Growth-promoting factors result in the degradation of p27, activation of the Cdk2–cyclin E complex and transition to the S phase. The ubiquitin protein-degrading system is responsible for the degradation of p27.

2.5.2 THE S PHASE

Entry into the S phase is determined by a putative cytoplasmic signal that is most likely an active Cdk-cyclin complex. The entrance into S phase from G1 and progression through S phase to G2 depends on the function of specific Cdk-cyclin complexes. Cdk2 initially binds cyclin E as the cells proceed into the S phase. Cyclin A activates Cdk2 and phosphorylation of proteins required for DNA replication.

2.5.3 THE G2/M PHASE

The G2/M phase represents a critical checkpoint where cells decide whether to enter mitosis. The critical proteins involved in the G2/M checkpoint include Cdc2 and cyclin B, which form a complex. The Cdc2–cyclin B complex is essential for the entry into and exit from the M phase, which involves the activation and deactivation of the Cdc2-cyclin B complex through a series of phosphorylation and dephosphorylation steps.

2.5.4 THE M PHASE

The sudden activation of the Cdc2–cyclin B complex by dephosphorylation, which occurs at the G2/M border, results in the phosphorylation of a variety of proteins required for mitosis. Three checkpoints are key to the orderly entry into and exit from mitosis, with each daughter cell receiving an exact copy of the parental genome. These three checkpoints are (1) the transition from G2 to M concurrent with the activation of the Cdc2–cyclin B complex; (2) the M phase checkpoint that occurs during metaphase (the point that regulates the timing of the separation of the chromatids and the initiation of anaphase); and (3) the immediate proteolytic destruction of cyclin B at the onset of anaphase with the concomitant inactivation of Cdc2 (which allows the cell to exit the M phase and enter a new G1 phase). These checkpoints are regulated by the ubiquitin pathway.

2.5.5 THE ROLE OF P53 AND P21 IN THE CONTROL OF CELL DAMAGE

The orderly progression within the cell cycle and the cell's ability to sense any perturbation in its normal state are crucial to normal cell growth and development. Cells have evolved negative regulatory mechanisms that sense physiological disturbances, DNA damage, hypoxia, nutrient depletion and viral infection. Either the cell can arrest the cycle at a particular stage or, in some instances, the cell will undergo programmed cell death, which is called apoptosis.

The DNA-binding protein, p53, orchestrates the negative regulatory mechanisms that take effect when the cell is damaged. The p53 protein is a tumor-supressor protein and activates transcription of the gene encoding the Cdk inhibitor, p21. The p21 protein binds to multiple cyclin-Cdk complexes and blocks kinase activity. This inhibits the phosphorylation of proteins required for the various stages of the cell cycle. The binding of p21 to the G1 cyclin-Cdk complexes is central to the cessation of the G1 phase that follows DNA damage by radiation. This gives the DNA repair mechanisms time to correct the damage. Another function of p21 is to bind *proliferating cell nuclear antigen (PCNA)*. PCNA is a cofactor required for full expression of DNA polymerase δ (see Section 2.6). DNA replication is inhibited when p21 is bound to PCNA. The roles that p53 and p21 play in damage control in cells are illustrated in Figure 6.

Mutations that lead to the loss or alteration of p53 activity result in cancer development. Abnormal p53 levels are associated with the loss of the cell's ability to halt the progression of the cell cycle under the aforementioned adverse conditions. Therefore, the cell continues to proliferate, and this results in a defective phenotype.

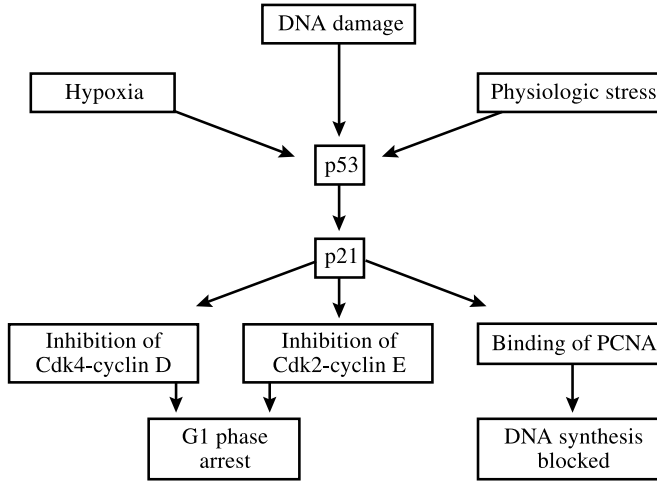

FIGURE 6. The control of damage by p53 and p21. Cellular damage results in increased p53 activity. p53 functions as a transcription factor and induces the transcription of p21, a cyclin-dependent kinase inhibitor (CKI). The p21 interacts with multiple cyclin-dependent kinase (Cdk)-cyclin complexes, inhibits the kinase activity, and halts the cells in G1 phase. p21 also binds proliferating cell nuclear antigen (PCNA), inhibiting DNA synthesis.

2.6 DNA Replication[14-19]

As described earlier, the replication of DNA occurs during the S phase of the cell cycle. The S phase occupies approximately 30% of the cell-cycle time. The replication of DNA is a semi-conservative process, wherein each parental strand of the DNA helix serves as a template for the synthesis of a new and complementary daughter strand. In human diploid cells, this involves the replication of 6 billion base pairs of DNA.

Many enzymes and proteins are involved in DNA replication. The key enzyme is *DNA polymerase*, which catalyzes the ligation of the deoxyribonucleoside 5′-triphosphates (dNTPs) to generate the growing DNA chain. Eukaryotic cells contain 5 types of DNA polymerases: α, β, γ, δ, ϵ. The properties of the various human DNA polymerases are described in Table 2. The DNA polymerase γ is restricted to the mitochondria, where it is responsible for mitochondrial DNA replication. The other four DNA polymerases are localized in the nucleus. DNA polymerase δ is the major replicating enzyme in human cells.

TABLE 2. The structural and functional properties of human DNA polymerases

DNA polymerase	*Size (catalytic subunit) [kilodaltons]*	*Location*	*Function in the cell*
α	160–185	Nucleus	Lagging strand replication
β	40	Nucleus	DNA repair
γ	125	Mitochondria	Replication of mitochondrial DNA
δ	125	Nucleus	Leading and lagging strand replication
ϵ	210–230	Nucleus	DNA repair (?)

The process of DNA replication on each chromosome is initiated at designated positions, referred to as *origins of replication (ori)*. Each human chromosome has multiple ori placed at every 150–200 kilobase pairs. There are approximately 30,000 initiation sites found over the entire human genome. Thus, multiple sections of the genome are replicated simultaneously. Each small replicating unit is termed a *replicon,* and has its own ori site where DNA synthesis is initiated. The process of DNA replication proceeds bi-directionally on the chromosome until each replicon comes into contact with the next one. Thus, an entire chromosome can be replicated completely during the S phase of the cell cycle.

As the two parent DNA strands unwind and separate, DNA replication begins at ori and proceeds down the two DNA strands (Figure 7). Because of the inherent properties of DNA polymerase, daughter strand synthesis can proceed from the ori only in the 5′-to-3′ direction. Thus, one strand is synthesized in a 5′-to-3′ direction and the opposite strand is also synthesized in the 5′-to-3′ direction. As there is no DNA polymerase that can synthesize DNA in a 3′-to-5′ direction, a DNA strand cannot be used as a template in the 3′-to-5′ direction. Thus, replication of the 3′-to-5′ strand above the ori is accomplished by synthesis of short fragments of DNA, called *Okazaki fragments.* The Okazaki fragments are approximately 200 nucleotides in length, and are synthesized in a 5′-to-3′ direction. The resulting fragments are then joined by an enzyme called *DNA ligase* to give one continuous DNA strand. The DNA strand that is synthesized continuously in the 5′-to-3′ direction is called the *leading strand of DNA synthesis*, since it starts at a fixed point and dictates DNA synthesis. The strand of DNA that is synthesized in the 5′-to-3′ direction in short pieces (i.e., discontinuously) is called the *lagging strand of DNA synthesis.*

The *replication fork* refers to that part of the DNA molecule that is being replicated at a given time, and represents the region between the unreplicated segment of the DNA molecule and a newly replicated portion of

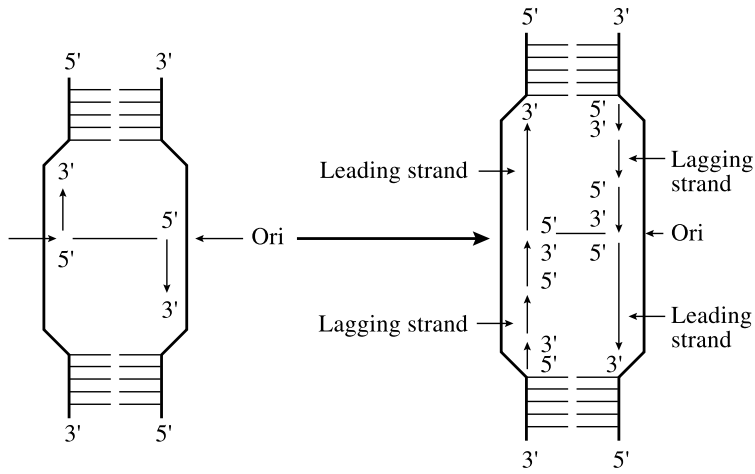

FIGURE 7. Replicon. DNA polymerase can synthesize DNA only in a 5′-to-3′ direction. For both strands of the DNA helix to serve as templates, one strand (i.e., the leading strand) is synthesized continuously in a 5′-to-3′ direction while the other strand (i.e., the lagging strand) is synthesized discontinuously in short fragments but still in a 5′-to-3′ direction. The short DNA fragments (Okazaki fragments) are subsequently joined together by DNA ligase.

DNA. Since DNA is synthesized bi-directionally, each replicon contains two replication forks. A specific initiator protein has the ability to recognize the origin sequence and signals the initiation of DNA synthesis. It has been hypothesized that this initiator protein binds the ori sequence, and attracts the DNA-replicating complex to this particular site on the DNA molecule.

All DNA polymerases must have a *primer* (i.e., a free 3′-OH end of a polynucleotide). The primer in DNA replication is not DNA, but rather is a small segment of RNA measuring 5 to 10 nucleotides in length, which is synthesized by the enzyme *DNA primase*. DNA primase initiates the synthesis of an RNA molecule at the ori, and DNA polymerase uses this RNA primer to add deoxyribonucleotides to the 3′-OH group of the RNA and synthesizes a new DNA strand that is complementary to the template strand. After completion of DNA synthesis, the RNA molecule is removed from the DNA helix and the resulting gap in the DNA is filled by a DNA polymerase.

The various proteins that play an important role on the process of DNA replication are listed in Table 3. The separation of the two strands of DNA is catalyzed by an enzyme called *DNA helicase*, which breaks the hydrogen bonds holding the DNA strands together. The DNA helix is subsequently

TABLE 3. Proteins involved in DNA replication

Protein	*Function*
DNA helicase	Unwinds DNA and breaks hydrogen bonds
Single-stranded DNA-binding protein (RPA)	Binds single-stranded DNA to prevent hydrogen bonding
Proliferating cell nuclear antigen (PCNA)	Stimulates DNA polymerase δ activity
DNA polymerase δ	Leading and lagging strand DNA replication and 3′–5′ exonuclease proofreading
DNA polymerase α/DNA primase complex	Synthesis of RNA primers and lagging strand synthesis
DNA ligase	Seals 3′ terminal hydroxyl and 5′ terminal phosphate groups of adjacent nucleotides in DNA
Ribonuclease H1 (RNase H1)	Removes RNA from RNA-DNA hybrid
DNA topoisomerase	Relaxes DNA by breaking and resealing phosphodiester bonds

unwound, and the strands remain separated through the action of a protein called *replication protein A (RPA)*. RPA is a single-stranded DNA-binding protein (Figure 8). The DNA helicase acts at the edge of the replication fork, opening and unwinding the DNA as replication proceeds along the DNA molecule. As the helicase unwinds the DNA at the replication fork, the DNA helix downstream becomes tightly wound and supercoiled. The tension on the DNA molecule is released by the action of *DNA topoisomerase,* which breaks phosphodiester bonds, unwinds the downstream DNA helix, and then reseals it by forming new phosphodiester bonds. Both DNA helicases and DNA topoisomerases play a pivotal role in the process of DNA replication and transcription.

The DNA polymerases catalyze the formation of phosphodiester bonds between the adjacent deoxyribonucleotides in the DNA molecule. All DNA polymerases catalyze the synthesis of DNA only in the 5′-to-3′ direction. DNA polymerase δ is the major replicating protein in human cells, and is involved in both leading and lagging strand replication. DNA polymerase α is complexed with another protein, the DNA primase. Together, these proteins are involved in the replication of the lagging strand. DNA primase makes the small RNA primers with DNA polymerase α. Deoxyribonucleotides are added to the 3′ terminal of the primer for a short distance of about 30 nucleotides. The DNA polymerase α/DNA primase complex subsequently falls off the DNA molecule, and is replaced by DNA polymerase δ, which continues the synthesis of the growing DNA chain. The RNA primers used by DNA polymerases must be removed from the DNA

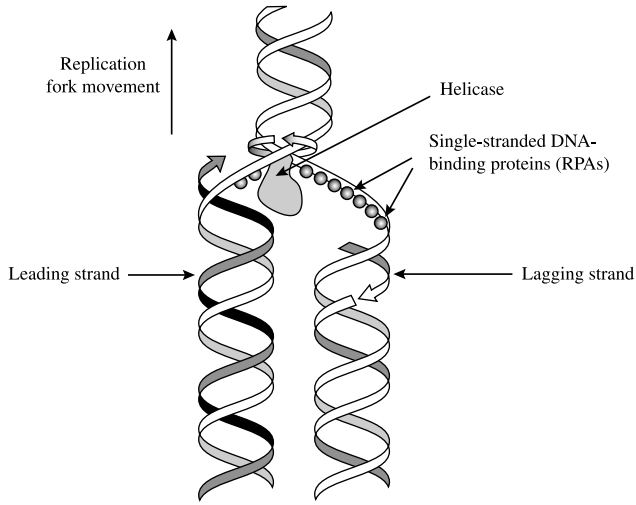

FIGURE 8. The replication of a DNA molecule, illustrating the interaction of the helicase and DNA-binding proteins at the replication fork.

molecule. This is accomplished by the action of the enzyme *RNase H1,* which specifically degrades RNA present in a DNA/RNA hybrid. DNA polymerase later completes the DNA synthesis of the lagging strand by filling in the gap. Then the ligation of the 3′-OH terminus of the DNA of one Okazaki fragment with the 5′ terminal phosphate of DNA of the adjacent fragment occurs through the formation of a phosphodiester bond. This reaction is catalyzed by DNA ligase.

DNA polymerase β and ϵ serve in the process of DNA repair, and are not directly involved in replicating the entire genome. Finally, DNA polymerase γ is responsible for replicating the circular double-stranded DNA found in mitochondria.

An additional protein involved in the replication of DNA in human cells is termed *proliferating cell nuclear antigen (PCNA).* PCNA forms part of the DNA polymerase δ complex and stimulates the activity in the DNA polymerase. The interactions of various proteins involved in DNA synthesis in the lagging strand are depicted in the model shown in Figure 9.

Some DNA polymerases (e.g., DNA polymerase δ) have intrinsic 3′-5′ *exonuclease* activity, which removes bases sequentially from the end of the DNA molecule (i.e., the 3′ end). This nuclease activity plays a critical role in preventing mistakes in base pairing during DNA replication. For example, if

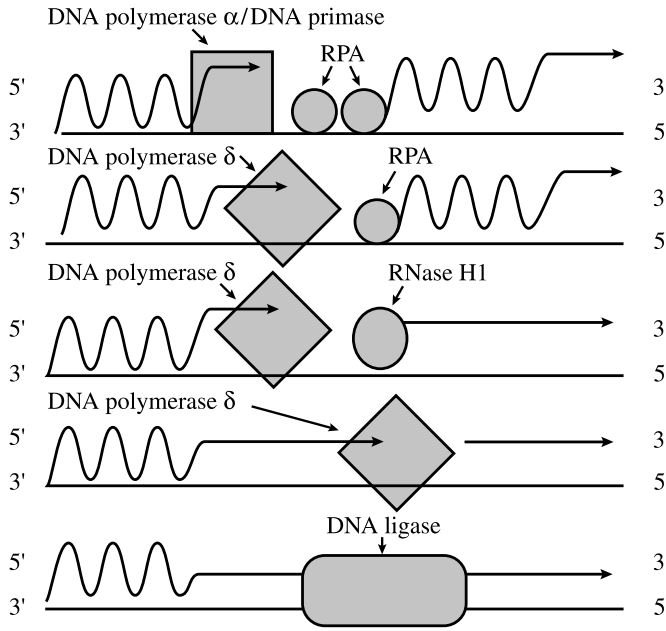

FIGURE 9. Model for DNA replication in human cells. Replication protein A (RPA), a single-stranded DNA-binding protein, separates the DNA strands to allow the DNA polymerase α/DNA primase complex to bind to the DNA and initiate synthesis of an RNA primer (indicated by the wavy line). DNA polymerase α adds approximately 30 deoxyribonucleotides to the 3′ end of the RNA primer. The DNA polymerase δ displaces the RNA polymerase α/DNA primase complex and extends the DNA strand by adding deoxyribonucleotides to the 3′ end of the newly synthesized DNA strand. Upon completion of DNA synthesis, RNase H1 removes the RNA primer. The DNA polymerase δ fills in the gap using the opposite DNA strand as a template. Finally, the two Okazaki fragments are joined together. This reaction is catalyzed by DNA ligase.

a C on the new DNA strand binds to an A on the template strand, subsequent replications of this mistake result in a CG base pair molecule instead of an AT base pair. The substitution of one base pair by another leads to a mutation in the DNA molecule which may have an impact upon cellular function. The 3′-5′ exonuclease recognizes these mispairs as soon as they occur and removes the newly inserted incorrect base. The DNA polymerase then inserts the proper base into the growing DNA chain. This exonuclease activity of DNA polymerase is termed the *proofreading function.*

As mentioned in Section 2.3, the ends (telomeres) of all chromosomes maintain the overall integrity of the chromosomes. Telomeres consist of the

base sequence TTAGGG, whose elements are randomly repeated 100 to 1,000 times. Because DNA polymerases function only in the 5′-to-3′ direction, they are unable to copy the extreme 5′ ends of linear DNA molecules. These sequences (i.e., telomeres) are replicated by the action of the enzyme telomerase, which is a reverse transcriptase. Reverse transcriptases synthesize DNA from an RNA template. Telomerases carry their own template RNA complementary to the telomere repeat sequences. The RNA template allows telomerase to generate multiple copies of the telomeric repeat sequences, thus maintaining telomeres in the absence of a conventional DNA template to direct their synthesis.

Despite the accuracy of DNA replication, cellular genomes are far from static. Gene rearrangements and mutations are required to maintain genetic diversity among individuals. To this end, recombination between homologous chromosomes occurs during meiosis and allows parental genes to be rearranged in new combinations in the next generation of cells. The rearrangements of DNA sequences within the genome create novel combinations of genetic information. In some instances DNA rearrangements are programmed to regulate gene expression during the cellular processes of differentiation and development. A striking example of this is the rearrangement of antibody genes during the development of the immune system. A key feature of both immunoglobulins and T-cell receptors is their enormous diversity, which allows different antibody or T-cell receptor molecules to recognize a variable array of antigens. These diverse antibodies and T-cell receptors are encoded by unique lymphocyte genes that are formed during the development of the immune system as a result of site-specific recombination between distinct segments of immunoglobulin and T-cell receptor genes.

2.7 Mutations and DNA Repair Mechanisms[20-26]

Mutations are the result of permanent changes in the base sequence of the DNA molecule, and are central to the pathogenesis of all human genetic diseases. The various classes of mutations that occur in DNA molecules are listed in Table 4. Many of the concepts concerning the different types of mutations that occur in DNA, and the potential mechanisms associated with the production of these mutations, were originally developed in bacterial cell model systems. Recently, our knowledge base has expanded in the area of the molecular basis of mutations in eukaryotic cells. Studies of diseased human cells have established common mechanisms by which DNA undergoes mutation. More importantly, DNA repair mechanisms have been defined.

Many of the mutations that occur in DNA are the result of *single-base-pair substitutions* in which one base pair (e.g., an adenine–thymine pair) is replaced by a second base pair (e.g., a guanine–cytosine pair). The substitution of one

TABLE 4. The classes of mutations found in human DNA

Class	*Result*
Single-base pair substitutions (point mutations)	
Altered structure of gene product	
Missense mutation	Single amino acid replacement in the protein
Nonsense mutation	A termination codon in the middle of the gene results in premature termination of protein synthesis
RNA-splicing mutation	The protein may be missing part or all of an exon sequence
Altered quantity of gene product	
Mutations in regulatory sequences	Transcription of the gene is altered, which can reduce or eliminate the gene product
Mutations in RNA processing and translation	The stability of messenger RNA is altered, which may reduce the amount of gene product
Insertions or deletions	
One or two base pairs (frameshift mutations)	The addition or deletion of one or two base pairs can affect the reading frame of the gene, resulting in a grossly altered or absent gene product
Large number of base pairs	Large pieces of the DNA may be lost, or large segments of DNA may insert into the middle of a gene, resulting in loss of function
Expansion of trinucleotide repeat sequences	
	Unstable trinucleotide repeats can suddenly expand in number, resulting in the alteration of production or structure of a particular gene product
Chromosomal alterations	Inversions, translocations, duplications or gene amplification may result

base pair by a second base pair elicits a change of codon that can lead either to a *missense mutation* (where one amino acid replaces another amino acid in a protein) or to a *nonsense mutation* (where one of the terminator codons appears in the middle of a gene). With a nonsense mutation, there is no transfer of an RNA molecule to recognize these codons, and protein synthesis terminates at the site of the nonsense codon. This leads to the production of a truncated polypeptide.

A mutation that alters the splice acceptor or splice donor sequences can result in apparent splicing of an RNA transcript. This leads to the production of an mRNA that may be missing a substantial part of a particular exon, and thus codes for a mutant protein. Other base pair substitutions can occur in regulatory sequences required for the binding of transcription factors or RNA polymerase. In this instance, the quantity of the product produced by the gene

that is controlled by these sequences is dramatically altered. In the extreme case, base pair substitutions can lead to a complete absence of the gene product, or to a dramatic increase in the amount of a particular gene product.

Frameshift mutations are caused by the addition or deletion of one or two base pairs within the coding sequence of a gene. This alters the reading frame of the mRNA. Thus, the mRNA is translated out of the frame from the site of the insertion or deletion of the base pair. This results in the production of a protein that is altered in its amino acid sequence, starting from the point of the insertion or deletion of the base pair and continuing to the end of the protein. Often, the altered reading frame also leads to the production of a termination codon in the middle of the gene. This results in premature cessation of protein synthesis.

The insertion and deletion of many base pairs can also occur with DNA molecules. Deletion mutations can occur in a chromosome with the loss of hundreds to thousands of base pairs from the DNA, with the result that the deleted genetic material is permanently lost. Large insertions of DNA sequences have been described. These are caused by transposon-like elements, often repetitive DNA sequences such as long interspersed nuclear element (LINE) repeats.

In summary, the possible changes in DNA that give rise to mutations may be illustrated by considering the following literary masterpiece:

Wild type	The cat sat on the mat.
Substitution	The **r**at sat on the mat.
Insertion (single)	The cat s**h**at on the mat.
Insertion (multiple)	The cat**tle** sat on the mat.
Deletion (single)	The c**.**t sat on the mat.
Deletion (multiple)	The cat **… ..** the mat.
Inversion (small)	The **tac** sat on the mat.
Inversion (large)	**Tam eht no tas tac eht.**

DNA polymerases catalyze the proper pairing of A to T and G to C with very high accuracy. However, mispairing occurs at a frequency of approximately 10^{-5} bases. For example, if an AC pair forms instead of an AT pair, and if such a mispair remains in the DNA molecule, the initial AT pair that has become an AC pair now gives rise to a GC pair during the next replication cycle. In order to keep the mutation rate at a low level, eukaryotic cells have devised mechanisms for correcting base mispairs before they become a permanent feature of the DNA.

Bases that are present in DNA molecules can undergo spontaneous damage or modification. One frequent form of modification occurs with the purine

bases A and G. Purine residues may be lost from the DNA molecules by a process called *depurination.* The glycosidic bond between the deoxyribose and the base is hydrolyzed, which leads to a gap in one of the DNA strands. This damage must be corrected before the DNA is replicated, otherwise a mutation ensues. The bases C, A and G are capable of undergoing spontaneous *deamination*, wherein the base loses an amino group and its structure is changed. For example, when cytosine is deaminated it becomes uracil. This leads to the presence of uracil in DNA instead of cytosine. Uracil now appears with an adenine residue during the next replication cycle. The original GC pair, which after deamination is now a GU pair, subsequently becomes an AT pair. Ultraviolet rays from sunlight are a common mutagenic agent that causes bond formation between adjacent pyrimidines on the same DNA strand. The most frequent type of pyrimidine dimer is the TT dimer. The presence of a TT dimer in the DNA molecule blocks DNA replication and leads to the death of the cell if it is not removed. The 3′-5′ exonuclease activity associated with DNA polymerase δ and ϵ is responsible for cleaving mispaired nucleotides from the 3′ end of newly replicated DNA strands. This allows the polymerase a second opportunity to add the correct base. The entire process is known as the proofreading function.

If base mispairing remains in the DNA, it leads to a mutuation at the next DNA replication cycle. However, eukaryotic cells have evolved a mechanism to deal specifically with persistent base mispairing immediately after replication. Human cells have a *methyl-directed mismatch repair system,* which appears to be similar to that of bacterial strains. The methyl-directed mismatch repair system scans the DNA molecule, and when base mispairs as well as insertions and deletions are detected, correction of the error occurs on the nonmethylated, newly synthesized DNA strand. This allows the repair system to correct the nascent strand that has a normal base in the wrong location, and prevents the mispaired bases from giving rise to a permanent mutation.

DNA molecules are methylated at specific sites, either on an A or a C residue. In human cells, C residues located in CpG islands are methylated. Methylation is a postreplication event. During the initial period of DNA replication, one strand (i.e., the template strand) is methylated, while the newly synthesized DNA strand is not methylated.

Mutator (mut) proteins are involved in methyl-directed mismatch repair. Human homologues have been identified for MutS (hMSH2 and GTBP) and MutL (hMLH1 and hPMS2), but at this time, there are no known homologues for MutH. Methyl-directed mismatch repair appears to be similar in bacteria and humans. In human cells, mismatches are recognized by the protein hMSH2 or a dimer composed of hMSH2 and GTBP. Base mispairing creates

a bulge in the DNA, which is recognized and bound by the MutS protein. The MutS protein that is bound to the mismatch recruits the MutL homologue to the site. MutH cleaves the nonmethylated DNA strand. This is followed by the stepwise removal of nucleotides by an exonuclease, and the resulting gap in the DNA molecule is repaired by DNA polymerase using the base sequence in the template strand. The final phosphodiester bond is sealed by DNA ligase.

One of the most common hereditary cancers, HNPCC (hereditary nonpolyposis colon cancer), arises from mutations in the methyl-directed mismatch repair system. HNPCC affects 1 in 200 people in North America and accounts for approximately 15% of all colon cancers. There are at least five genetic loci involved in the human mismatch repair process. These include hMSH2, hMLH1, hPMS1, and hPMS2 and the GTBP gene. Cells with HNPCC are characterized by *microsatellite instability*. Microsatellites are repetitive nucleotide sequences (di-, tri- or tetranucleotides) located throughout the human genome. The presence of these repeats in the DNA is a "road block" to the DNA polymerase molecule during DNA replication. When DNA polymerase is confronted with a long, repetitive sequence of DNA, it produces a strand of DNA with extra bases that are not base-paired with the template and that loop away from the DNA helix. The mismatch repair system recognizes these loops as defective and removes them. The loops remain if the repair system is defective. Microsatellite instability signals that the cell has developed a *mut* phenotype and has an increased rate of overall mutation. These cells also develop mutations in such genes as the p53 gene or other tumor supressor genes at a much higher rate than do normal cells.

Another type of DNA mutation is incurred through damage to bases of a DNA molecule that is not undergoing replication. Cells have evolved two major repair systems to deal with this type of DNA damage. The first system is called *base excision repair*. When a uracil residue occurs in a DNA molecule, it is recognized by uracil-DNA glycosylase and is removed from the DNA, leaving behind a gap. The lack of a base in the DNA helix is recognized by specific endonucleases known as *AP endonucleases* (which recognize *ap*urinic and *ap*yrimidinic sites in DNA). The AP endonuclease cleaves the DNA at the site of the missing base. The resulting gap is repaired by DNA polymerase, using the base present in the complementary strand as a template. This is followed by ligation via DNA ligase. If the uracil residue is not removed, it eventually results in a GU mismatch, and the original GC pair becomes an AT pair or a mutation. A more general repair mechanism, known as *nucleotide excision repair,* repairs bulky distortions in the DNA molecule. The overall scheme for nucleotide excision repair resembles that of base excision repair and methyl-directed mismatch repair. All systems have specific

proteins that recognize the damaged area of DNA, as well as specific proteins involved in the removal of the damage from the DNA. Following removal of the damage, the gap is filled by repair synthesis. This is catalyzed by DNA polymerase, and sealing is accomplished by DNA ligase.

Xeroderma pigmentosum (XP) is a rare autosomal recessive disorder characterized by skin neoplasms. Skin cells from XP patients are unable to repair DNA damage caused by exposure to ultraviolet (UV) light. UV light damages DNA, resulting in the formation of dimers between adjacent pyrimidines on the same DNA strand (e.g., TT dimer). These TT dimers distort the DNA helix and result in the cessation of replication and transcription at that point until the dimer is removed. The nucleotide excision repair system removes these TT dimers. The initial step is the recognition of the damage by the XPA protein, which binds along with XPF-ERCC1 protein and the single-stranded DNA-binding protein RPA. Helicase activity unwinds the helix and stimulates the excision activity of two endonucleases, XPF and XPG, which cut the DNA. This creates a large gap in the DNA molecule, and the 3′ hydroxyl terminus is recognized by DNA polymerase δ or ϵ, which carries out repair synthesis using the undamaged DNA strand as a template. The final nick is sealed by DNA ligase.

A new type of mutation has been recently described which results in a number of human genetic diseases. These mutations are the result of the expansion of trinucleotide repeats (CAG, CTG, CGG or GAA) found throughout the human genome. Long runs of these repeat triplets are found in exons at the 5′ or 3′ end of genes. Individuals affected with one of the expansion disorder diseases have an increase in the number of copies of the trinucleotide repeats. The expansion of the repeat sequences can alter either the structure or function of a particular protein. One of the best characterized examples of this is the trinucleotide CAG, which codes for the amino acid glutamine. In *Huntington's disease* the CAG repeat is located in the coding region of the first exon at the 5′ end of the gene. These repeats are translated, and appear as a long stretch of glutamines within the structure of the protein such that the mutant protein has a range of 40 to 100 glutamines at that particular site. All of the CAG repeat diseases are autosomal dominant disorders characterized by late-onset neuronal loss.

3. Eukaryotic Gene Transcription and Post-transcriptional RNA Processing

3.1 Chromatin Structure and Transcription[27-31]

The DNA present in all eukaryotic cells is tightly associated to histones, forming *chromatin*. Moreover, the packaging of eukaryotic DNA into chromatin

has important ramifications in terms of its availability to serve as a template for transcription. Thus, chromatin structure is a critical aspect of eukaryotic gene expression. Actively transcribed genes are situated in regions of decondensed chromatin. The tight coiling of DNA around the *nucleosome* poses a major obstacle to transcription: the tight coiling impedes the ability of transcription factors to bind to DNA, as well as impeding the ability of RNA polymerase to gain access to the DNA template. This inhibitory effect of nucleosomes is overcome by the action of *nucleosome remodeling factors*. These remodeling factors disrupt chromatin structure, thus allowing transcription factors to gain purchase to nucleosome DNA and coordinate the assembly of the transcription complex with the promoter. A multiprotein complex, initially identified in yeast as the SWI/SNF (switch/sucrose nonfermenting) complex, has been localized in mammalian cells. SWI/SNF disrupts the nucleosome array and facilitates the transcription of DNA that was previously unavailable to the transcription complex.

Eukaryotic transcriptional activators play a dual role in modulating gene expression. In addition to promoting transcription by interacting with basal transcription factors, they stimulate changes in chromatin structure that alleviate repression by histones. The ability of RNA polymerase to transcribe chromatin templates is facilitated through the acetylation of histones, and by the association of the nonhistone chromosomal proteins HMG-14 and HMG-17 with the nucleosomes of actively transcribed genes. The signals that target HMG-14 and HMG-17 to actively transcribe genes remain an enigma.

3.2 Cis-Acting Elements[27, 30-32]

This discussion of the transcriptional control of gene expression is focused on the role of RNA polymerase II, the enzyme responsible for transcribing protein-encoding genes into mRNAs. The production of each mRNA in human cells involves complex interactions of proteins (*trans-acting factors*) with specific sequences on the DNA (*cis-acting elements*). Cis-acting elements are short base sequences adjacent to, or within, a particular gene. Alternatively, they can be sequences that occur several thousand base pairs away from a particular gene. Cis-acting elements are sequences required for the recognition of a gene by RNA polymerase II. These sequences also serve as binding sites for the proteins that regulate the rate and specificity of transcription. The initiation of transcription is dictated by sequences that are present in each gene. The major cis-acting sequences of a gene are illustrated in Figure 10, and include the following:

1. The *core promoter element* is situated 5′ to the gene and consists of the sequences where the transcription complex containing the RNA polymerase II assembles on the DNA molecule. There are two fixed sequence elements: the *initiator element (Inr),* which determines the transcription

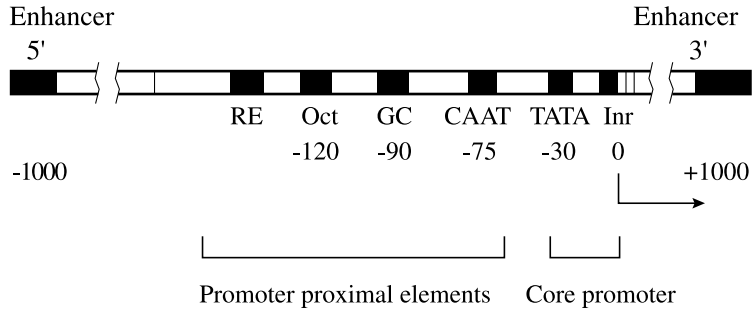

FIGURE 10. The localization of cis-acting sequences in a typical human gene. The core promoter is composed of TATA and initiator (Inr) sequences. The TATA sequence, located 30 base pairs upstream of the Inr sequence, is the binding site for the TATA-binding protein (TBP). The Inr sequence is where RNA polymerase II binds and initiates transcription. The promoter proximal elements are located 50 to several hundred base pairs upstream of the Inr site and include the common sequences CAAT, GC and Oct. These sequences are the binding sites for upstream transcription factors. Sequences in the promoter proximal regions are the response elements (RE), which are the binding sites for inducible transcription factors. Situated thousands of pairs away, either 5′ or 3′ to the gene of interest, are enhancer elements that bind activators.

start site, and the *TATA element,* which is located 25–30 base pairs upstream from the Inr. The promoter initiation site defines the location and the direction of transcription.

2. The *promoter proximal elements* are composed of two types of cis-acting sequences located 50 to a few hundred base pairs upstream from the start site. The first type of promoter proximal element comprises a class of base sequences (e.g., CAAT or GC) found in many genes, and these sequences function as binding sites for proteins called *upstream transcription factors.* The second type of promoter proximal element is the *response element (RE).* The RE contains sequences that are found in promoters controlled by a particular stimulus, e.g., genes that respond to particular glucocorticoid stimulation or iron response elements (IRE) implicated in intestinal iron absorption.
3. The *promoter distal elements* are cis-acting sequences found thousands of base pairs away from the start site of transcription. These distal sites are known as *enhancers* or *silencers* and are situated either upstream or downstream from the gene that they regulate. Enhancers, like promoters, act by binding transcription factors that subsequently regulate RNA polymerase. Because of the looping of the DNA helix, this allows a transcription factor bound to a distant enhancer to lie in relative proximity to the

upstream promoter and interact with RNA polymerase or basal transcription factors at the promoter. The binding of specific transcriptional regulatory proteins to enhancers is a mechanism responsible for controlling gene expression during development and cell differentiation. In addition, this mechanism also serves to mediate the response of cells to hormones and growth factors.

Transcription is initiated by the binding of a variety of transcription factors and the enzyme RNA polymerase to the promoter site. A large number of transcription factors serve to recruit the RNA polymerase to the promoter site. Transcription factors bind to sequences in the promoter site on the DNA molecule or they can bind to one another in several different areas to determine whether RNA polymerase will or will not transcribe a particular gene. The structural features of typical transcription factors are illustrated in Figure 11. Transcription factors are characterized by the following shared features: (1) binding to specific DNA sequences; (2) interaction with other transcription factors to regulate transcription; (3) a DNA-binding domain made up of the amino acid sequences that recognize and bind specific DNA sequences; and (4) a transactivation domain comprising the amino acid sequences required for the activation of transcription.

Transcription factors may have similar DNA-binding domains but different transactivating domains. Thus, they bind the same sequence of DNA but activate transcription in a different manner. Alternatively, transcription factors have similar transactivating domains but different DNA-binding domains. In this case, the transcription factors bind to different sequences of DNA, although the process of activation is similar. RNA polymerase catalyzes the formation of a phosphodiester bond by attaching the 5′-phosphate of the incoming ribonucleotide to the 3′-hydroxyl of the growing RNA chain. Multiple RNA transcripts may be synthesized from a single DNA molecule through the sequential binding of additional RNA polymerase to the promoter sequence.

3.3 Trans-Acting Transcription Factors[27, 30-36]

Trans-acting transcription factors bind to cis-acting elements on the DNA and interact with other transcription factors. These proteins control the initiation of transcription and comprise the following:

1. *General transcription factors* are polypeptides that assemble at the core of the promoter site and recruit RNA polymerase II to that site to form the pre-initiation complex.
2. *Upstream transcription factors* are proteins that bind the common cis-acting sequences proximal to many promoters, such as the sequences CAAT and GC.

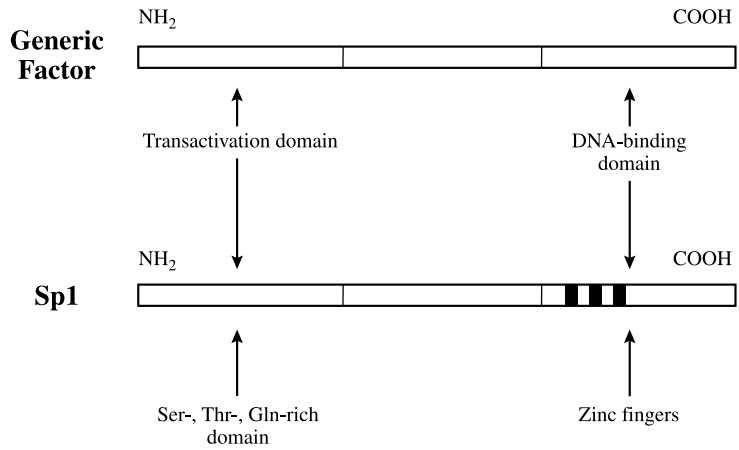

FIGURE 11. Common functional domains in transcription factors. Many transcription factors contain two common functional domains. The transactivation domain represents the amino acid sequence of the protein that interacts with other protein factors and is responsible for activating the transcription of genes. The DNA-binding domain comprises amino acid sequences that are responsible for interacting with and binding to specific DNA sequences. The upstream transcription factor (Sp1) binds to GC sequences through its DNA-binding domain, which includes three zinc finger motifs. The transactivation domain of Sp1 is rich in the amino acids serine, threonine and glutamine, and interacts with the TAFIID 110 subunit of TFIID.

3. *Inducible transcription factors* are proteins that respond to external stimuli that activate them and in turn promote their binding to the response element (RE) sequences. This results in increased transcription of genes containing the particular response element sequence.
4. *Activator proteins* are transcription factors that bind enhancers and increase transcriptional initiation of a particular gene.
5. *Repressor proteins* are transcription factors that silence and inhibit transcriptional initiation of a particular gene.

The ability of proteins to bind DNA is a reflection of their amino acid sequences and the formation of specific motifs. A well-characterized DNA-binding domain is the *zinc finger domain*. This contains repeats of cysteine and histidine residues that bind zinc ions within the DNA-binding domain. Zinc finger domains are common among transcription factors that regulate RNA polymerase II promoters, including the common transcription factor Sp1, the general transcription factor TFIIA and the glucocorticoid receptors. The *helix-*

turn-helix motif is found in the *homeodomain* proteins, among other eukaryotic cell proteins. These play a central role in the regulation of gene expression during embryonic development. The molecular cloning and analysis of these genes have shown that they contain conserved sequences of 180 base pairs (homeoboxes) that encode the DNA-binding domains (homeodomains of transcription factors). Homeobox genes are highly conserved across a variety of species. Finally, leucine zipper and helix-loop-helix proteins are two other families of DNA-binding proteins that contain DNA-binding domains formed by dimerization of two polypeptide chains. They appear to play important roles in regulating tissue specific and inducible gene expression.

3.4 Initiation of Transcription by RNA Polymerase II[37-40]

A set of *basal transcription factors* interact with the cis-acting core promoter sequences to form a *basal transcription complex* (Figure 12) during the process of the initiation of transcription by RNA polymerase II. These transcription factors are named TFII for transcription factors associated with RNA polymerase II, followed by a letter (A, B, D, E, F or H). Other transcription factors bind to DNA sequences that control the expression of distinct genes and are thus responsible for regulating gene expression.

TFIID is the first TF to bind to the core promoter sequence, and is made up of a variety of proteins. These include a *TATA-binding protein* (TBP) that recognizes the TATA sequence at all promoter sites. The remaining proteins in TFIID are called TBP-associated factors (TAFs). Once TFIID is bound at the TATA sequence, a pre-initiation complex is formed with the recruitment of TFIIA, TFIIB, TFIIF/RNA polymerase II, TFIIE and TFIIH (Figure 12). The synthesis of mRNA then proceeds with the movement of RNA polymerase II away from the promoter region, and elongation of the mRNA transcript.

3.4.1 ACTIVATION OF TRANSCRIPTION[30, 31]

A variety of short cis-acting sequences (Figure 10) that are located upstream of the TATA sequence facilitate the efficient and specific recognition of the core promoter by the basal transcription complex. The sequences include the common sequences found in RNA polymerase II promoters, including CAAT, Oct and GC. Specific upstream TFs recognize these sequences and bind to the DNA through a set of interactions among the DNA-binding domain of the TF, the DNA sequence, and the amino acid sequence of the transcription factor. For example, the upstream transcription factor Sp1 binds to GC sequences and subsequently interacts with the TFIID bound at the TATA box to activate transcription.

The activation mechanism for transcription of some classes/families of genes is shared in common under specific conditions. For example, exposure of cells to glucocorticoids or phorbol esters elicits a specific induction of the

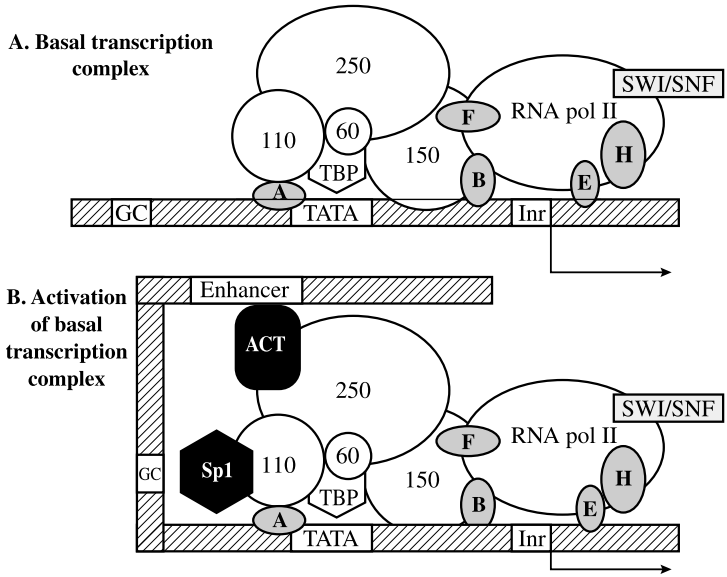

FIGURE 12. Model of the initiation of transcription by RNA polymerase II. The binding of the general transcription factors (GTFs) is illustrated in panel A, which depicts the formation of the basal transcription complex. RNA polymerase II (RNA pol II) binds to the core promoter. The TATA-binding protein (TBP), a subunit of TFIID, binds to the TATA sequence and facilitates the binding of the TBP-associated factors (TAFs). TBP and some of the TAFs are indicated as 250, 110, 150 and 60. Once TFIID is bound to the TATA sequence, the other GTFs (A, B, F, E and H) and RNA polymerase II bind to the core promoter, thus forming the basal transcription complex. Also indicated is the SWI/SNF multiprotein complex, associated with RNA polymerase II. This multiprotein complex is necessary for the disruption of chromatin structure. The activation of the basal transcription complex is illustrated in panel B: the activation of transcriptional initiation by Sp1 bound to the GC sequence and interacting with TAFIID 110. Further activation results from the binding of an activator protein (ACT) to an enhancer sequence located 1,000 base pairs from the core promoter. The ACT is brought into close proximity with the basal transcription complex by looping away from the DNA between the enhancer sequence and the core promoter to allow the activator to interact with TAFIID 250.

transcription of all the genes induced by these molecules. These inducible responses are attributed to upstream RE sequences in special promoters that function as binding sites for specific inducible transcription factors. An example of inducible control is the binding of the factor AP1 (made up of subunits encoded by *fos* and *jun*) to the TRE sequence (TGACTCA) in genes that are activated by phorbol esters, growth factors or cytokines. In the absence of phorbol ester, AP1 is phosphorylated, and then cannot bind to DNA (i.e., it is

inactive). The activation of AP1 involves its dephosphorylation, such that it may bind to promoters containing TRE sequences. The binding of AP1 increases the rate of initiation of transcription.

Another example is steroid hormones, which bind to specific receptors to form an activated complex that is capable of binding to RE sequences found in specific genes. Steroid-receptor proteins comprise a DNA-binding domain that contains zinc finger motifs, and a hormone-binding domain. Activated steroid-receptor proteins are essentially TFs that, when bound to RE sites in the DNA, activate transcription of a specific class of genes through activation of the initiation of transcription by RNA polymerase II. All genes that contain the common RE sequence are simultaneously activated. This allows the cell to coordinate the inducible expression of multiple genes collectively in response to specific hormone signals.

One important class of membrane protein receptors has intrinsic tyrosine kinase activity. The ligands of these receptors include growth factors and cytokines, both of which regulate cell growth. Important to this class of receptors are the *signal transducers and activators of transcription (STATs)*. STATs are transcription factors that reside in the cytoplasm in an inactive form. The binding of cytokines to membrane-bound receptors leads to phosphorylation of the receptor by activation of the receptor tyrosine kinase activity. This provides a binding site for the STAT proteins. The bound STAT proteins are phosphorylated on tyrosine residues and undergo dimerization prior to migration to the nucleus. There they act as TFs by binding to specific DNA sequences upstream of the TATA sequence.

3.4.2 EUKARYOTIC REPRESSORS[41]

Gene expression in eukaryotic cells is regulated by repressors as well as by activators. Repressors bind to specific DNA sequences and inhibit transcription through a variety of mechansims. In some instances the repressors simply interfere with the binding of other transcription factors to DNA. Other repressors have been shown to compete with activators in binding to specific regulatory sequences. As a result, their binding to a promoter or enhancer blocks the binding of the activator, thereby inhibiting transcription. Other repressors contain specific functional domains, called repression domains, that inhibit transcription through protein–protein interactions.

The regulation of transcription by repressors as well as by activators extends the repertoire of mechanisms that control the expression of eukaryotic genes. One important role of repressors is the inhibition of expression of tissue-specific genes in appropriate cell types. Other repressors play key roles in the control of cell proliferation and differentiation in response to growth factors as well as hormones. Such intricate control is especially important

when considering the coordination required for maintaining the vertical crypt-villus and horizontal jejunoileocolonic axis of the gut.

3.5 Post-Transcriptional Processing and the Regulation of Eukaryotic Gene Expression[42-57]

The human genome contains coding information for approximately 100,000 different RNA molecules. However, within a single cell different genes are expressed at different times through a process known as *differential gene expression.* Differential gene expression occurs in response to signals that occur during cell development, proliferation and differentiation. The orderly, programmed expression of every gene thereby plays a central role in cellular and whole-organ homeostasis. Thus, it is not surprising that cells have evolved elaborate mechanisms that specifically control gene expression for particular genes. The pivotal step in all cells for the regulation of gene expression is at the level of transcription. The complex task of regulating gene expression in the many differentiated cell types in higher eukaryotes is a reflection of the combined actions of a diverse array of transcriptional regulatory proteins.

While the cellular events associated with the regulation of transcription represent the predominant step in the regulation of eukaryotic gene expression, additional levels of control include the following: (1) controlling the processing of mRNA by determining which exons present in the initial mRNA transcript are retained in the mature and fully functional mRNA; control mechanisms include either the alternative splicing of exons or the differential polyadenylation of the initial mRNA transcript; and (2) controlling the stability or the rate of degradation of the mature mRNA transcript. As well, the packaging of DNA into chromatin and its modification by methylation add further dimensions to the control of eukaryotic gene expression.

3.5.1 RNA PROCESSING[47]

The majority of newly synthesized RNAs are subsequently modified in a variety of ways to be converted to their functional forms. The regulation of the processing of RNA adds an additional level of control in eukaryotic gene expression.

RNA polymerase I is devoted to the transcription of rRNAs in the nucleolus. The processing of the 45S initial transcript, pre-rRNA, involves methylation of the RNA as well as ribonuclease-mediated cleavage of segments of the initial transcript to yield the 28S, 18S and 5.8S rRNAs (Figure 13).

The 5S tRNA is transcribed from a separate gene by RNA polymerase III, and the large precursor (pre-tRNA) undergoes cleavage and methylation. The processing of the 3′ end of tRNA involves the addition of a CCA terminus,

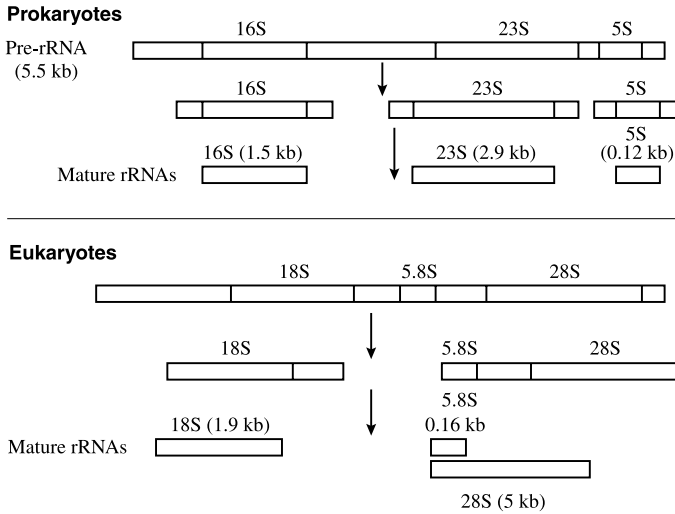

FIGURE 13. The processing of ribosomal RNAs. Prokaryotic cells contain three RNAs (16S, 23S and 5S) that are formed through cleavage of a pre-rRNA transcript. Eukaryotic cells contain four RNAs. One of these (5S rRNA) is transcribed from a separate gene; the other three (18S, 28S and 5.8S) are derived from a common pre-rRNA. Following cleavage, the 5.8S rRNA (which is unique to eukaryotes) becomes hydrogen-bonded 28S rRNA.

such that all tRNAs have the sequence CCA at the 3′ end. This sequence is the site of an amino acid attachment to the tRNA during protein synthesis.

In eukaryotic cells, the mRNA synthesized in the nucleus by RNA polymerase II is exported to the cytoplasm before it can be used as a template for protein synthesis. The initial products of transcription in eukaryotic cells (pre-mRNAs) are extensively modified prior to export from the nucleus. The processing of eukaryotic mRNAs is illustrated in Figure 14. This processing involves the modification of both ends of the mRNA, as well as the removal of introns from its mid portion. The 5′ end of pre-mRNA is modified by the addition of a *7-methylguanosine (m^7G) Cap*. The 5′ Cap has several putative functions, including (1) protecting the RNA from 5′-to-3′ exonuclease degradation; (2) facilitating transport to the cytoplasm; (3) facilitating RNA splicing; and (4) assisting in the alignment of mRNAs on the ribosomes during translation.

The 3′ end of most eukaryotic mRNAs is modified by a processing reaction called *polyadenylation*. The signal for polyadenylation is the hexanucleotide sequence AAUAAA. This AAUAAA sequence is recognized by a protein complex that cleaves the RNA chain 15 to 30 nucleotides farther downstream.

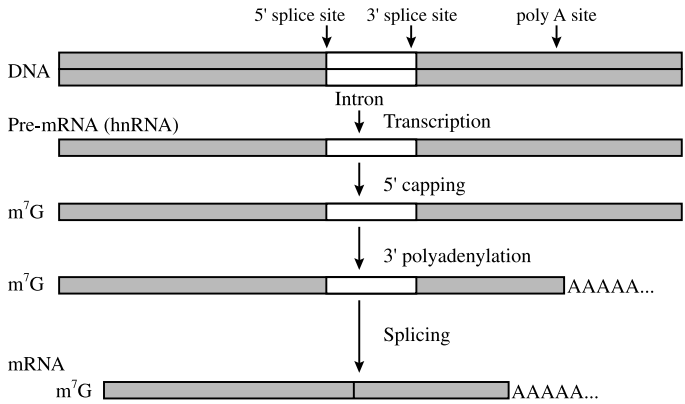

FIGURE 14. The processing of eukaryotic messenger RNAs. The processing of mRNA involves modification at the 5′ end by capping with 7-methylguanosine (m^7G), modification at the 3′ end by polyadenylation, and removal of introns by splicing. The 5′ Cap is formed by the addition of a GTP in reverse orientation to the 5′ end of the mRNA, forming a 5′-to-5′ linkage. The added G is then methylated at the N-7 position, and the methyl groups are added to the riboses of the first one or two nucleotides in the mRNA.

Subsequently, a poly A polymerase adds a poly A tail of approximately 200 nucleotides to the transcript. The initiation of polyadenylation heralds termination of transcription by RNA polymerase. The poly A tails have been envisaged to have several potential functions, including (1) facilitating transport of mRNA molecules to the cytoplasm; (2) stabilizing mRNAs in order to prevent degradation; and (3) facilitating translation by enhancing the recognition of the mRNA by the translational machinery.

Untranslated regions (UTRs) are found at both the 5′ and 3′ ends of the mRNA. UTRs represent sequences in the exons that remain in the mRNA but are not translated into protein. The 5′ and 3′ UTRs contain signals that are necessary for the processing of the RNA and subsequent translation into protein.

3.5.2 SPLICING MECHANISMS[46, 49, 50]

Most genes contain multiple introns, which account for about 10 times more pre-mRNA sequences than do the exons. Thus, the most striking modification of the pre-mRNAs involves the removal of introns by a process known as *splicing*. Splicing involves endonucleolytic cleavage and removal of intronic RNA, and end-to-end ligation (i.e., splicing) of exonic RNA segments (Figure 15). The mechanism of RNA splicing is critically dependent on the *GT-AG rule*: introns start with GT and end with AG. The sequences adjacent to the

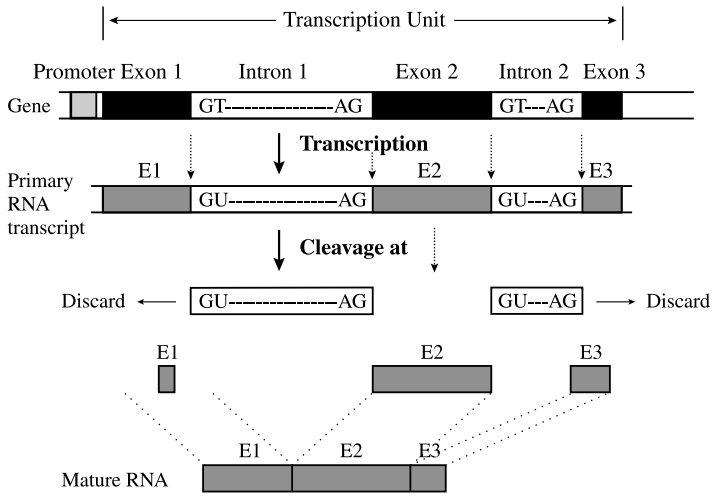

FIGURE 15. Splicing of primary RNA transcripts. RNA splicing involves endonucleolytic cleavage, removal of intronic RNA segments and splicing of exonic RNA segments.

GT and AG dinucleotides are highly conserved, and an additional conserved sequence situated just before the terminal AG at the end of the intron is the so-called *branch site*. The splicing mechanism is depicted in Figure 19 (Section 3.5.7), and involves the following steps: (1) cleavage at the 5′ splice junction; (2) joining of the 5′ end of the intron to an A within the intron (i.e., branch site) to form a lariat-shaped structure; and (3) cleavage at the 3′ splice site leading to the release of the lariat-like intronic RNA, and splicing of the exonic RNA segments. Splicing occurs in large complexes called *spliceosomes*. The RNA components of the spliceosomes are *small nuclear RNAs (snRNAs)*. These snRNAs range in size from approximately 50 to 200 nucleotides and are complexed with protein molecules to form *small nuclear ribonucleoprotein particles (snRNPs)*. SnRNPs play an important role in the splicing process. The snRNA part of the snRNP carries out the "intellectual task" of recognizing the splice and branch sites of the larger RNA molecule. In contrast, the protein part of the snRNP does the "manual labor" of cutting and reattaching the RNA molecule.

The central role that splicing plays in the processing of pre-mRNA affords another mechanism for regulation of gene expression by the control of the activity of the cellular splicing machinery. Since most pre-mRNAs contain multiple introns, different mRNAs can be produced from the same gene by

different combinations of the 5′ and 3′ splice sites. The possibility of joining exons in various combinations provides a novel mechanism for the control of gene expression through the generation of multiple mRNAs (and thus multiple proteins) from the same pre-mRNA. This process is termed *alternative splicing*, and occurs frequently in genes of higher eukaryotes. Alternative splicing affords an important mechanism for the tissue-specific and developmental regulation of eukaryotic gene expression. In the case of transcriptional regulatory proteins, alternative splicing of pre-mRNAs yields products with dramatically different functions (e.g., the ability to act as either activators or repressors of transcription). An important variation of the theme of splicing is a phenomenon known as *trans-splicing,* where exons originating from two separate transcripts are ligated together. The biological significance of trans-splicing remains to be elucidated.

3.5.3 EXON SELECTION DURING SPLICING[51-53]

An additional level of control of gene expression occurs through the process of *exon splicing* during the processing of the pre-mRNA. The cell determines which exons present in the pre-mRNA are conserved in the final mRNA. This allows for the production of more than one protein from the same gene. For example, the same gene encodes calcitonin and the calcitonin gene–related peptide (CGRP). These proteins differ with respect to their amino acid sequence, function and tissue localization. The synthesis of these different proteins using the same genetic information occurs by a combination of alternative polyadenylation and differential exon selection. This is illustrated in Figure 16.

3.5.4 RNA EDITING[54, 55]

The protein-coding sequences of some RNAs are altered by RNA-processing events other than splicing. The best-characterized example is the editing of the mRNA for apolipoprotein B, where tissue-specific RNA editing gives rise to two different forms of apolipoprotein B (Figure 17). Apo B100 is synthesized in the liver by translation of the unedited mRNA, whereas a smaller protein specific to the intestine, apo B48, is synthesized as a result of translation of an edited mRNA where a C in a single codon has been changed to a U. This nucleotide substitution alters the codon for glutamine (CAA) in the unedited mRNA to a translation termination codon (UAA) in the edited mRNA. This results in the synthesis of the shorter apo B protein. This tissue-specific editing of apo B results in the expression of structurally and functionally different proteins in the liver and intestine. The full-length apo B100 produced by the liver transports lipids of the circulation, whereas apo B48 mediates the absorption of dietary lipids by the small intestine.

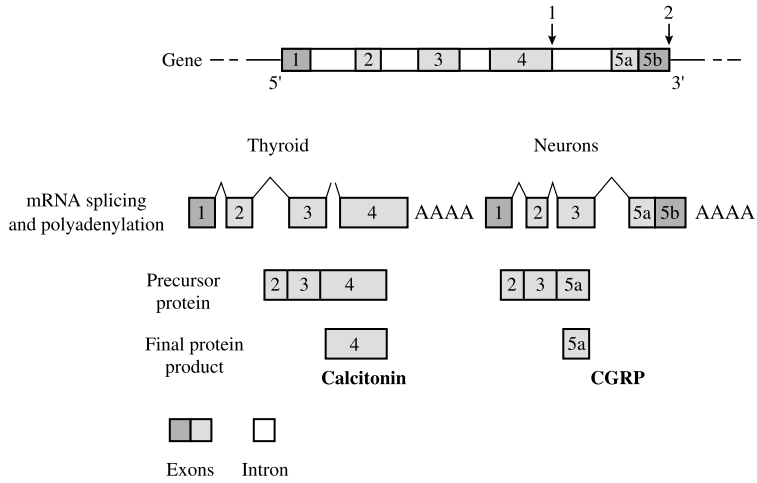

FIGURE 16. The role of exon selection in the production of two proteins from the same gene. The calcitonin gene contains two polyadenylation signals and six exons (1, 2, 3, 4, 5a, 5b). In the thyroid, the upstream polyadenylation signal (arrow 1) is recognized, and this results in cleavage and polyadenylation of the mRNA at the 3′ end of exon 4, to produce a precursor mRNA containing exons 1, 2, 3 and 4. These four exons are spliced together, forming the mature mRNA, which codes for the calcitonin precursor peptide. This peptide is processed to yield calcitonin that contains amino acid sequence information only from exon 4. In neurons, the downstream polyadenylation signal (arrow 2) is recognized, resulting in cleavage and polyadenylation of the mRNA at the 3′ end of exon 5b to form a precursor mRNA containing exons 1, 2, 3, 4, 5a and 5b. During the splicing process, exon 4 is deleted, and the mature mRNA contains exons 1, 2, 3, 5a and 5b, which code for the calcitonin gene–related peptide (CGRP). The final processing gives CGRP, which contains the amino acid information that is found in exon 5a.

3.5.5 RNA DEGRADATION[44, 45, 48]

The final aspect of the processing of an RNA molecule is its eventual degradation. The intracellular level of any particular RNA species reflects a balance between synthesis and degradation. In this way the rate at which particular RNAs are degraded constitutes another potential level at which gene expression can be controlled. In eukaryotic cells different mRNAs are degraded at different rates, and this allows for the differential regulation of eukaryotic gene expression.

The degradation of most mRNAs is initiated by the trimming of the poly A tail. This is followed by the removal of the 5′ Cap and degradation of the RNA by nucleases. The mRNA half-life varies from 30 minutes to about 24 hours. The mRNAs with short half-lives usually encode for regulatory proteins. These mRNAs often contain specific AU-rich sequences situated

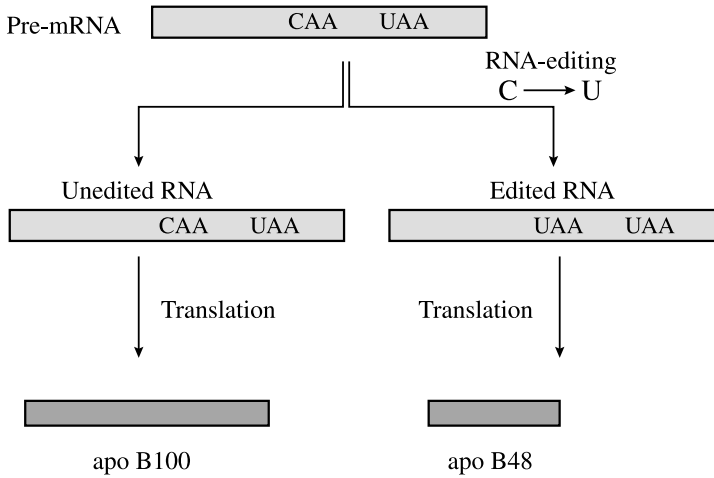

FIGURE 17. The editing of apolipoprotein B mRNA. In the human liver, unedited mRNA is translated to yield a 4,536-amino-acid protein called apo B100. In the human intestine, however, the mRNA is edited by a base modification that changes a specific C to a U. This modification changes the codon for a glutamine (CAA) to a termination codon (UAA), resulting in the synthesis of a shorter protein (apo B48, consisting of only 2,152 amino acids).

near the 3′ end, which appear to signal rapid degradation by promoting deadenylation at the 3′ poly A tail.

The stability of some mRNAs can be regulated in response to extracellular signals. For example, the level of abundance of the mRNA encoding the transferrin receptor, a cell-surface protein involved in iron uptake, is regulated by the availability of iron (Figure 18). This occurs by modulation of the stability of the transferrin-receptor mRNA. When iron is replete, the transferrin-receptor mRNA is rapidly degraded by specific nuclease cleavage that occurs at a sequence near the 3′ end. When the supply of iron is rate-limiting, the transferrin-receptor mRNA is stabilized, and this leads to an increased synthesis of transferrin receptor. Thus, more iron is transported into the cell. The regulation of the transferrin receptor is mediated by a protein that binds to specific sequences, called the iron-responsive element (IRE), which is located near the 3′ end of the transferrin-receptor mRNA. Binding protects the transferrin mRNA from cleavage and is controlled by the levels of intracellular iron.

3.5.6 PROMOTER SELECTION[56, 57]

The presence of more than one promoter within a particular gene can result in different amounts of the same gene product being produced in different

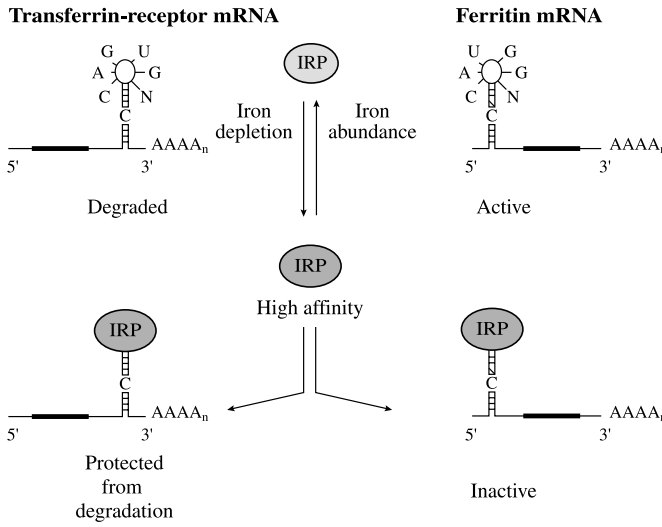

FIGURE 18. The role of iron in the regulation of protein synthesis in the liver. A stem-loop structure is located at the 3′ end of the transferrin-receptor mRNA. Additional structures located at the 3′ end include the iron-response element, which binds an iron regulatory protein (IRP) when the cell is depleted of iron. The binding of the IRP to the 3′ end of the transferrin-receptor mRNA protects the mRNA from degradation and results in an increase in the level of the transferrin-receptor mRNA and a corresponding increase in the level of the transferrin-receptor protein. At the 5′ end of the ferritin mRNA molecule is a stem-loop structure that binds IRP when iron is depleted in the cell. Binding of the IRP at the 5′ end of the ferritin mRNA inhibits the translation of this mRNA and results in a decreased level of ferritin protein. When iron levels are abundant, the ferritin mRNA no longer binds IRP and actively translates ferritin protein. At the same time, iron abundance inhibits IRP from binding to the 3′ end of the transferrin-receptor mRNA, and the mRNA is degraded. This results in a reduction of the level of transferrin-receptor protein.

tissues. Furthermore, tissue-specific availability of certain transcription factors also contributes to this process. For example, the α-amylase gene contains two promoter sites that control the expression of this gene in a tissue-specific manner. Salivary gland cells have very high levels of α-amylase, whereas hepatocytes have very low levels. The relative difference in amounts of α-amylase is controlled at the transcriptional level. In salivary gland cells, the first promoter site, located just 5′ to the first exon of the α-amylase gene, determines the start of transcription as well as the rate of gene transcription. This is a strong promoter, because it has the

ability to transcribe the gene at a high transcriptional rate. By contrast, in hepatocytes the available transcription factors do not recognize the first strong promoter of the gene, and divert the RNA polymerase II to the second and weaker promoter located just 5′ to the second exon of the α-amylase gene. This results in the same α-amylase protein being transcribed, albeit at lower levels. When the pre-mRNA is later spliced to form the mature mRNA, this 5′ untranslated exon in each cell type is spliced to the first exon containing the amino acid sequence information. The final result is that the mature mRNA in hepatocytes differs from that which is found in salivary gland cells with respect to the 5′ untranslated sequence only (the amino acid coding regions are identical).

3.5.7 ALTERNATIVE POLYADENYLATION SITES[43]

The differential production of the membrane form and the secreted form of immunoglobulin M (IgM) depends on the structure of the heavy-chain component of the antibody molecule. The membrane form of IgM contains a heavy chain with a carboxy terminal amino acid sequence rich in hydrophobic amino acids that facilitate its interaction and binding to the cell membrane. In contrast, the secreted form of the antibody contains a heavy chain devoid of this carboxy terminal amino acid sequence, and is unable to bind to the plasma membrane.

By using alternative polyadenylation signals within the gene, the precise type of heavy-chain mRNA is determined during B-cell development (Figure 19). When the mRNA encoding the membrane form of the heavy chain is produced, a polyadenylation signal present at the distal 3′ end of the message determines the site of cleavage and polyadenylation of the mRNA. After polyadenylation of the mRNA occurs, splicing of all of the exons follows, including the 3′ exon, which codes for the hydrophobic amino acid sequence located at the carboxy terminal end of the membrane-bound form of the heavy chain. This yields the mature mRNA. Translation of this particular mRNA produces a form of the heavy chain that has a hydrophobic tail and is found in the membrane-bound form of IgM. In contrast, in the cells in which the secreted form of the IgM molecule is produced, a second polyadenylation signal, which is recognized by the cell-specific polyadenylation system of mature B-cells, is located further upstream of the distal 3′ polyadenylation signal. In these cells, the cleavage and polyadenylation of the mRNA occurs at this second site, and the exons located 3′ to this site are no longer present in the mRNA produced. Following polyadenylation, the remaining exons are spliced together to yield an mRNA that encodes a heavy chain that is lacking the hydrophobic tail. Translation of this mRNA produces the heavy chain found in the secreted form of IgM.

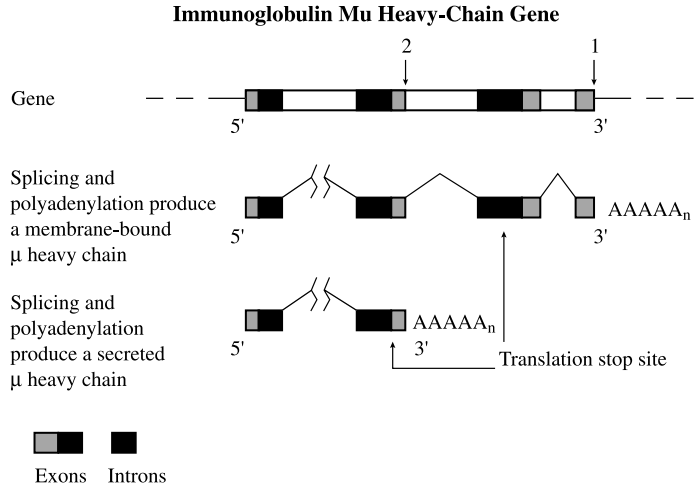

FIGURE 19. Alternative polyadenylation in the immunoglogulin μ heavy chain gene. Only some of the exons of the μ heavy chain gene are represented. In cells in which the membrane-bound form of the antibody is produced, a polyadenylation signal present at the distal 3′ end of the mRNA (indicated by arrow 1) determines the site of cleavage and polyadenylation of the mRNA. During the splicing process of this mRNA, all the exons, including the 3′ exon coding for the hydrophobic amino acids found at the carboxy terminal end of the membrane-bound form of the μ heavy chain, are spliced together to produce the mature mRNA. In cells in which the secreted form of the antibody is produced, the upstream polyadenylation signal (arrow 2) is recognized and determines the site of cleavage and polyadenylation of the mRNA found in mature B cells. The mRNA produced after splicing is lacking the exons located 3′ to this polyadenylation signal. This results in the production of a μ heavy chain devoid of a hydrophobic tail, which is thus secreted.

3.6 DNA Methylation and the Control of Transcription[42]

Not only is methylation important in DNA synthesis and repair, but it also represents another general mechanism associated with the control of eukaryotic gene transcription. Cytosine residues in eukaryotic DNA are modified by the addition of methyl groups. DNA is methylated specifically at the Cs that precede Gs (*CpG dinucleotides*). Methylation is correlated with reduced transcriptional activity of several genes. Distinct patterns of methylation are seen in different tissues. The DNA of inactive genes is more heavily methylated, as compared to the DNA of genes that are actively transcribed. Moreover, some genes contain high frequencies of CpG dinucleotides in the region of their promoters. Transcription of these genes is repressed by methylation through the action of a protein that binds specifically to methylated DNA and inhibits transcription.

4. Protein Synthesis and Post-translational Processing in Eukaryotic Cells

4.1 Translation of mRNA[58-63]

The tRNAs serve as carriers and adapters for the alignment of each of the 20 amino acids with their corresponding codons on the mRNA template. tRNAs consist of 70 to 80 nucleotides, with a characteristic "clover leaf" configuration that results from complementary base pairing between regions of the molecule. The tRNAs possess unique identifying sequences that allow the correct amino acid to be attached and aligned with the appropriate codon in the mRNA. All tRNAs have the sequence CCA at the 3′ end where free amino acids covalently attach to the ribose of the terminal adenosine residue. Recognition of the mRNA template occurs through interaction with an *anticodon loop*, located at the other end of the tRNA, which binds to the appropriate codon through complementary base pairing. The attachment of amino acids to specific tRNAs is mediated by *aminoacyl tRNA synthetases*. The three-base sequence on the anticodon loop is complementary to a specific codon found in the mRNA. For example, if the codon in the mRNA is GGC, it is recognized by the anticodon of the tRNA as CCG.

While there are 61 codons specifying amino acids, there are fewer than 61 tRNA molecules. Thus, some of the tRNA molecules are able to recognize more than one codon; this phenomenon is called *wobble*. Wobble effects are found with the third base of the codon.

4.2 The Steps in Protein Synthesis

Particles consisting of RNA and protein, known as *ribosomes*, are located in the cytoplasm and serve as the site of protein synthesis. The principal components of the protein synthesis machinery include mRNA, tRNAs, amino acids and ribosomes.

Each ribosome is composed of two subunits, the 40S (or small subunit) and 60S (or large subunit). The size of the entire particle is 80S. The 40S subunit is made up of the 18S rRNA and 30 different proteins. The 60S subunit is made up of the 5S, the 5.8S and the 28S rRNA as well as 50 different protein species. Ribosomal proteins are imported to the nucleolus from the cytoplasm, and begin to assemble on pre-rRNA prior to its cleavage. As the pre-rRNA is processed, additional ribosomal proteins and the 5S rRNA assemble to form pre-ribosomal particles. The pre-ribosomal particles are exported from the nucleus to the cytoplasm, yielding the 40S and 60S ribosomal subunits.

The ribosome physically moves down the mRNA in the 5′-to-3′ direction, with the sequential addition of amino acids from tRNAs to form the nascent polypeptide. Amino acids are attached to tRNA by a process called

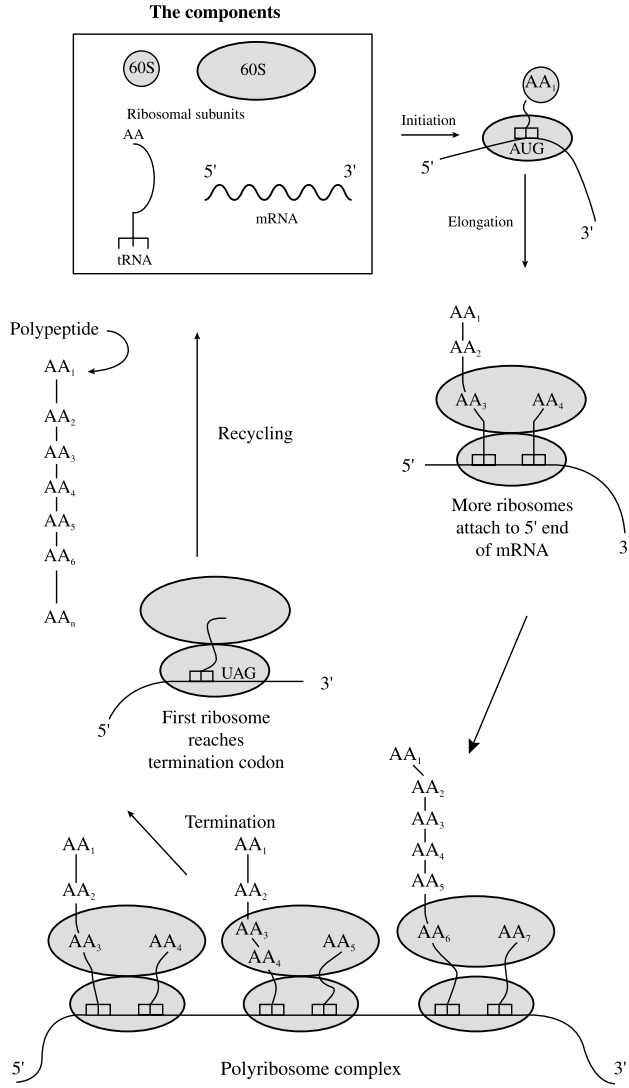

FIGURE 20. Overview of translation. Translation involves three stages. Initiation occurs when the ribosomal subunits and charged tRNA associate with an mRNA molecule to form the initiation complex. Elongation occurs when additional tRNA molecules bring additional amino acids to the mRNA, where they are added in a stepwise fashion to the growing polypeptide chain. Termination occurs when a stop codon appears in the mRNA, and the completed polypeptide is released from the ribosome.

charging, which is mediated by *aminoacyl tRNA synthetases*. For each of the 20 amino acids, there are 20 different aminoacyl tRNA synthetases. When the protein is completed, it is released along with the ribosome and tRNA molecules, which are free to begin the cycle again.

Protein synthesis comprises three specific steps: *initiation, elongation* and *termination.* Each of these steps involves specific proteins, and the energy for this process is derived from either ATP or GTP. These steps are illustrated in Figure 20.

4.2.1 INITIATION OF TRANSLATION

In eukaryotes, the initiation of protein synthesis involves approximately 10 different proteins (Figure 21). The initiation factors eIF3 and eIF1A bind to the 40S ribosomal subunit. The initiation factor eIF2 binds to GTP to form a complex that binds a tRNA charged with initiator methionine. The 5′ Cap of the mRNA is recognized by eIF4, which brings the mRNA to the ribosome. The eIF2-Met-tRNA-GTP complex subsequently interacts with the 40S subunit at the 5′ end of the mRNA. After binding to the 5′ end of the message, the 40S subunit with the eIF2-Met-tRNA-GTP complex moves down the mRNA. This process is known as *scanning.* Scanning continues until the complex reaches the first AUG (i.e., the initiator codon) on the mRNA. Then, the 60S ribosomal subunit binds to the complex to form the final ribosomal structure. This process requires GTP as an energy source. The formation of this final structure signals the completion of the initiation step. eIF2 and GDP are released from the complex and are able to reinitiate the cycle. When the initiator codon (AUG) is located, eIF5 triggers the hydrolysis of GTP bound to eIF2, followed by the release of eIF2 (complexed to GDP) and other initiation factors. The 60S ribosomal subunit then joins the 40S complex to form the 80S initiation complex.

4.2.2 PEPTIDE ELONGATION

The various steps involved in the elongation phase of protein synthesis are illustrated in Figure 22. The ribosome has three sites for tRNA binding designated the P (peptidyl), A (aminoacyl) and E (exit) sites. The initiator Met-tRNA is bound at the P site. The first step in elongation is the binding of the next aminoacyl tRNA to the A site by pairing with the second codon on the mRNA. The aminoacyl tRNA is escorted to the ribosome by an *elongation factor* (eEF1α), which is complexed to GTP. The GTP is hydrolyzed to GDP after the correct aminoacyl tRNA is inserted into the A site of the ribosome, and the elongation factor bound to GDP is released.

Once the eEF1α has left the ribosome, the peptide bond is formed between the initiator met-tRNA at the P site and the second aminoacyl tRNA at the

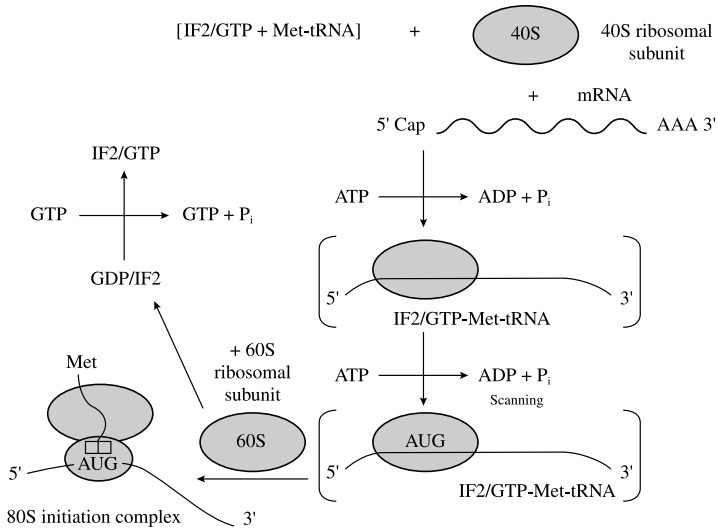

FIGURE 21. Initiation of protein synthesis. The initiation factor 2 (IF2) complex with GTP binds to a tRNA charged with methionine (Met). This complex interacts with the small 40S ribosomal subunit at the 5′ end of an mRNA molecule. After binding to the 5′ end of the mRNA, the 40S subunit scans the mRNA until it reaches the first AUG codon. At this point, the 60S ribosomal subunit binds to the complex to form the final initiation complex.

A site. This reaction is catalyzed by the large ribosomal subunit. The result is the transfer of methionine to the aminoacyl tRNA at the A site of the ribosome, forming a peptidyl tRNA at this position and leaving the uncharged initiator tRNA at the P site. The next step in elongation is translocation, which requires the elongation factor eEF2, and is again coupled to the hydrolysis of GTP. During translocation, the ribosome moves 3 nucleotides along the mRNA, positioning the next codon in an empty A site. This step translocates the peptidyl tRNA from the A site to the P site, and the uncharged tRNA from the P site to the E site. The ribosome is then left with a peptidyl tRNA at the P site, and an empty A site. The binding of a new aminoacyl tRNA to the A site then causes the release of the uncharged tRNA from the E site. This leaves the ribosome ready for the next amino acid in the growing polypeptide chain.

4.2.3 TERMINATION OF TRANSLATION

Elongation of the polypeptide chain continues until a termination codon (stop or terminator codon) is translocated into the A site of the ribosome. The *release factor (eRF)* recognizes all three termination codons. The eRF binds

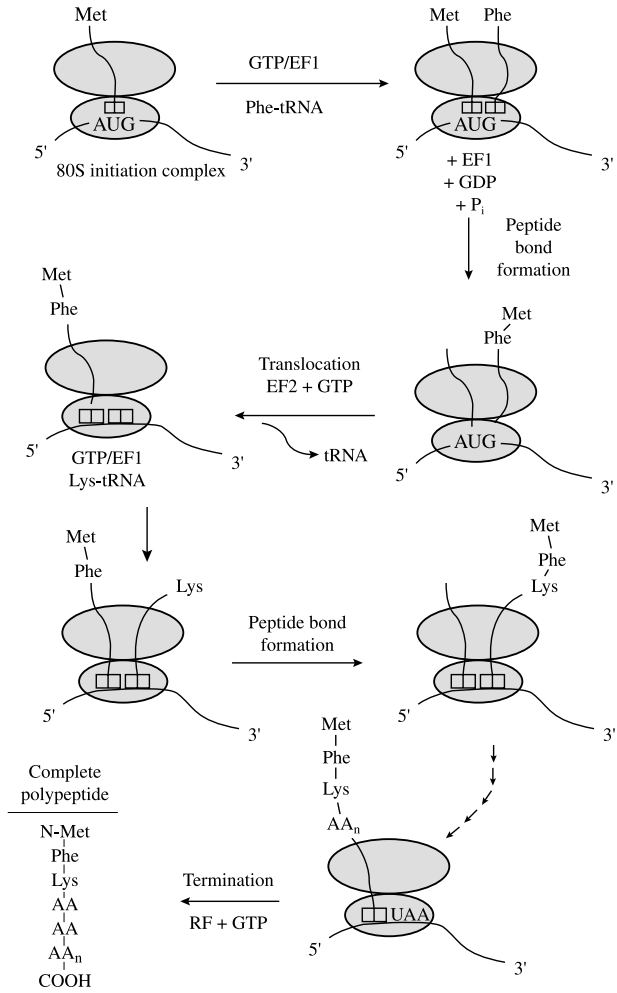

FIGURE 22. Elongation and termination of protein synthesis. Protein synthesis is initiated by the binding of methionine (Met)-tRNA to the AUG codon in mRNA bound to the ribosome. A second aminoacyl-tRNA interacts with the elongation factor 1 (EF1) and GTP followed by the binding of this complex to the second codon on the mRNA. In the presence of peptidyl transferase, a peptide bond is formed between methionine and phenylalanine (Phe). Subsequently, in the presence of GTP, EF2, and an enzyme known as translocase, the ribosome translocates one codon along the tRNA in the 5′-to-3′ direction. This is followed by the release of the uncharged tRNA and the exposure of a new codon. The next aminoacyl-tRNA binds, and the cycle is repeated until a termination codon is encountered. In the presence of release factor (RF), the completed peptide is released from the ribosome.

to a termination codon at the A site and stimulates the hydrolysis of the bond between the tRNA and the polypeptide chain at the P site. This results in the release of the completed polypeptide from the ribosome.

The mRNAs are usually translated by a series of ribosomes, spaced at intervals of about 100 to 200 nucleotides. The group of ribosomes bound to an mRNA molecule is called a *polyribosome* (*polysome*), and each ribosome within the group functions independently to synthesize a separate polypeptide chain.

4.3 Regulation of Translation[58-63]

Although transcription is the primary level at which gene expression is controlled, the translation of mRNA represents an additional regulatory control point in eukaryotic cells. One of the best examples of translational regulation in eukaryotic cells is the cellular mechanisms associated with the regulation of ferritin synthesis. The translation of ferritin mRNA is regulated by the supply of iron (Figure 18). More ferritin is synthesized when iron is abundant, and this regulation is mediated by a protein that binds to the iron-responsive element (IRE) in the 5′ untranslated region of ferritin mRNA. In the presence of iron, the repressor no longer binds to the IRE, and ferritin translation is able to proceed.

The regulation of ferritin translation by iron is similar to the regulation of the stability of transferrin receptor mRNA, which is regulated by protein binding to an IRE in the mRNA's 3′ untranslated region. The same protein binds to the IREs of both the ferritin and the transferrin receptor mRNAs. However, the consequences of the binding of this protein to the two IREs are quite different (Figure 18). The protein bound to the transferrin receptor IRE protects the mRNA from degradation, rather than inhibiting its translation. These distinct effects probably result from the different locations of the IRE in the two mRNAs. Thus, binding of the same regulatory protein to different sites on mRNA molecules can have distinct effects on gene expression, in one case inhibiting translation, and in the other case, stabilizing the mRNA to increase protein synthesis. In the case of the ferritin mRNA, the IRE blocks translation by interfering with 5′ Cap recognition and binding of the 40S ribosomal subunit. This protein binding to the same sequence in the 3′ UTR of transferrin-receptor mRNA protects the mRNA from nuclease degradation and prolongs its half-life.

4.4 Post-Translational Processing of Proteins[64-72, 79, 82-84, 87]

Newly synthesized polypeptides are subsequently folded into three-dimensional structures. In many instances, multiple polypeptide chains are assembled into a functional complex. Many proteins undergo further modifications,

which include the covalent attachment of carbohydrates and lipids that are critical for determining the function and correct localization of proteins within the cell.

Earlier studies suggested that protein folding is a self-assembly process determined primarily by its amino acid sequence. However, more recent studies have shown that the proper folding of proteins is mediated by the activities of a group of proteins called *molecular chaperones*. Chaperones catalyze protein folding by assisting the self-assembly process: the folded conformation of a protein is determined solely by its amino acid sequence. Chaperones bind to and stabilize partially folded polypeptides. In the absence of chaperones, unfolded or incompletely folded polypeptides are unstable within the cell and aggregate into insoluble complexes. Some chaperones bind to nascent polypeptides that are still being translated on ribosomes. This prevents incorrect folding of the amino terminal region of the polypeptide before the synthesis of the chain is terminated. This interaction is important for proteins in which the carboxy terminal region is required for correct folding of the amino terminus. Other classes of chaperones stabilize unfolded polypeptide chains during their intracellular transport to organelles such as the mitochondria. Finally, chaperones are also involved in the assembly of proteins that consist of multiple polypeptide chains.

Many of the molecular chaperones were originally identified as *heat-shock proteins (Hsp)*, a group of proteins that are expressed in cells that have been subjected to increased temperature or other forms of environmental stress. The heat-shock proteins appear to stabilize and facilitate the refolding of proteins that have been partially denatured as a result of exposure to increased temperature. However, many heat-shock proteins are expressed under normal growth conditions. They function as molecular chaperones required for polypeptide folding and transport under normal conditions, as well as under conditions of environmental stress. Members of the Hsp-70 family stabilize unfolded polypeptide chains during translation as well as during intracellular transport to subcellular compartments such as the endoplasmic reticulum and mitochondria. These proteins bind to short segments of seven or eight amino acid residues of unfolded polypeptides and maintain the polypeptide chain in an unfolded conformation, thereby preventing aggregation. Proteins in the Hsp-60 family facilitate the folding of proteins into their native conformations. In several instances, members of the Hsp-70 and Hsp-60 families act together in a sequential fashion, and may therefore represent a general pathway of protein folding.

In addition to molecular chaperones, cells contain enzymes that catalyze protein folding by breaking and reforming covalent bonds. The formation of disulfide bonds between cysteine residues is an important step in the

stabilization of the folded structures of many protein species. In this regard, *protein disulfide isomerase (PDI)* catalyzes the breakage and reunion of these bonds. Disulfide bonds are usually restricted to secreted proteins and some membrane proteins. In eukaryotic cells, disulfide bonds form in the endoplasmic reticulum where the activity of PDI is correlated with the level of protein secretion. Another example of an enzyme that plays a pivotal role in protein folding is *peptidyl-prolyl-isomerase,* which catalyzes the isomerization of peptide bonds that involve proline residues.

Proteolysis is a critical step in the maturation of many proteins. A simple example of proteolysis is the removal of the initiator methionine residue from the amino terminus of many polypeptides after the growing polypeptide chain leaves the ribosome. As well, proteolytic modification of the amino terminus plays a central role in the translocation of many proteins across the membranes. This includes the translocation of secreted proteins as well as proteins destined for targeting to the plasma membrane, lysosomes and mitochondria of eukaryotic cells.

Active enzymes and hormones are formed via proteolytic processing of larger precursors. For example, insulin is synthesized as a large precursor polypeptide (pre-proinsulin) containing an amino terminal sequence that targets the polypeptide chain to the endoplasmic reticulum (ER). Proinsulin is formed through the removal of the signal sequence during transfer to the ER. Proinsulin is subsequently converted to insulin, which consists of two chains held together by disulfide bonds, by proteolytic removal of an internal peptide.

The levels of proteins within cells reflect a balance between synthesis and degradation. The differential rates of protein degradation represent an important aspect of cell regulation. Rapidly degraded proteins function primarily as regulatory molecules, such as transcription factors. The rapid turnover of these proteins is necessary to allow their levels to respond quickly to external stimuli. Two major pathways mediate protein degradation: the ubiquitin-proteasome pathway and lysosomal proteolysis. The major pathway for selective protein degradation employs *ubiquitin* as a marker that targets cytoplasmic and nuclear proteins for rapid degradation. Ubiquitin is a 76-amino-acid polypeptide that attaches to the amino group of lysine residues. The ubiquinated proteins are recognized and degraded by a multi-subunit protease complex called *proteasome*. Ubiquitin is subsequently released and recycled. The other major pathway for protein degradation involves the transport of proteins to lysosomes, where they are taken up and degraded by proteases.

4.5 The Cellular Compartmentalization of Protein Sorting and Intracellular Transport[70, 77-79, 95]

Eukaryotic cells are distinct from prokaryotic cells by the presence of

membrane-delimited compartments wherein specific cellular activities occur. The sorting and targeting of proteins to their appropriate destinations such as the plasma membrane, the endoplasmic reticulum or the Golgi complex are key features in the maintenance of these specific cellular activities.

Proteins destined for the endoplasmic reticulum, the Golgi appartatus, lysomes, the plasma membrane, and cellular secretion are synthesized on ribosomes that are bound to the ER membrane. Nascent polypeptide chains are transported from the cytoplasm into the ER, where protein folding and further processing occur prior to transport to the Golgi apparatus via ER-derived vesicles. In the Golgi appartatus, proteins are further processed and sorted for transport to the plasma membrane or to lysosomes, or export from the cell as secretory proteins. The various cellular compartments associated with protein sorting and transport are depicted in Figure 23.

Proteins synthesized on free ribosomes either remain in the cytoplasm or are transported to the nucleus, mitochondria or peroxisomes. Proteins destined for transport to the nucleus are responsible for important aspects of genome structure and function. These include histones, DNA and RNA polymerases, transcription factors and splicing factors. These proteins are targeted to the nucleus by specific *nuclear localization signals* that direct their transport through the *nuclear pore complex*. The first nuclear localization signal characterized was that of the SV40 viral T antigen. The amino acid sequence Pro-Lys-Lys-Lys-Arg-Lys-Val is necessary for the nuclear transport of the T antigen and other types of cytoplasmic proteins. Proteins are transported through the nuclear pore complex, a process mediated by the action of a nuclear receptor called *importin*.

4.5.1 PROTEIN TARGETING TO THE ENDOPLASMIC RETICULUM[70, 77-79, 85-87, 89, 90, 92, 95]

Ribosomes that participate in the synthesis of proteins that are ultimately destined for secretion are targeted to the ER. This targeting is directed by the amino acid sequence of the newly synthesized polypeptide chain, rather than by the intrinsic properties of the ribosome. A *signal sequence* spans about 20 amino acids, including a stretch of hydrophobic residues, and is located at the amino terminus of the polypeptide chain. As they emerge from the ribosome, signal sequences are recognized and bound by a *signal-recognition particle (SRP),* which consists of six polypeptides and a small cytoplasmic RNA. The binding of the SRP inhibits translation and targets the complex (polypeptide chain, SRP, ribosome) to the rough ER. This is mediated by binding to the SRP receptor on the ER membrane. Binding to the receptor releases the SRP from the ribosome and the signal sequence of the polypeptide chain. The ribosome subsequently binds to the *protein translocation complex* of the

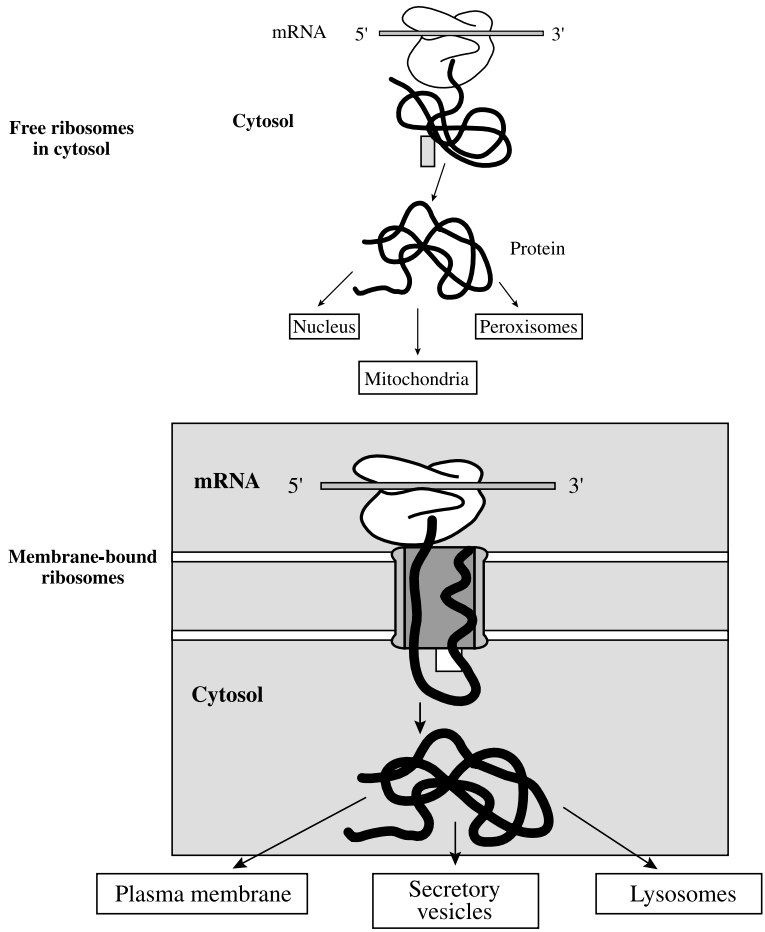

FIGURE 23. Overview of protein sorting. Proteins synthesized on free ribosomes either remain in the cytoplasm or are transported to the nucleus, mitochondria, chloroplasts or peroxisomes. By contrast, proteins synthesized on membrane-bound ribosomes are subsequently translocated into the ER while their translation is in progress. They may be either retained within the ER or transported to the Golgi apparatus and, from there, to lysosomes or to the plasma membrane, or secreted outside the cell within secretory vesicles.

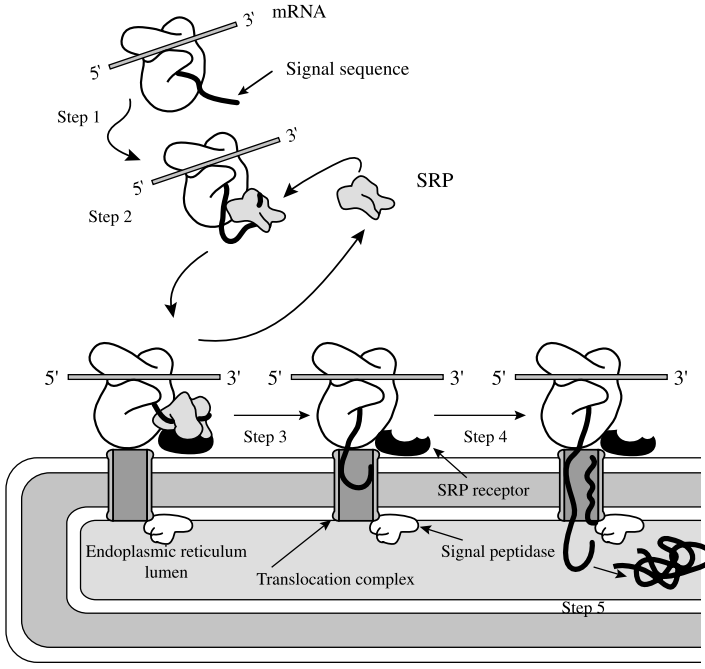

FIGURE 24. The targeting of secretory proteins to the ER. Step 1: As the signal sequence emerges from the ribosome, it is subsequently recognized and bound by the signal-recognition particle (SRP). Step 2: The SRP escorts the complex to the ER membrane, where it binds to the SRP receptor. Step 3: The SRP is subsequently released, the ribosome binds to a membrane translocation complex, and the signal sequence is inserted into a membrane channel. Step 4: Translation resumes, and the growing peptide chain is translocated across the ER membrane. Step 5: A signal peptidase catalyzes the cleavage of the signal sequence, and this releases the polypeptide into the ER lumen.

ER membrane, and the signal sequence is inserted into an ER membrane channel. Translation resumes, and the growing polypeptide chain is translocated across the membrane into the ER lumen. The signal sequence is cleaved by the action of *signal peptidase*, and the polypeptide is liberated into the ER lumen. The sec-61 complex comprises three membrane-spanning proteins and is the principal component of the ER protein-conducting channel in mammalian cells. The targeting of secretory proteins to the ER is illustrated in Figure 24.

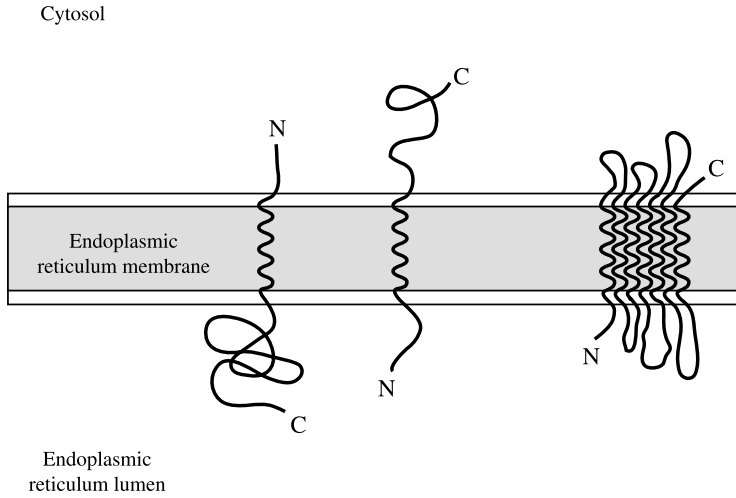

FIGURE 25. The possible orientations of membrane proteins. Integral membrane proteins span the membrane via α-helical regions of 20 to 25 hydrophobic amino acids, which can be inserted in a variety of orientations. The proteins at left and center each span the membrane only once, but they differ in whether the amino (N) or carboxy (C) terminus is on the cytoplasmic side. On the right is an example of a protein that has multiple membrane-spanning regions.

Proteins destined for incorporation into the plasma membrane, ER membranes, Golgi or lysosomes are inserted initially into the ER membrane, instead of being liberated into the ER lumen. These proteins then proceed to their final destination along the secretory pathway: ER → Golgi → plasma membrane or lysosomes. The proteins are transported along this pathway as membrane constituents, which differentiates the process from that of secretory proteins. These integral membrane proteins are embedded in the plasma membrane by hydrophobic regions that span the phospholipid bilayer of the membrane. The orientation of proteins inserted into the ER, Golgi, lysosomal and plasma membranes is established as the polypeptide chain is inserted into the ER. The ER lumen is topologically equivalent to the exterior of the cell membrane, such that the domains of plasma membrane proteins that are exposed at the level of the cell surface correspond to the regions of polypeptide chains that are translocated into the ER.

A variety of orientations of membrane proteins are found in eukaryotic cells. Transmembrane proteins are observed with either the carboxy or amino termini exposed to the cytosol (Figure 25). Other proteins have multiple membrane-spanning regions called α-*helical regions,* which consist of 20 to 25 hydrophobic amino acids. Some integral membrane proteins span the

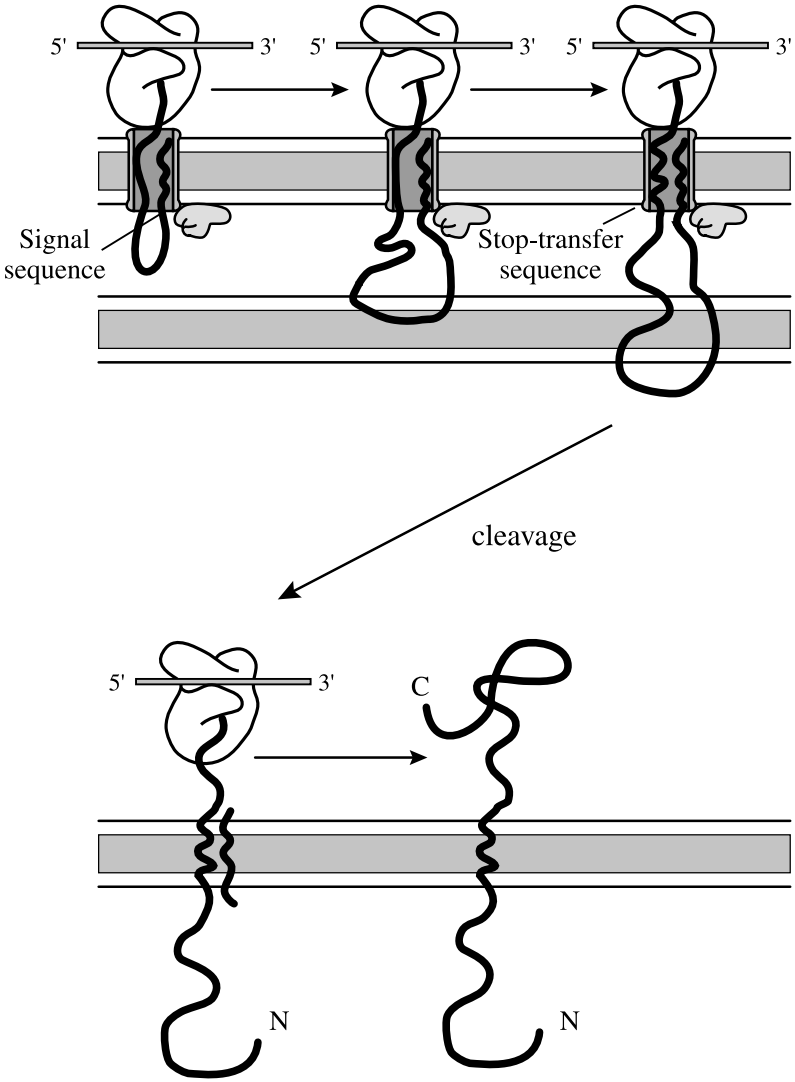

FIGURE 26. The insertion of a membrane protein with a cleavable signal sequence and a single stop-transfer sequence. The signal sequence is cleaved as the poplypeptide chain is exposed within the ER lumen. However, translocation of the polypeptide chain across the membrane is halted by a stop-transfer sequence that anchors the protein to the membrane. The ribosome is released from the membrane, and continued translation results in a membrane-spanning protein with its C terminus on the cytoplasmic side.

plasma membrane only once, while others have multiple membrane-spanning regions. As well, some proteins are oriented in the membrane with their amino terminus on the cytoplasmic side, and others have their carboxy terminus exposed to the cytoplasm. Two additional features of membrane proteins have been discovered, which play a key role in determining the orientation of membrane proteins: the *stop-transfer sequence* and the *internal signal sequence*. The consequences of these sequences in determining membrane protein orientation are illustrated in Figures 26–28.

4.5.2 PROTEIN PROCESSING IN THE ENDOPLASMIC RETICULUM[66-68, 72, 87, 91]

A variety of modifications to polypeptides at the level of ER include folding and assembly, as well as covalent modifications.

The proteolytic cleavage of the internal signal sequence takes place as the polypeptide chain is translocated across the ER membrane. The translocation occurs while translation is still in progress, and molecular chaperones facilitate the folding of the polypeptide chains. The *binding protein (BiP)* is a member of the Hsp-70 family of chaperones that mediate protein folding and the assembly of multi-subunit proteins within the lumen of the ER (Figure 29). The correctly assembled proteins are released from BiP and are available for export to the Golgi apparatus. By contrast, abnormally folded or improperly assembled proteins remain bound to BiP and are retained within the ER, where they are subsequently degraded. Disulfide bond formation represents an important aspect of protein folding and assembly within the ER. This process is facilitated by the enzyme *disulfide isomerase*, which is located within the lumen of the ER.

Some proteins are anchored within the plasma membrane by *glycosylphosphatidylinositol (GPI)* anchors, which are assembled in the ER membrane. The GPI anchors are added immediately after completion of protein synthesis to the carboxy terminus of some proteins, which are subsequently transported to the cell surface via the secretory pathway. Their orientation within the ER dictates that GPI anchor proteins reside outside the cell.

4.5.3 TRANSPORT OF PROTEINS FROM THE ENDOPLASMIC RETICULUM[70, 71, 78, 79, 88, 94, 95]

Proteins travel along the secretory pathway in transport vesicles derived from the ER. These proteins subsequently fuse with the membrane of the Golgi apparatus. The subsequent steps in the secretory pathway involve vesicular transport between the different Golgi compartments, and from the Golgi to the plasma membrane or lysosomes. The Golgi apparatus consists of a series of membrane-delimited cisternae and associated vesicles. Proteins derived from

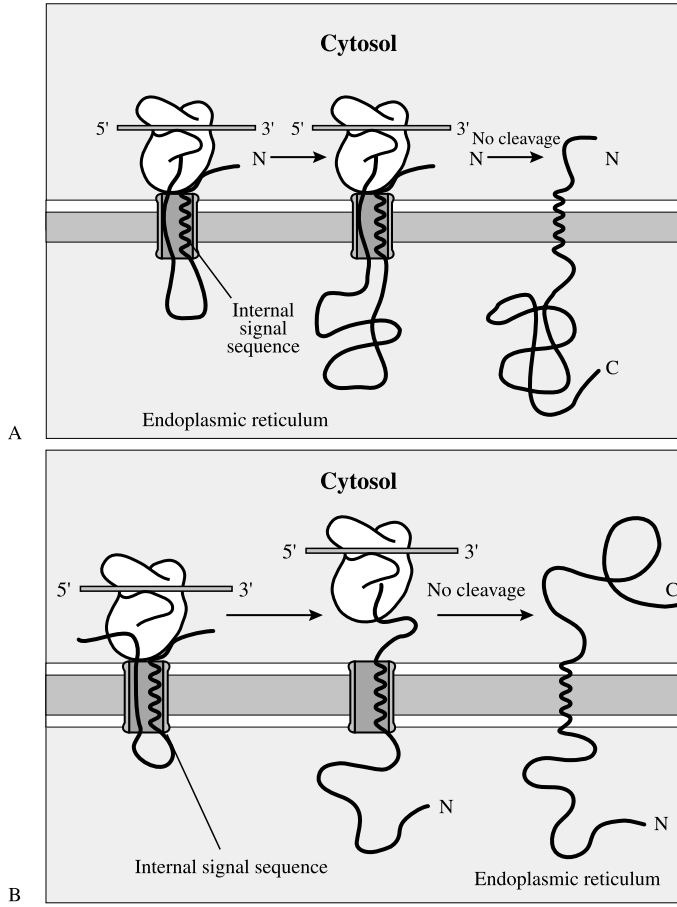

FIGURE 27. The insertion of membrane proteins with an internal non-cleavable signal sequence. Internal non-cleavable signal sequences result in the insertion of polypeptide chains in either orientation in the ER membrane.

A. The signal sequence directs insertion of a polypeptide such that its N terminus is exposed on the cytoplasmic side. The remainder of the polypeptide is translocated into the ER as translation proceeds. The signal sequence is not cleaved, so it acts as a membrane-spanning sequence that anchors the protein to the membrane with its C terminus within the ER lumen.

B. Other internal signal sequences are oriented to direct the transfer of the N terminal portion of the polypeptide across the membrane. Continued translation results in a protein that spans the ER membrane with its N terminus in the lumen and its C terminus in the cytoplasm. This orientation is the same as that resulting from insertion of a protein that contains a cleavable signal sequence followed by a stop-transfer sequence.

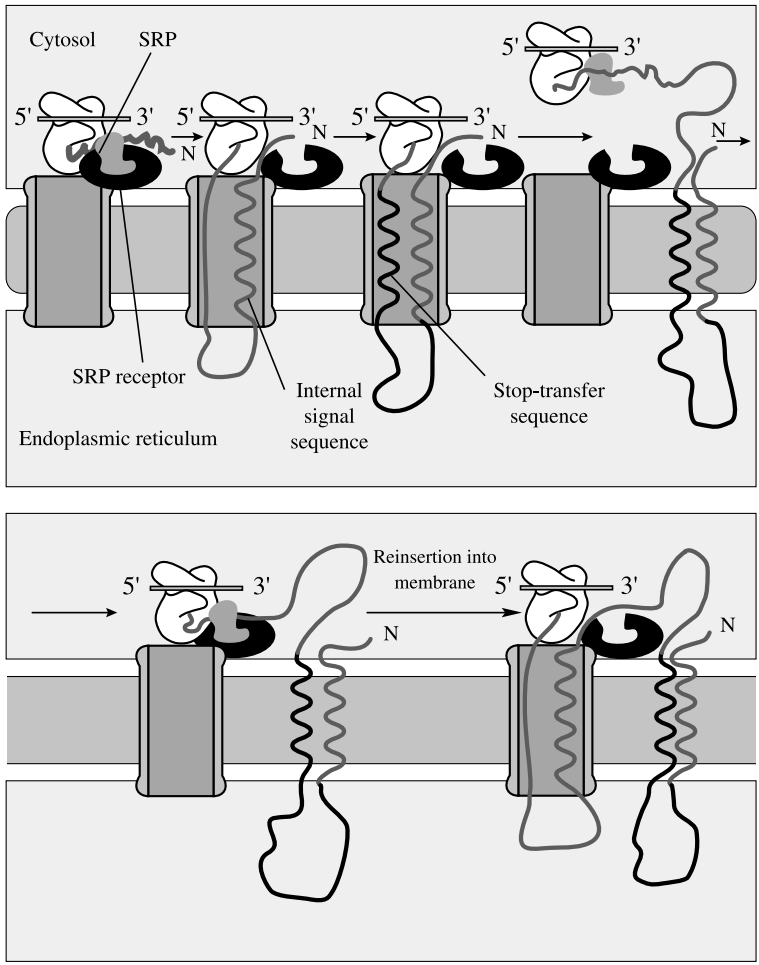

FIGURE 28. Insertion of a protein that spans the membrane multiple times. In this example, the internal signal sequence results in insertion of the polypeptide chain with its N terminus on the cytoplasmic side of the membrane. A stop-transfer sequence then causes the polypeptide chain to form a loop within the ER lumen, and translation continues in the cytoplasm. A second internal signal sequence triggers reinsertion of a polypeptide chain into the ER membrane, forming a loop within the cytoplasm. This process can occur many times and results in the insertion of proteins with multiple membrane-spanning regions.

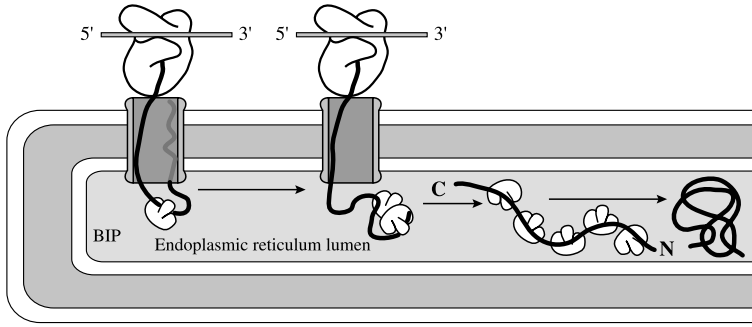

FIGURE 29. Protein folding in the ER. The molecular chaperone BiP binds to polypeptide chains as they cross the ER membrane and facilitates protein folding and assembly within the ER lumen.

the ER enter the Golgi at the cis-face and exit the Golgi from its trans-face. Proteins marked for residence within the ER are recognized by the Golgi and are returned to the ER. Other proteins are carried by transport vesicles to the trans-Golgi network where the final stages of protein modification are completed prior to their being targeted to lysosomes and to the plasma membrane.

Most proteins travel from the ER to the Golgi. However, some proteins particular to the functioning of the ER must be retained within that organelle (e.g., BiP, signal peptidase, protein disulfide isomerase). Targeting sequences specifically designate proteins destined for retention in the ER or transport to the Golgi (Figure 30). The proteins that are retained in the ER lumen contain the targeting sequence KDEL (single-letter amino acid code; Lys-Asp-Glu-Leu) at their carboxy terminus. The retention of certain transmembrane proteins within the ER is dictated by the carboxy terminal sequence KKXX. Soluble ER proteins are packaged into vesicles and are transported into the Golgi where they are subsequently retrieved and returned to the ER via a recycling pathway. Thus, proteins bearing the KDEL and KKXX sequences bind to specific recycling receptors in the Golgi membrane and are selectively transported back to the ER (Figure 31). Proteins destined for transport from the ER are selectively packaged into transport vesicles targeted to the Golgi apparatus. Thus, protein export from the ER is controlled not only by retention/retrieval signals, but also by targeting signals that mediate the selective transport to the Golgi.

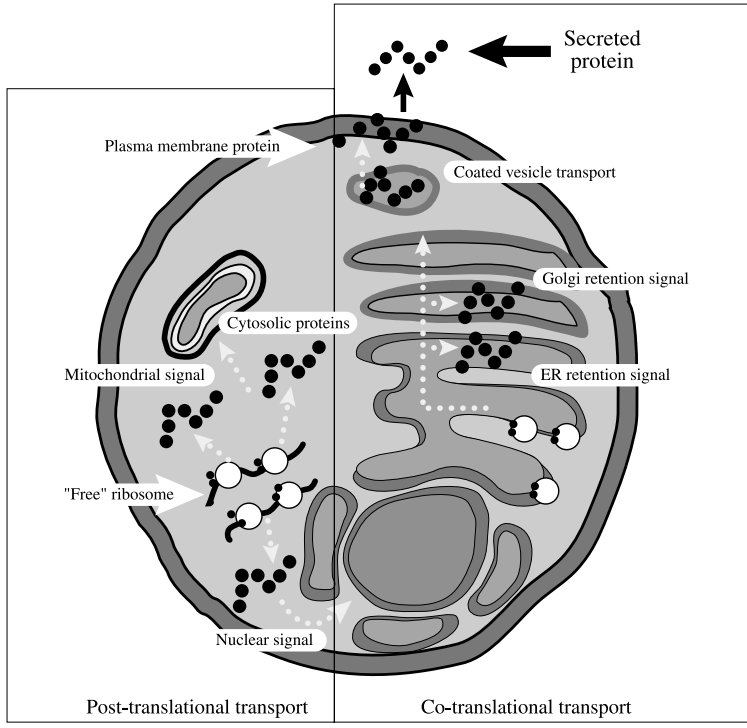

FIGURE 30. Proteins that are localized post-translationally are released into the cytoplasm after synthesis on "free ribosomes." Some have signals for targeting to the nucleus or mitochondria. Proteins that are localized co-translationally associate with the ER membrane during synthesis so that their ribosomes are "membrane-bound." The proteins pass into the ER, travel to the Golgi and then to the plasma membrane, unless they possess the signals that cause retention in one of the compartments along the pathway. They may also be directed to other organelles, such as lysosomes. Transport along this pathway occurs by way of secretory vesicles.

4.5.4 PROTEIN GLYCOSYLATION[71, 78, 85, 94]

Protein glycosylation takes place on specific asparagine residues (N-linked glycosylation) while a translation is taking place. The oligosaccharide is synthesized on a *dolichol carrier* anchored to the ER membrane. The membrane-bound enzyme oligosaccharyl transferase transfers the oligosaccharide unit to acceptor asparagine residues in the consensus sequence (Asn)-X-Ser/Thr.

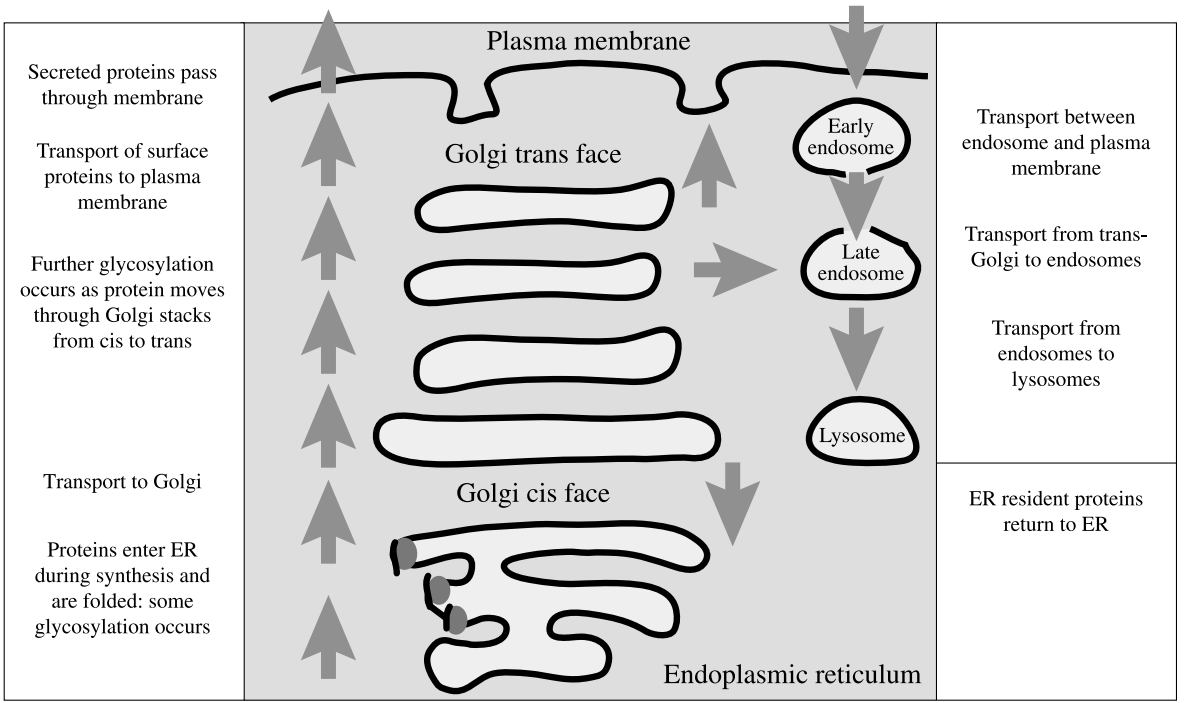

FIGURE 31. Proteins that enter the ER are transported to the Golgi and subsequently to the plasma membrane. Specific signals cause proteins to be returned from the Golgi to the ER, to be retained within the Golgi, to be retained in the plasma membrane, or to be transported to endosomes and lysosomes. Proteins may be transported between the plasma membrane and endosomes.

Thereafter, three glucose residues and one mannose residue are trimmed while the protein is still within the ER. The sequence of steps associated with protein glycosylation in the ER is illustrated in Figure 32.

The N-linked oligosaccharides are processed within the Golgi complex in an ordered sequence of reactions. The first modification is the removal of three additional mannose residues. This occurs on proteins destined for secretion or for targeting to the *plasma membrane*. This is followed by the sequential addition of an N-acetylglucosamine residue, the removal of two more mannoses, and the addition of fucose as well as two more N-acetlyglucosamines. Finally, three sialic acid residues and three galactose moieties are added; these reactions occur at the level of the trans-Golgi network. The processing of the N-linked oligosaccharide of *lysosomal* proteins differs from that of secretory and plasma membrane proteins. The proteins destined for incorporation into lysosomes are modified by mannose phosphorylation, followed by the removal of the N-acetylglucosamine group, leaving mannose 6-phosphate residues on the N-linked oligosaccharide. These phosphorylated mannose residues are specifically recognized by the mannose 6-phosphate receptor in the trans-Golgi that directs the trafficking of these proteins to lysosomes. Proteins can also be modified by the addition of carbohydrates to the side chains of serine and threonine residues within specific sequences of amino acids (O-linked glycosylation). The serine or threonine is usually linked directly to N-acetylgalactosamine to which other sugars can be subsequently added.

4.5.5 PROTEIN SORTING AND TRANSPORT FROM THE GOLGI APPARATUS[78, 79, 93, 95]

Proteins are transported from the Golgi apparatus to their ultimate destinations via the secretory pathways. This involves sorting the proteins into different kinds of transport vesicles that bud from the trans-Golgi network and deliver their contents to the appropriate cellular addresses. In the absence of specific targeting signals, proteins are delivered to plasma membranes by *bulk flow*. This transports proteins in a nonselective fashion from the ER to the Golgi and ultimately to the cell surface. This bulk flow pathway accounts for the incorporation of new proteins and lipids into the plasma membrane as well as for the continuous secretion of certain proteins from the cell.

The bulk flow pathway leads to continuous, unregulated protein secretion. In contrast, in some cell types, a distinct, regulated secretory pathway exists in which specific proteins are secreted in response to particular stimuli. Examples of regulated secretion include the release of hormones and neurotransmitters, and the release of digestive enzymes from the pancreatic acinar cells. These proteins are packaged into specialized secretory vesicles, which store their contents until specific signals direct their fusion with the plasma

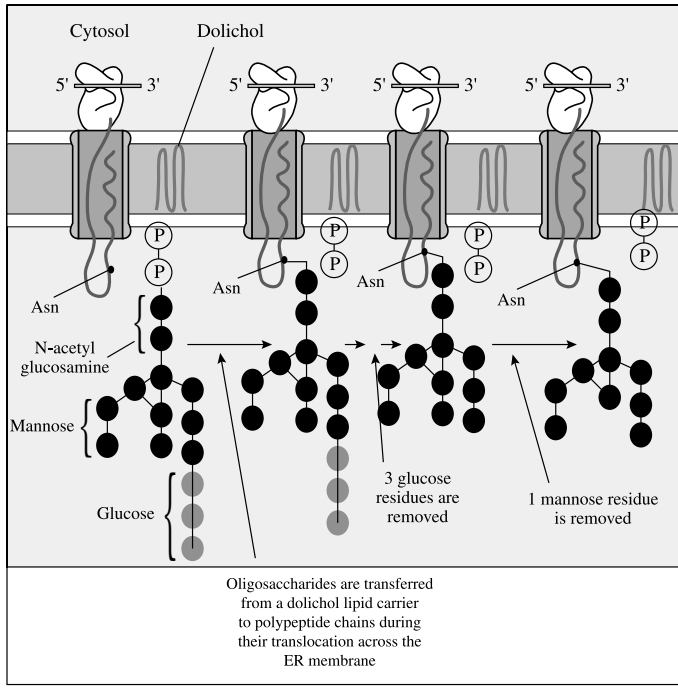

FIGURE 32. The sequential process of protein glycosylation in the ER.

membrane. The sorting of proteins into the regulated secretory pathway involves the recognition of signal patches shared by multiple proteins that enter this pathway.

Proteins that function within the Golgi complex must be retained within that organelle. Retention of Golgi membrane proteins is based on the trans-membrane domains of those particular proteins. Golgi membrane proteins have short trans-membrane α-helices of about 15 amino acids, which contribute to the retention of these proteins within the Golgi complex. As well, signals in the cytoplasmic tails of some Golgi proteins mediate the retrieval of these proteins from subsequent compartments along the secretory pathway.

The plasma membrane of polarized epithelial cells, such as the enterocyte, is divided into apical and basolateral domains. Each domain contains compartment-specific proteins related to the unique functions of each domain. In some types of epithelia, membrane proteins are sorted at the level of the trans-Golgi network for *selective transport* to the domains of the plasma membrane.

The GPI anchor is one signal that directs proteins to the apical membrane domain.

A specific receptor in the trans-Golgi network recognizes mannose 6-phosphate residues. The resulting complexes comprise the receptor plus lysosomal enzyme, and are packaged into transport vesicles destined for lysosomes.

4.5.6 VESICULAR TRANSPORT [73-76, 80, 81, 96-98]

The first step in vesicular transport is the formation of a vesicle by a process of "budding" from the membrane. The cytoplasmic surfaces of these transport vesicles are coated with proteins. Three types of coated vesicles that participate in vesicular transport have been characterized. *Clathrin-coated* vesicles are responsible for the uptake of molecules from the plasma membrane by endocytosis, as well as the transport of molecules from the trans-Golgi network to lysosomes (Figure 33). The two remaining types of coated vesicles that arise from the ER and Golgi complex are called *non-clathrin-coated* or *COP-coated* vesicles. COP-I-coated vesicles arise from the Golgi apparatus, whereas COP-II-coated vesicles bud from the ER. The COP-II-coated vesicles transport material from the ER to the Golgi, whereas COP-I-coated vesicles mediate transport between Golgi stacks, recycling from the Golgi to the ER, and possibly other transport processes.

The binding of clathrin to membranes is mediated by *adaptins.* These adaptins are responsible for the assembly of clathrin-coated vesicles at the plasma membrane and at the trans-Golgi network, as well as being responsible for selecting specific molecules to be incorporated into the vesicles.

Distinct protein complexes comprise the coats of COP-I- and COP-II-coated vesicles. The components of the COP-I coat interact with the KKXX motif that is responsible for the retrieval of ER proteins from the Golgi apparatus, and is consistent with the role for COP-I-coated vesicles in recycling from the Golgi to the ER. The budding of clathrin-coated and COP-I-coated vesicles from the trans-Golgi network requires the activity of a GTP-binding protein called *ARF (ADP-ribosylation factor)* (Figure 34). ARF is related to Ras proteins, which function as oncogenes in human cancers. ARF bound to GTP associates with the Golgi membranes and is required for the binding of either COP-I-coat components or clathrin adaptins.

Several other Ras-related GTP-binding proteins have also been characterized in the secretory process. These include more than 30 Ras-related proteins (termed Rab proteins) that are implicated in vesicular transport in eukaryotic cells.

Two types of events characterize fusion of the vesicle with its target. First, the transport vesicles recognize the correct target membrane. Second, the vesicle and target membranes fuse, thus delivering the contents of the vesicle

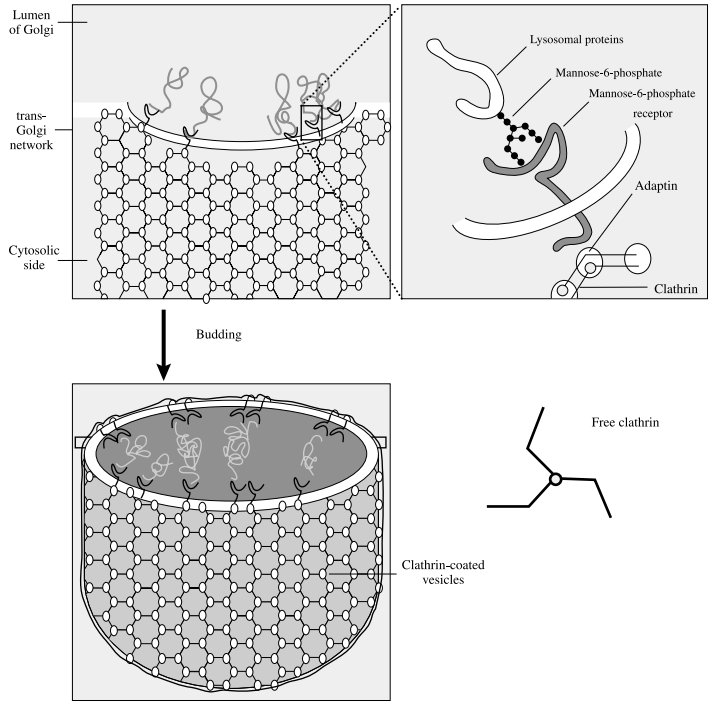

FIGURE 33. The incorporation of lysosomal proteins into clathrin-coated vesicles. Proteins targeted for delivery to lysosomes are marked by mannose-6-phosphates, which bind to mannose-6-phosphate receptors in the trans-Golgi network. The mannose-6-phosphate receptors span the Golgi membrane and function as binding sites for cytoplasmic adaptins, which in turn bind clathrin. Clathrins comprise three protein chains that associate with each other to form a lattice structure that distorts the membrane and promotes vesicle budding.

to the target organelle. Recognition between the vesicle and its target is mediated by interactions between unique pairs of transmembrane proteins. In contrast, fusion between the vesicle and target membranes arises from the action of general fusion proteins.

Biochemical analyses of reconstituted vesicular transport systems from mammalian cells have defined two classes of proteins involved in vesicle

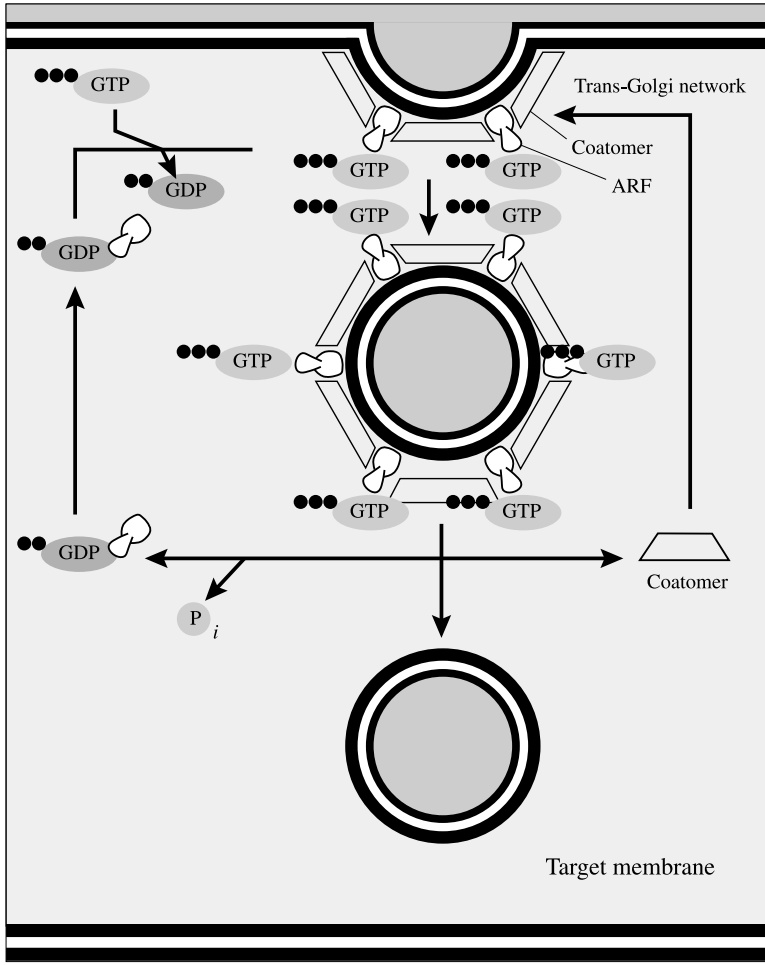

FIGURE 34. The role of ARF (ADP ribosylation factor) in the formation of COP-coated vesicles. ARF alternates between GTP-bound and GDP-bound states. When bound to GTP, ARF associates with the membrane of the trans-Golgi network and promotes the binding of COP-coat protein (coatomer). This leads to the budding of vesicles. The hydrolysis of the bound GDP then converts ARF to the GDP-bound state. This leads to the disassembly of the vesicle coat prior to fusion with the target membrane. The GDP-bound ARF is subsequently reconverted to the GTP-bound state. This is mediated by the action of a Golgi membrane protein that promotes a GDP–GTP exchange process. This leads to another cycle of coatomer assembly.

fusion: NSF and SNAPs. NSF (N-ethylmaleimide-sensitive fusion) is a soluble cytoplasmic protein that binds to membranes with other proteins called SNAPs (soluble NSF attachment proteins). NSF and SNAPs bind to families of specific membrane receptors called SNAP receptors or SNAREs. According to the SNARE hypothesis, interactions between specific vesicle SNAREs (v-SNAREs) and target SNAREs (t-SNAREs) membranes dictate the specificity of the vesicle fusion. Following specific vesicle–target interaction, the SNARE complex recruits NSF and SNAPs, resulting in the fusion of the vesicle and target membranes. For example, transport from the ER to the Golgi requires SNAREs that are located on both the vesicle and target membranes. These interactions are additionally regulated by the Rab GTP-binding proteins that are essential for vesicle transport. The SNARE hypothesis provides a central framework for understanding the molecular mechanisms of vesicle docking and fusion.

The major functions of lysosomes relate to the digestion of material taken up from outside the cell by *endocytosis.* Lysosomes are formed by the fusion of transport vesicles arising from the trans-Golgi network with endosomes, which contain the molecules taken up by endocytosis at the level of the plasma membrane. Acid hydrolyases are targeted to lysosomes by mannose 6-phosphate residues, which are recognized by mannose 6-phosphate receptors in the trans-Golgi network and packaged into clathrin-coated vesicles. After removal of the clathrin coat, these transport vesicles fuse with endosomes, and the acidic internal pH results in dissociation of the hydrolyases from the mannose 6-phosphate receptor. The hydolyases are thus released into the lumen of the endosome. The endosome then matures into a lysosome as it acquires a full compliment of acid hydrolyases that digest the molecules taken up by endocytosis.

4.6 Conclusion: Cystic Fibrosis as a Paradigm of Mutations Leading to Alterations in Transciptional and Post-Transcriptional Processing of an Integral Membrane Transport Protein[99-101]

The largest family of membrane transport proteins consists of the *ABC transporters*, so designated because they contain a basic structural unit characterized by six transmembrane domains followed by a highly conserved ***A**TP **b**inding **c**assette*. One of the most important members of the ABC family of transporters is the gene responsible for cystic fibrosis. This gene encodes a protein, the *cystic fibrosis transmembrane regulator (CFTR)*, which functions as a Cl^- channel in epithelial cells.

Cystic fibrosis (CF) is the most common (1 in 2,500 newborns) lethal recessive genetic disease of Caucasians. The fundamental physiological abnormality in CF is characterized by failure of cyclic adenosine monophosphate (cAMP) regulation of chloride transport across epithelial cell

membranes. The CFTR maps to chromosome 7 and comprises 27 exons (i.e., 230 kb of DNA) that encode a glycosylated protein containing 1,480 amino acids with a molecular mass of 170 kilodaltons. The CFTR gene product has two transmembrane domains, each containing six membrane-spanning segments, two nucleotide-binding domains (NBD) and a regulatory (R) domain (Figure 35). The hydolysis of ATP occurs at the NBD sites, while the R domains play an inhibitory role in keeping the Cl^- channel closed. The closed state of the Cl^- channel arises through the dephosphorylation of the R domain.

The CFTR is restricted to the apical membrane domain of epithelial cells, where it functions as a cAMP-dependent channel that allows the selective transport of chloride ions across the epithelial cell membrane. The binding of ATP leads to the gating of the Cl^- channel. As well, the CFTR is regulated by phosphorylation, which is accomplished by the action of a cAMP-dependent protein kinase A (PKA). The phosphorylation of the R domain results in a conformational change that leads to the opening of the chloride channel. The phosphorylated R domain plays a stimulatory role by enhancing the interaction of NBDs with ATP. The binding of ATP by the NBDs and its subsequent hydrolysis control the opening and closing of the chloride channel. The activated CFTR conducts Cl^- out of the epithelial cell and functions as a regulatory switch that allows cAMP to inhibit Na^+ absorption through Na^+ channels, and stimulate Cl^- secretion through channels distinct from the CFTR.

Chloride conductance at the apical membrane domain is dramatically reduced in CF. This is explained on the basis of quantitative or qualitative alterations in the CFTR, such that the clinical phenotype of CF patients is characterized by the inability of epithelial cells to transport or secrete chloride. The specific deletion of 3 bp in exon 10 results in the loss of a phenylalanine residue at position 508 within one of the ATP-binding domains of the CFTR protein (ΔF508). This particular mutation is associated with 70% of the mutant alleles in CF. More than 800 additional mutations within the CF gene comprise the remaining 30% of the mutant alleles in CF.

The ΔF508 mutation, for example, results in defective post-translational processing and intracellular trafficking of the CFTR such that it does not reach the apical membrane domain. Other mutations in the CFTR reduce its function in CF patients by a variety of mechanisms that act at one or several points in the flow of DNA to RNA to protein. Five classes of CFTR mutations have been described, and the molecular consequences of these different classes of mutations are illustrated in Figure 35. However, the various classes of CFTR mutations are not mutually exclusive. For example, in the ΔF508 CF mutation the deletion of phenylalanine leads to misprocessing of the CFTR but also failure of the CFTR protein to respond normally to activation signals.

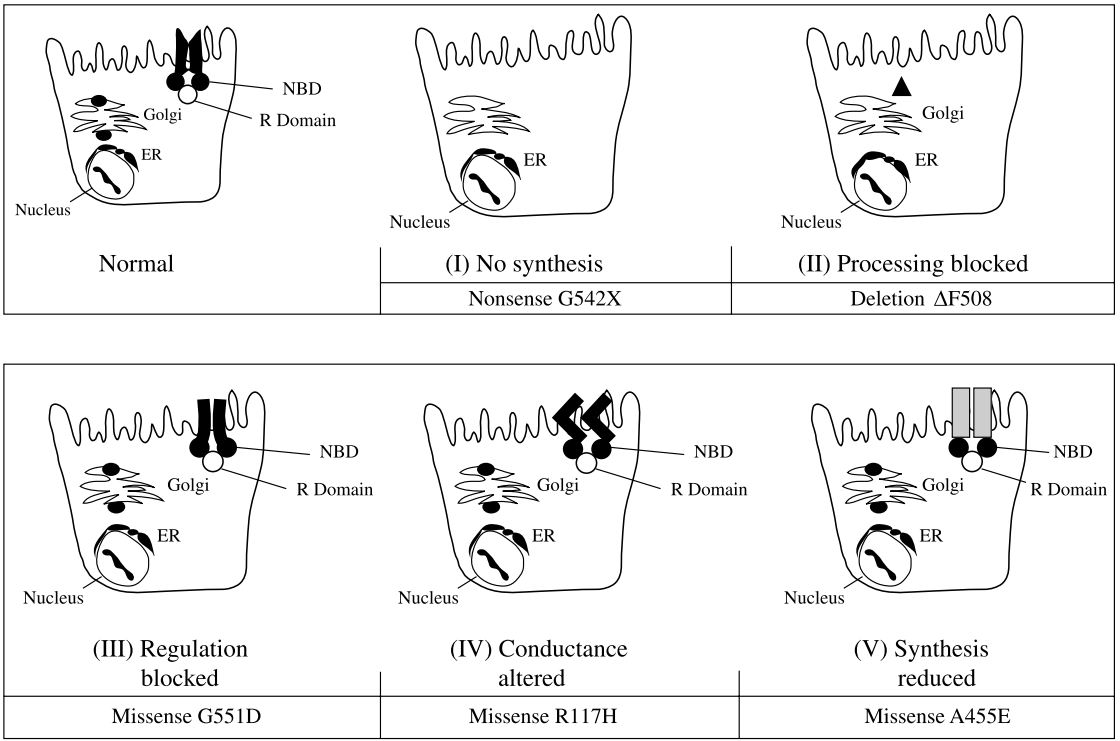

FIGURE 35. The five classes of CFTR gene mutations and the influence of these mutations on the expression of the CFTR gene product. The CFTR comprises a nucleotide-binding domain (NBD) and regulatory (R) domain.

In summary, mutations in the CFTR gene lead to alterations in transcription, post-transcriptional processing, translation and post-translational processing of the CFTR membrane protein along the secretory pathway. Importantly, the various types of CFTR mutations underscore the importance of each of these critical steps in the regulation of CFTR gene expression.

ABBREVIATIONS

DNA	Deoxyribonucleic acid
A	Adenine
G	Guanine
C	Cytosine
T	Thymine
U	Uracil
RNA	Ribonucleic acid
mRNA	Messenger RNA
tRNA	Transfer RNA
rRNA	Ribosomal RNA
hnRNA	Heterogeneous nuclear RNA
snRNA	Small nuclear RNA
UTR	Untranslated regions
SINEs	Short interspersed nuclear elements
LINEs	Long interspersed nuclear elements
G phase	Gap phase
S phase	Synthetic phase
M phase	Mitotic phase
Cdk	Cyclin-dependent kinases
CKI	Cyclin-dependent kinase inhibitors
INK	Inhibitor of Cdk
KIP	Kinase inhibitory protein

References

General References

1. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. Molecular cell biology. 3d ed. New York: WH Freeman, 1995.
2. Alberts B, Bray D, Johnson A, et al. Essential cell biology – an introduction to the molecular biology of the cell. New York: Garland, 1998.
3. Cooper GM. The cell – a molecular approach. Washington, DC: ASM, 1997.
4. Lewin B. Genes VI. New York: Oxford UP, 1997.
5. Glick BR, Pasternak JL. Molecular biotechnology – principles and applications of recombinant DNA. 2d ed. Washington, DC: ASM, 1998.

6. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular biology of the cell. 3d ed. New York: Garland, 1994.
7. Jameson JL. Principles of molecular medicine. New Jersey: Humana, 1998.
8. Strachan T, Read AP. Human molecular genetics. New York: Wiley-Liss, 1996.

The Cell Cycle

9. Hartwell LH, Kastan MB. Checkpoints: controls that ensure the order of cell cycle events. Science 1989; 246:629–634.
10. Murray AW. Creative blocks: cell-cycle checkpoints and feedback controls. Nature 1992; 359:599–604.
11. Norbury C, Nurse P. Animal cell cycles and their control. Annu Rev Biochem 1992; 61:441–470.
12. Morgan DO. Principles of CDK regulation. Nature 1995; 374:131–134.
13. Levine AJ. The tumor suppressor genes. Annu Rev Biochem 1993; 623–651.

DNA Replication

14. Blackburn EH. Telomerases. Annu Rev Biochem 1992; 61:113-129.
15. Diller JD, Raghuraman MK. Eukaryotic replication origins: control in space in time. Trends Biochem Sci 1994; 19:320–325.
16. Heintz NH, Dailey L, Held P, Heintz N. Eukaryotic replication origins as promoters of bidirectional DNA synthesis. Trends Genet 1992; 8:376-381.
17. Kelman Z, O'Donnell M. DNA polymerse III holoenzyme: structure and function of a chromosomal replicating machine. Annu Rev Biochem 1995; 64:171–200.
18. Roca J. The mechanism of DNA topoisomerases. Trends Biochem Sci 1995; 20:156–160.
19. Zakian VA. Telomeres: beginning to understand the end. Science 1995; 270:1601–1607.

Mutations and DNA Repair

20. Kolodner RD. Mismatch repair: mechanisms and relationships to cancer susceptibility. Trends Biochem Sci 1995; 20:397–401.
21. Leach FSE. Mutations of *mutS* homology in hereditary nonpolyosis colorectal cancer. Cell 1993; 75:1215–1225.
22. Modrich P. Mismatch repair, genetic stability, and cancer. Science 1994; 266:1959–1960.
23. Sancar A. Mechanisms of DNA excision repair. Science 1994; 266:1954–1956.
24. Seeberg E, Eide L, Bjoras M. The base excision repair pathway. Trends Biochem Sci 1995; 20:391–397.
25. Tanaka K, Wood RD. Xeroderma pigmentosum and nucleotide excision repair of DNA. Trends Biochem Sci 1994; 19:83–86.
26. Davis MM. T cell receptor gene diversity and selection. Annu Rev Biochem 1990; 59:475–496.

Eukaryotic Gene Transcription

27. Buratowski S. Mechanisms of gene activation. Science 1995; 270:1773–1774.
28. Grunstein M. Histones as regulators of genes. Sci Amer 1992; 267(4):68–74B.

29. Paranjape SM, Kamakaka, RT, Kadonaga, JT. Role of chromatin structure in the regulation of transcription by RNA polymerase II. Annu Rev Biochem 1994; 63:265–297.
30. Tjian R. Molecular machines that control genes. Sci Amer 1995; 272(2):54–61.
31. Tjian R, Maniatis T. Transcriptional activation: a complex puzzle with few easy pieces. Cell 1994; 77:5–8.
32. Goodrich JA, Cutler G, Tjian R. Contacts in context: promoter specificity and macromolecular interactions in transcription. Cell 1996; 84:825–830.
33. Beato M, Herrlich P, Schutz G. Steroid hormone receptors: many actors in search of a plot. Cell 1995; 83:851–857.
34. Gehring WJ, Qian YQ, Billeter M, et al. Homeodomain-DNA recognition. Cell 1994; 78:211–223.
35. Maniatis T, Goodbourn S, Fischer JA. Regulation of inducible and tissue-specific gene expression. Science 1987; 236:1237–1244.
36. Pabo CO, Sauer RT. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Annu Rev Biochem 1992; 61:1053–1095.

Eukaryotic RNA Polymerases and Basal Transcription Factors

37. Buratowski S. The basics of basal transcription by RNA polymerase II. Cell 1994; 77:1–3.
38. Conaway RC, Conaway JC. General initiation factors for RNA polymerase II. Annu Rev Biochem 1993; 62:161–190.
39. Young RA. RNA polymerase II. Annu Rev Biochem 1991; 60:689–715.
40. Zawel L, Reinberg D. Common themes in assembly and function of eukaryotic transcription complexes. Annu Rev Biochem 1995; 64:533–561.
41. Hanna-Rose W, Hansen U. Active repression mechanisms of eukaryotic transcription repressors. Trends Genet 1996; 12:229–234.

Post-Transcriptional Processing of RNA

42. Bird A. The essentials of DNA methylation. Cell 1992; 70:5–8.
43. Staudt LM, Lenardo MJ. Immunoglobulin gene transcription. Annu Rev Immunol 1991; 9:373–398.
44. Beelman CA, Parker R. Degradation of mRNA in eukaryotes. Cell 1995; 81:179–183.
45. Foulkes NS, Sassone-Corsi P. More is better: activators and repressors from the same gene. Cell 1992; 68:411–414.
46. Green MR. Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. Annu Rev Cell Biol 1991; 7:559–599.
47. Keller W. No end yet to messenger RNA 3′ processing! Cell 1995; 81:829–832.
48. Klausner RD, Rouault TA, Harford JB. Regulating the fate of mRNA: the control of cellular iron metabolism. Cell 1993; 72:19–28.
49. Maniatis T. Mechanisms of alternative pre-mRNA splicing. Science 1991; 251:33–34.
50. McKeown M. Alternative mRNA splicing. Annu Rev Cell Biol 1992; 8:133–155.
51. Bennett MM, Amara SG. Molecular mechanisms of cell-specific and regulated expression of the calcitonin/alpha-CGRP and beta-CGRP genes. Ann NY Acad Sci 1992; 657:36–49.
52. Zandberg H, Moen TC, Baas PD. Cooperation of 5′ and 3′ processing sites as well as intron and exon sequences in calcitonin exon recognition. Nucleic Acids Res 1995; 23(2):248–255.

53. Lou H, Cote GJ, Gagel RF. The calcitonin exon and its flanking intronic sequences are sufficient for the regulation of human calcitonin/calcitonin gene-related peptide alternative RNA splicing. Mol Endocrinol 1994; 8(12); 1618–1626.
54. Chan L. Apolipoprotein B messenger RNA editing; an update. Biochimie 1995; 77(1–2): 75–78.
55. Chan L, Chang BH, Nakamuta M, Li WH, Smith LC. Apobec-1 and apolipoprotein B mRNA editing. Biochim Biophys Acta 1997; 1345(1): 11–26.
56. Schibler U, Hagenbuchle O, Wellauer PK, Pittet AC. Two promoters of different strengths control the transcription of the mouse alpha-amylase gene Amy-1a in the parotid gland and the liver. Cell 1983; 33: 501–508.
57. Sierra F, Pittet AC, Schibler U. Different tissue-specific expression of the amylase gene Amy-1 in mice and rats. Mol Cell Biol 1986; 6(11):4067–4076.

Translation of mRNA

58. Hershey JWB. Translation control in mammalian cells. Annu Rev Biochem 1991; 60:717–755.
59. Klausner RD, Rouault TA, Harford JB. Regulating the fate of mRNA: the control of cellular iron metabolism. Cell 1993; 72:19–28.
60. Kozak M. Regulation of translation in eukaryotic systems. Annu Rev Cell Biol 1992; 8:187–225.
61. Merrick WC. Mechanism and regulation of eukaryotic protein synthesis. Microbiol Rev 1992; 56:291–315.
62. Noller HF. Ribosomal RNA and translation. Annu Rev Biochem 1991; 60:191–227.
63. Rhoads RE. Regulation of eukaryotic protein synthesis by initiation factors. J Biol Chem 1993; 268:3017–3020.

Protein Folding and Processing

64. Casey PJ. Protein lipidation in cell signaling. Science 1995; 268:221–225.
65. Clarke S. Protein isoprenylation and methylation at carboxy-terminal cysteine residues. Annu Rev Biochem 1992; 61:355–386.
66. Dalbey RE, von Heijne G. Signal peptidases in prokaryotes and eukaryotes – a new protease family. Trends Biochem Sci 1992; 17:474–478.
67. Englund PT. The structure and biosynthesis of glycolsyl phosphatidylinositol protein anchors. Annu Rev Biochem 1993; 62:121–138.
68. Freedman RB, Hirst TR, Tuite MF. Protein disulphide isomerase: building bridges in protein folding. Trends Biochem Sci 1994; 19:331–336.
69. Gething MJ, Sambrook J. Protein folding in the cell. Nature 1992; 355:33–45.
70. Gierasch LM. Signal sequences. Biochemistry 1989; 28:923–930.
71. Hart GW, Haltiwanger RS, Holt GD, Kelly WG. Glycosylation in the nucleus and cytoplasm. Annu Rev Biochem 1989; 58:841–874.
72. Hartl FU. Molecular chaperones in cellular protein folding. Nature 1996; 381:571–580.

Protein Sorting and Transport

73. Fischer von Mallard GB, Stahl CL, Sudhof TC, Jahn R. Rab proteins in regulated exocytosis. Trends Biochem Sci 1994; 19:164–168.
74. Mellman I. Protein mediators of membrane fusion. Cell 1995; 82:869–872.

75. Novick P, Brennwald P. Friends and family: the role of Rab GTPases in vesicular traffic. Cell 1993; 75:597–601.
76. Pelham HRB. About turn for the COPs? Cell 1994; 79:1125–1127.
77. Pryer NK, Wuestehube LJ, Schekman R. Vesicle-mediated protein sorting. Annu Rev Biochem 1992; 61:471–516.
78. Rothman JE. Mechanisms of intracellular protein transport. Nature 1994; 372:55–62.
79. Rothman JE, Wieland FT. Protein sorting by transport vesicles. Science 1996; 272:227–234.
80. Schekman R, Orci L. Coat proteins and vesicle budding. Science 1996; 271:1526–1533.
81. Whiteheart SW, Kubalek EW. SNAPs and NSF: general members of the fusion apparatus. Trends Cell Biol 1995; 5:64–68.

Protein Degradation

82. Ciechanover A. The ubiquitin-proteasome proteolytic pathway. Cell 1994; 79:13–21.
83. Dice JF. Peptide sequences that target cytosolic proteins for lysosomal proteolysis. Trends Biochem Sci 1990; 15:305–309.
84. Goldberg AL. Functions of the proteasome: the lysis at the end of the tunnel. Science 1995; 268:522–523.

The Endoplasmic Reticulum

85. Abeijon C, Hirschberg CB. Topography of gylcosylation reactions in the endoplasmic reticulum. Trends Biochem Sci 1992; 17:32–36.
86. Gilmore R. Protein translocation across the endoplasmic reticulum: a tunnel with toll booths at entry and exit. Cell 1993; 75:589–592.
87. Hendrick JP, Hartl FU. Molecular chaperone functions of heat-shock proteins. Annu Rev Biochem 1993; 62:349–384.
88. Hurtley SM. Golgi localization signals. Trends Biochem Sci 1992; 17:2–3.
89. Rapaport TA. Transport of proteins across the endoplasmic reticulum membrane. Science 1992; 258:931–936.
90. Sanders SL, Schekman R. Polypeptide translocation across the endoplasmic reticulum membrane. J Biol Chem 1992; 267:13791–13794.
91. Udenfriend S, Kodukula K. How glycosylphosphatidylinositol anchored membrane proteins are made. Annu Rev Biochem 1995; 64:563–591.
92. Walter P, Johnson AE. Signal sequence recognition and protein targeting to the endoplasmic reticulum membrane. Annu Rev Cell Biol 1994; 10:87–119.

The Golgi Apparatus

93. Burgess TL, Kelly RB. Constitutive and regulated secretion of proteins. Annu Rev Cell Biol 1998; 3:243–293.
94. Machamer CE. Targeting and retention of Golgi membrane proteins. Curr Opin Cell Biol 1993; 5:606–612.
95. Pelham HRB, Munro S. Sorting of membrane proteins in the secretory pathway. Cell 1993; 75:603–605.

Lysosomes

96. Dunn WA, Jr. Autophagy and related mechanisms of lysosome-mediated protein degradation. Trends Cell Biol 1991; 266:21327–21330.
97. Kornfeld S. Structure and function of the mannose 6-phosphate/insulinlike growth factor II receptors. Annu Rev Biochem 1992; 61:307–330.
98. Kornfeld S, Mellman I. The biogenesis of lysosomes. Annu Rev Cell Biol 1989; 5:483–525.

Cystic Fibrosis

99. Collins FS. Cystic fibrosis: molecular biology and therapeutic implications. Science 1992; 256: 774–779.
100. Zielenski J, Tsui LC. Cystic fibrosis: genotypic and phenotypic variations. Annu Rev Genet 1995; 29:777–807.
101. Davis PB, Drumm M, Konstan MW. Cystic fibrosis. Am J Respir Crit Care Med 1996; 154:1229–1256.

Acknowledgments

This work was supported by operating grants from the Medical Research Council of Canada and the Crohn's and Colitis Foundation of Canada. Dr. Gary E. Wild is a senior clinician scientist of the Fonds de la recherche en santé du Québec. Dr. Wild wishes to extend his appreciation to Drs. David Fromson, John Southin, Howard Bussey and Bruce Brandhorst of the McGill Biology Department. Their tireless efforts in the area of undergraduate science education fostered a sense of inquiry and collegiality that guided a cohort of students through the early recombinant DNA era.